esomar Congress 2026

METAMORPHOSIS

# PAPERS

Copyright © **Esomar** 2026

# Table of Contents


## Editorial
Introduction to the Compendium of Papers
*Adrian Terron*

## The Sourdough Principle
What Happens When You Stop Throwing Away Your Starter
*Sushma Panchawati · Kelli Johnson · Kristina Lenz · Scott Peiser*

## When AI Sees but Doesn't Understand
Evidence From 50+ Parallel Tests of AI Versus Human Eye-Tracking
*Patrick Young · Olivier Blanchet*

## Smart Labels, Real Homes, True Usage
Closing the Say-Do Gap in Product Testing with Passive Observations
*Nirmit Behal · Karolina Dobrzyńska · Bert Nijhuis · Susan Purcell · Omkar Shinde · Pratik Rane · Shreeraj Dhekane · Jinho Cho · Nikolai Reynolds*

## Simulate the "What If" Before You Build
Synthetic Populations as an Exploration-First Layer for Customer Insight, and the Governance that Keeps It Honest
*Mariana Camillo · Aline Souza, Leonardo Aragão · Fernanda Laurentys Airoldi · Claudia Furniel · Danny Wymmer*

## Market Research from Evaluator to Architect
How an Insight Team Re-Engineered the Casting Decision: A Vodafone Case Study
*Dirk Ziems · Christian Escher*

## Tuned to Fit
Reporting Standards for LLM-Based Silicon Samples
*Jan Polišenský · Matej Koreň · Radek Hranický · Kamila Zahradníčková*

## Metamorphosis in Practice: From Static Insights to Creative Springboards
How Culture-Led Insight Sprints Helped Unilever Build Briefs That Inspire Stronger Creative Outcomes Faster
*YeeMei Chan · Lanny Li*

## Decoding the Gold Standard
From Niche Perspective to Sophisticated Balance: How AB InBev and Ipsos Decoded the Wellness Mainstream
*Marcela Ayres · Gustavo Araujo de Castro · Fernanda Wajchenberg · Tiago Fernandes*

## From Tool to Teammate
Embedding AI into the Everyday Workflow of Insights and Marketing Teams at Philips
*Mathilde Beljaarts · Fenny Léautier*

## Historical Survey Data into New Insights Using Semantic Similarities
*Douglas Ethé-Sayers · Kristina O'Keeffe · István Hidegkuti PhD*

## Digital Twins: From Theory to Practice
An Exploration of the Capabilities of Digital Twin and LLM-Generated Insights for Commercial Research
*Jon Puleston · David Priestley · Maciek Ozorowski · Davide Baldo*

## Research During Wartime

How Insight Infrastructure Survives a Mega-Crisis
*Evgeniya Blyznyuk*

**Ten Times the Research. The Same Team. A Different Practice.**
*Chris Courtney-Smith · Leanne Tomasevic*

**Not New Methods, New Craft**
Integrating Present-Forward and Future-Back Insight
*Rachel Hansen · Kimberly Krenek*

**Talk to My Twin**
When Human and Synthetic Intelligences Debate the Future of Insight
*Christine Antoine Simonet · Mathilde Guinaudeau · Benoit Hubert*

**AI-Enabled Insight and Foresight Transformation at Global Scale**
Redesigning the Craft of Insight for Speed, Scale and Strategic Impact
*Tomasz Soluch*

**From Synthetic Data to Synthetic Epistemology**
*Konrad Suchecki*

**Mondelez International Gastronomy Radar with Nextatlas**
Predicting the Next Wave of Food Flavours and Ingredients
*Mario Coletti · Serhiy Kalinovsky*

**The Mars Digital Twin Report Card**
Measuring What AI Personas Get Right (and Wrong)
*Adam Bai · Johanna Thompson · Richard Owen*

**Simulating Tomorrow for Coca-Cola's Advanced Hydration Portfolio**
Future twin communities in action
*Thomas Troch · Annelies Verhaeghe · Adam Palenicek*

**Embracing the Unknown...**
In Search of a Predictive Model Grounded in Real Human Understanding. Madness – or Method?
*Muriel Esposito · Laurence Minisini*

**Decoding the dine-in micro-moments that build brand love for Nando's BA and Nando's**
*Anneri Venter · Simone Kakana*

**Deep Breath**
Uncovering the Missing Dimension in Consumer Emotions
*Graham Page*

**Marketing's Most Expensive Oversight**
Teaching Researchers and Marketeers How to Engage with the Neurodiverse Community
*Alexander Holmes · Jessica Garlick · Heidi Kim · Sebastian Giugovaz*

**The Ripple Effect**
Turning a 60+ Market Brand Tracker into a Strategic Decision System
*Alexander Shmelkov · Sylvia van der Waal · Muge Vardar*

**Citizen Intelligence in Action**
Turning Mobility Insights into Public Transport Change

*Raquel Marzo · Josep Chiner · Amparo López-Vicente · Carol Soriano · Juan F. Giménez · Elisa Signes · José Solaz · Raquel Alario*

**Anticipating the Future Consumer**
Using Digital Behavioural Data and AI to Size Global Macro Forces and Detect Emerging Signals
*Emily Dobbing · Giulia Iorio-Ndlovu*

**Where Fraud Meets Its Friction**
Preventing Survey Fraud at Reward Redemption
*Frank Kelly · Justin Hill · Mardi Hartzog*

**We Do Not Understand People**
The Empathy Illusion in Financial Anxiety Research
*Jaime Solórzano Villela*

**From Raw Text to Executive Insight**
Evaluating LLM Summarisation Workflows for Customer Complaint Analytics
*Joanna Quan · Wenqi Shen · Chitvan Trivedi*

**A Dual Metamorphosis**
How Market Research Evolves with the Evolving Brand, Social Purpose and Public Opinion
*Linda Di Luzio PhD · Kathy Cheng*

**The Missing Metric**
Why Social Health Is the Next Frontier of Workplace Well-Being
*Ellen Carnasciali · Ferdinando Gomes · Enric Cid · Federico Zambrano*

**Detect, Trace, Learn**
Turning suspect interviews into actionable fraud patterns
*Sebastian Berger · Julia Mittermayr*

**Human + Machine**
Rebuilding Representation in AI-Driven Research
*Mario X. Carrasco*

**The Strategic Growth Gap**
How Insight is Rewiring Growth Decisions
*Lucas Hulsebos*

**Dialogues Within**
A Dialogical Self-Exploration of Men Becoming Fathers
*Sandeep Dutta · Paroo Tikoo*

**The Wellbeing Shift**
A Human–AI Metamorphosis of Brand Positioning
*Charlene Van Zyl, Karlien Kriegler, David Wright*

**Closing the Protection Gap**
From Insight to Intervention in Preventing Online Child Sexual Exploitation
*Joy Ganvik · Anju Ahuja*

**When Machines Lose Their Moral Nerve**
*Ross Denton*

**Augmenting, Not Replacing**

Integrating Simulated Audiences into Real Research Workflows
*Romani Patel · Kate Weymer · Robert Graves*

## When Grief Walks Into Work
From Mixed-Method Research to AI-Driven Tools That Support Grievers at Scale
*Gina Pingitore, PhD · Pete Trujillo, MSAI*

## Designing the Future of Sport
Mapping Emerging Communities to Shape Creative Territories
*Sarah Laurier · Rodolphe Pal*

## Not Another Gen-Z Brief
How to Form a New Contract with Older Adults
*Scott Percival · Drago Djourov*

## Conversations with Agents
Turning complex consumer data into confident innovation bets
*Dave Lundahl · John Ennis · Mimi Sherlock · Greg Stucky*

## Metamorphosis of Expertise
How AI is Redefining Value in the Insights Industry
*Kamila Zahradníčková, MSc*

## Agentic AI Application for Rapid Creative Production Rooted in Consumer Insights
*Minh Nguyen, PhD · Takumi Akuzawa · Togo Kida · Toshiaki Kamiya · Yuki Tanaka · Takamata Shoshi · Otsuki Erika · Takeuchi Ami · Munish Bhamoriya*

## Gen Alpha Outsmarts Research
Why They Won't Tell Us The Truth
*Jaime Solorzano · Rebecca Boden*

## Creative Consequences
Why the World’s Strongest Brands Create the World’s Best Advertising
*Vera Sidlova · Polly Wyn-Jones · Horacio Gonzalez · Felipe Thomaz · Neta Livneh*

# Editorial

## Introduction to the Compendium of Papers

*Adrian Terron*

Esomar

Office address:

Burgemeester Stramanweg 105

1101 AA, Amsterdam

Phone: +31 20 664 21 41

Email: info@Esomar.org

Website: www.Esomar.org

Publication Date: September 2026

Esomar Publication Series Volume S424 Congress 2026

ISBN 92-831-0337-8

Copyright © **Esomar** 2026

Copyright © **Esomar** 2026



# Editorial

## Introduction to the Compendium of Papers

*Adrian Terron*

### The term “metamorphosis” has never been about change for the sake of change

It has never been about the clichés that accompany “change" as a phenomenon: its inevitability, its constancy, its suddenness. What its spirit is meant to capture is the process of change: its trials, its triumphs, its tantalising ability to glimpse a future unknown but immeasurably exciting. “Meta” means to go beyond, “morph” means to change and “osis” refers to the continuous. Together, they help us transcend boring dogma and creaking convention to leap over the unsolved problems of human understanding that continue to constrain us. Today, the latest in a series of human inventions arms us with a powerful catalyst to govern this change and determine its course, by using it as collaborator rather than competitor, and by choosing to harness it correctly rather than letting it lash us to its whim. Whether you call it artificial intelligence, computational systems or by any assortment of terms, it brings up fundamental questions of trust, accountability and effectiveness. Questions that place us squarely in the position of deciding how we want to employ it to enhance the outcomes that best serve us as individuals, organisations and communities.

The papers in this edition of the Esomar Congress offer us a new role to adopt: that of alchemists. The kind that don't need the "Philosopher's Stone" to turn substances into gold for personal profit. Instead, the kind that can use community, the Esomar community, to convert everything we already have into so much more. Whether it is re-architecting human experience, revitalising the representation of previously ignored segments or revisiting how to unleash our innate curiosity and creativity, the collection of impactful ideas in this edition is tinged with an undeniably transformational potential. As you read through them, I hope you see what the programme committee has seen in them: the startling ability to make change, and the continuous learning that accompanies it: the connective tissue that makes the insights community the centrifugal force for positive outcomes for everyone.

### Adrian Terron

Esomar Congress Programme Committee Chair.

Head of Corporate Brand and Marketing Strategy, Tata Group, India.

Copyright © **ESOMAR** 2026

esomar

# Copyright

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system of any nature, or transmitted or made available in any form or by any means, electronic, mechanical, photocopying, recording or otherwise, without the prior written permission of Esomar. Esomar will pursue copyright infringements.
In spite of careful preparation and editing, this publication may contain errors and imperfections. Authors, editors and Esomar do not accept any responsibility for the consequences that may arise as a result thereof. The views expressed by the authors in this publication do not necessarily represent the views of Esomar.
By the mere offering of any material to Esomar in order to be published, the author thereby guarantees:

- that the author - in line with the ICC/Esomar International Code of Marketing and Social Research– has obtained permission from clients and/ or third parties to present and publish the information contained in the material offered to Esomar;
- that the material offered to Esomar does not infringe on any right of any third party; and
- that the author shall defend Esomar and hold Esomar harmless from any claim of any third party based upon the publication by Esomar of the offered material.

Published by Esomar, Amsterdam,
The Netherlands
Edited by: Rosemary Carlton-Willis

Copyright © **ESOMAR** 2026

# About Esomar

Esomar is the champion of the insights and analytics sector. It is the business community for every sector professional. Founded in 1947, the global membership association is a network reaching over 40,000 professionals and 750+ companies in 130+ countries. We support individual and corporate members supplying or using insights by helping them raise ethical standards, grow the demand for insights, and improve its uses and applications by all decision-makers.

## About Esomar Membership

Esomar is open to everyone, all over the world, who believes that high quality research improves the way businesses make decisions. Our members are active in a wide range of industries and come from a variety of professional backgrounds, including research, marketing, advertising and media.
Membership benefits include the right to be listed in the Esomar Directories of Research Organisations and to use the Esomar Membership mark, plus access to a range of publications (either free of charge or with discount) and registration to all standard events, including the Annual Congress, at preferential Members' rates.
Members have the opportunity to attend and speak at conferences or take part in workshops. At all events the emphasis is on exchanging ideas, learning about latest developments and best practice and networking with other professionals in marketing, advertising and research. CONGRESS is our flagship event, attracting over 1,000 people, with a full programme of original papers and keynote speakers, plus a highly successful trade exhibition. Full details on latest membership are available online at www.Esomar.org.

# Contact us

Esomar
Esomar Office:
Burgemeester Stramanweg 105
1101 AA, Amsterdam
The Netherlands
Tel.: +31 20 664 21 41

Email: info@Esomar.org
Website: www.Esomar.org

# The Sourdough Principle

## What Happens When You Stop Throwing Away Your Starter

*Sushma Panchawati*

*Kelli Johnson*

*Kristina Lenz*

*Scott Peiser*

## Esomar

Office address:

Burgemeester Stramanweg 105

1101 AA, Amsterdam

Phone: +31 20 664 21 41

Email: info@Esomar.org

Website: www.Esomar.org

Publication Date: September 2026

ESOMAR Publication Series Volume S424 Congress 2026

ISBN 92-831-0337-8

Congress 2026
Copyright © **ESOMAR** 2026

Copyright © **ESOMAR** 2026

Congress 2026

# The Sourdough Principle

## What Happens When You Stop Throwing Away Your Starter

*Sushma Panchawati*

*Kelli Johnson*

*Kristina Lenz*

*Scott Peiser*

### 1. Introduction

Every established brand keeps a sourdough starter without knowing it. Years of research, including tracking studies, concept tests, ethnographies and strategy decks pile up, get used once, and go cold. When a new question comes, the habit is to throw the starter out and buy fresh flour: commission a new study, recruit a new sample, wait weeks, spend heavily.

That habit is getting harder to justify. The old reason for always starting fresh was legitimate – there was too little relevant data. Most companies are now buried in information they never use. Forrester estimates that 60 to 73% of enterprise data goes unanalysed (Forrester Research, 2016); Gartner calls the unused remainder "dark data" (Gartner, n.d.). In one survey, three in four organisations admitted making decisions without consulting data they already held, and nearly half saw innovation work delayed or dropped as a result (Sisense, 2025).

The starter is in the fridge. We keep buying flour.

We are drowning in data, and still we keep buying more.

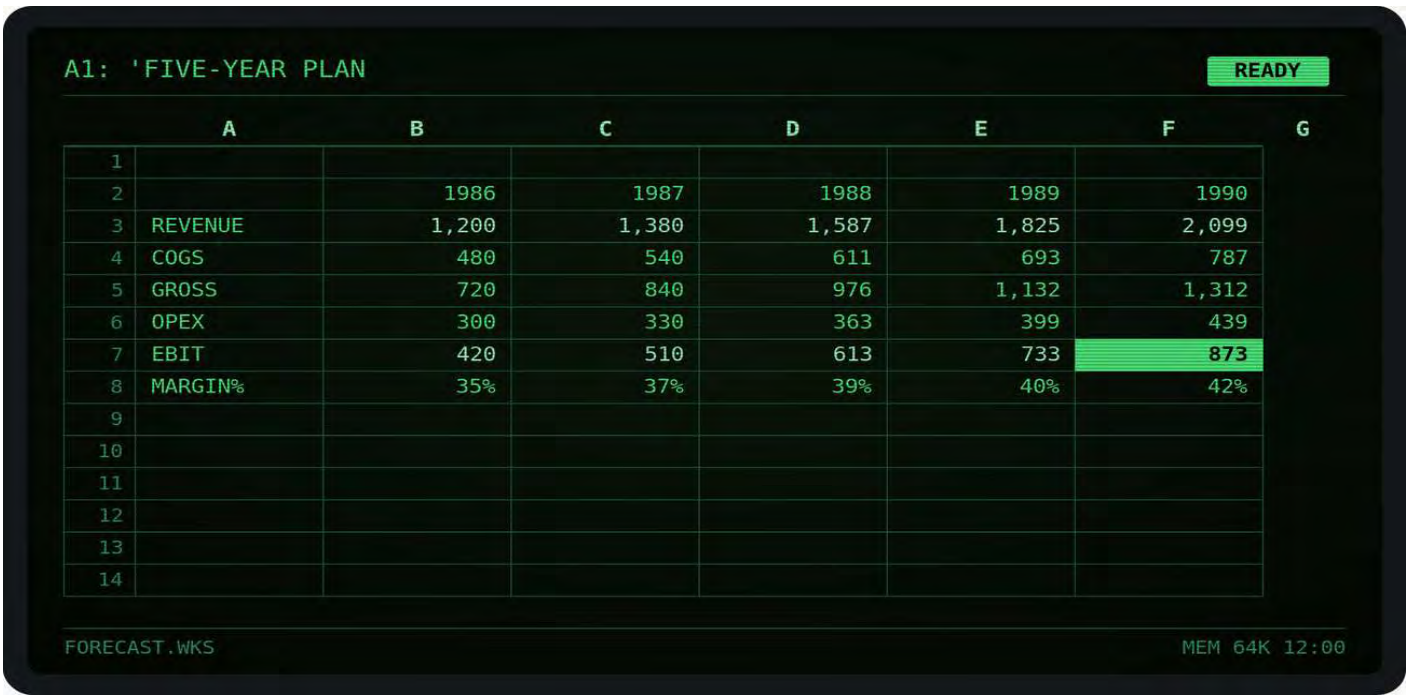
This paper is about what happens when a brand feeds the starter instead. KraftHeinz, Pepper Strategies and Qlue faced a fast-turn innovation brief for Philadelphia, one of the most loved cream cheese brands in America. Rather than start with new fieldwork, we started with what the brand already had: dozens of past studies, reconnected, fed with fresh consumer voices and worked through with AI. The result was a full innovation sprint in weeks instead of months – and it delivered enough ROI that Kraft commissioned five more briefs on the back of it.

A sourdough starter is never finished. Each bake feeds the next, and the culture gets richer with every loaf. Research can work the same way, if you keep it alive. Our argument is simple: when data is abundant, the value of research shifts from collecting what is new to curating what you already have – and that curation requires human judgment, which AI can sharpen but not entirely replace.

### 2. A Lesson from the Spreadsheet

Curating well is harder than it sounds, and it cannot be handed to a machine. The reason extends beyond AI. Every powerful tool produces answers faster than it produces understanding and tempts us to trust the output more than we should.

Steven Levy saw this in 1984, when the spreadsheet was new. For the first time, it let anyone model the future: lay a company's next five years across a grid and watch the numbers settle into a clean, confident answer. The trouble was that the answer rested on assumptions buried in the model, and once it appeared on screen it was treated as fact – because it looked authoritative and a computer had produced it.




| | A | B | C | D | E | F | G |
|---|---|---|---|---|---|---|---|
| 1 | | | | | | | |
| 2 | | 1986 | 1987 | 1988 | 1989 | 1990 | |
| 3 | REVENUE | 1,200 | 1,380 | 1,587 | 1,825 | 2,099 | |
| 4 | COGS | 480 | 540 | 611 | 693 | 787 | |
| 5 | GROSS | 720 | 840 | 976 | 1,132 | 1,312 | |
| 6 | OPEX | 300 | 330 | 363 | 399 | 439 | |
| 7 | EBIT | 420 | 510 | 613 | 733 | 873 | |
| 8 | MARGIN% | 35% | 37% | 39% | 40% | 42% | |
| 9 | | | | | | | |
| 10 | | | | | | | |
| 11 | | | | | | | |
| 12 | | | | | | | |
| 13 | | | | | | | |
| 14 | | | | | | | |

*Figure 1.*

A confident-looking answer is the easiest kind to stop questioning.

AI presents the same temptation, but it's even harder to resist. A weak line of reasoning now arrives complete and assured, so it travels further before anyone checks it (Duncan, 2026). For research, whose whole job is to earn trust, this is the hazard to design against.

Reading an archive well – deciding what still holds, what contradicts, what to believe – is the opposite of taking a confident output at face value. It is judgment. That is where our thesis begins.

## 3. Our Thesis

For most of its history, research has produced a record. A study captures what people said, in one place, about one question, at one time. It gets filed, and its value starts fading the day it lands. That is why the industry keeps commissioning more: a record goes stale, and freshness was the only way to keep it relevant and useful.

That can now change. A body of past research no longer has to sit as a record. It can be reasoned forward. It used to tell you only what consumers had already said. Now you can ask it new questions and get answers in the voice of the real people it was built from.

This is the difference between a photo album and a weather model. An album holds the past and answers one question: what happened.

Copyright © **ESOMAR** 2026


A weather model is also built from the past, but it does what the album cannot: it reasons about conditions based on past patterns and predicts what is likely in the future.

Research is moving from the album to the weather model.

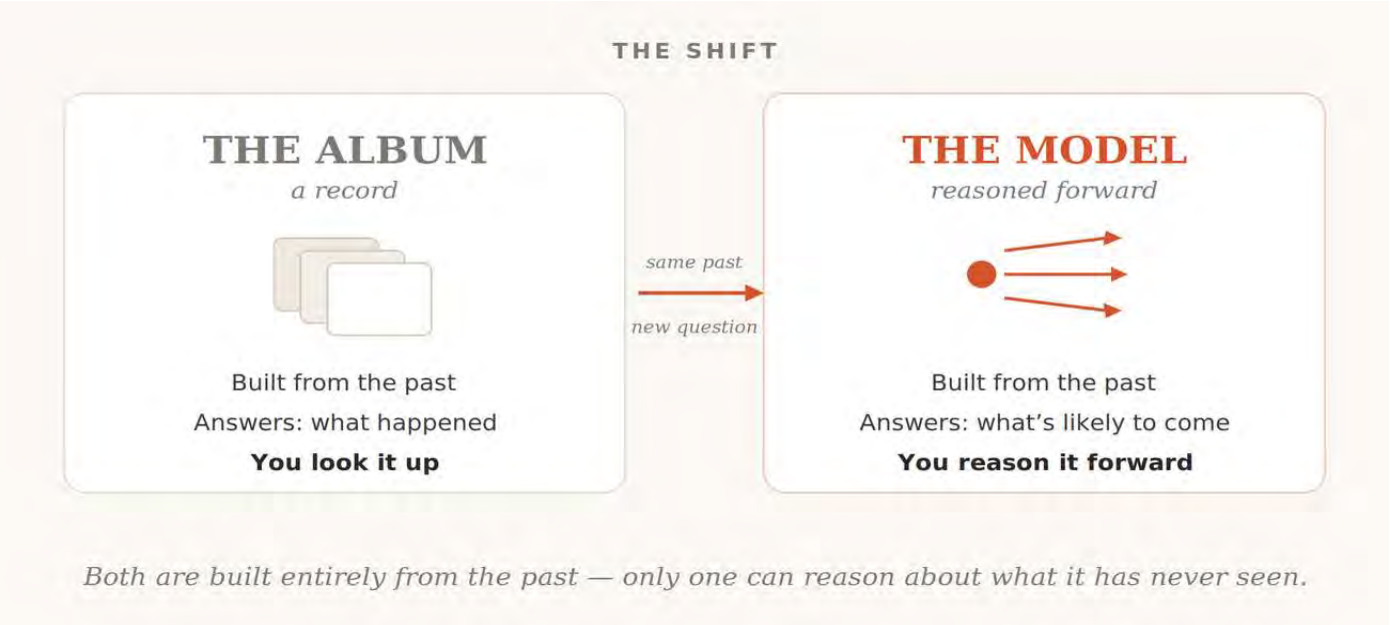



*Figure 2.*

What is new is not looking the past up. It is reasoning the past forwards.

Knowledge management systems and enterprise GPTs help with retrieval. Reasoning forward is different: you put a question to everything a brand knows about its audience, the market, the culture and the competition and get back a considered, multi-dimensional answer.

That shift changes the work in two ways.

1. Value moves from collecting to curating. If the model can answer, the scarce skill is no longer gathering more but building the model well and keeping it fed – connecting what a company already holds, and adding to it on purpose instead of starting over each time.
2. The work becomes judgment, not production. A model that answers fluently is not the same as one that answers truly, and the value now sits in the human ability to tell them apart and to catch the model speaking confidently about something it cannot know.

## 4. The Case: Philadelphia's Next Chapter

KraftHeinz came to us with one question: where should the brand Philadelphia go next? The brand has near-universal awareness, deep trust, and a hold on cream cheese so complete the two names are nearly the same word. But that dominance rests mostly on a single format. The question was whether the trust Philadelphia has earned in one product format could stretch to others, and in what shape.

The usual route would run in sequence: a landscape study, then qualitative work, then concept development, then validation. Months of work, each stage commissioned fresh. The timeline did not allow it. Philadelphia had been studied

for years. The intelligence already existed, scattered across documents and mostly unused since the projects that produced them.

The starter was in the fridge. It needed feeding, not replacing.

## *The Method: Feeding the Starter*

### *Step 1: Framing*

The brief was ambitious: where should one of the most established brands in its category extend next in terms of product innovation and portfolio extension? To answer this, we needed to explore the possibilities through multiple dimensions. The strategist sets the frame using their experience and expertise in solving similar problems. We arrived at five dimensions for evaluating which path the brand should take:

1. Brand permission: Can the brand credibly play here? Activated through the brand's equity, history and the meanings it already owns. We used brand tracks, brand stories, brand tracking & other related data points to map out the brand boundaries.
2. Category dynamics: Does it fit where the market is going? We used category reports, syndicated data, trend watch reports, and created a context of larger macro trends that affect food and dairy like the rise of protein and macros on TikTok and Instagram, the way GLP-1 medications are quietly rewriting how millions of people eat.
3. Consumer validation: Has something like this been tried before? Read from the brand's own past concept tests and consumer research. Before chasing a direction, this dimension checks the record for prior evidence: something that worked, something that failed, something never properly tested.
4. Occasion fit: Does it serve a real job to be done? Read from consumer behaviour and demand spaces related research data. A space only matters if it answers a job people are genuinely hiring food to do.
5. Business strategy: Is it aligned with the three-year plan? Read from the company's own strategy, R&D priorities and three-year plans. This was done to ensure whatever paths we came up with aligned with the company's larger goals, priorities and ambitions.

These five became the frame: the structure every later question was put through, and the bar every candidate space had to clear. Designing it is pure judgment, and it is where the strategist's experience really shines.

### *Step 2: Activate the Data*

With the frame set, the next move is to bring the data alive: we did this in two ways.

1. Using past wisdom: This is activated through a strategic agent, which works by reading through the different documents in context and reasoning, connecting and layering them to answer questions like: "What are the five biggest frustrations with snacking"; "Which categories within dairy are seeing an upward trend in terms of consumer chatter online"; and "Which of the following five categories come closest to Philadelphia cream cheese on the retail aisle".
2. Building digital twins of consumers: These are not synthetic respondents drawn from a population average, but were built qualitatively, from the brand's own past work and from real consumers, to map the actual target audience the brand cared about: their language, their priorities and the tensions they were already trying to resolve. Each twin had 300–400 data points that acted as narrow and specific context, giving the twin a detailed

understanding of the consumer's mental models about the category, their needs and perceptions, and the triggers and hold-backs that shape what they actually do, not what they say they do.

Why we used digital twins: An average synthetic persona reasons like no one in particular. A twin drawn from specific real people, simulates the person they were based on: their thinking, choices, behaviours, preferences and individual lifestyle context. Generic synthetic respondents track real people on simple measures but fall apart on harder analysis (Strat7 with Dunnhumby, 2025). These types of digital twins have been shown to reproduce real groups with a fidelity that generic averages never reach (Argyle et al., 2023).

### *Step 3: Iterative Simulations*

With the frame set and the data live, the real work begins. Step three is where the AI's volume meets the strategist's discipline. It has three parts: designing the questions, putting them to the data in a way that reasons forward, and running the analysis iteratively after each question, at a scale that an AI engine makes possible for the first time.

Answering a brief like this one – where should the brand innovate next – cannot be answered with one push button. These upstream, multi-dimensional problems need multi-dimensional and layered analysis.

The strategist designed the interrogation: we asked about 50 questions of the twins, and past research, that gave us the lay of the land, opportunity spaces, etc., in the beginning, and we asked more specific questions about each potential innovation space as it became clearer.

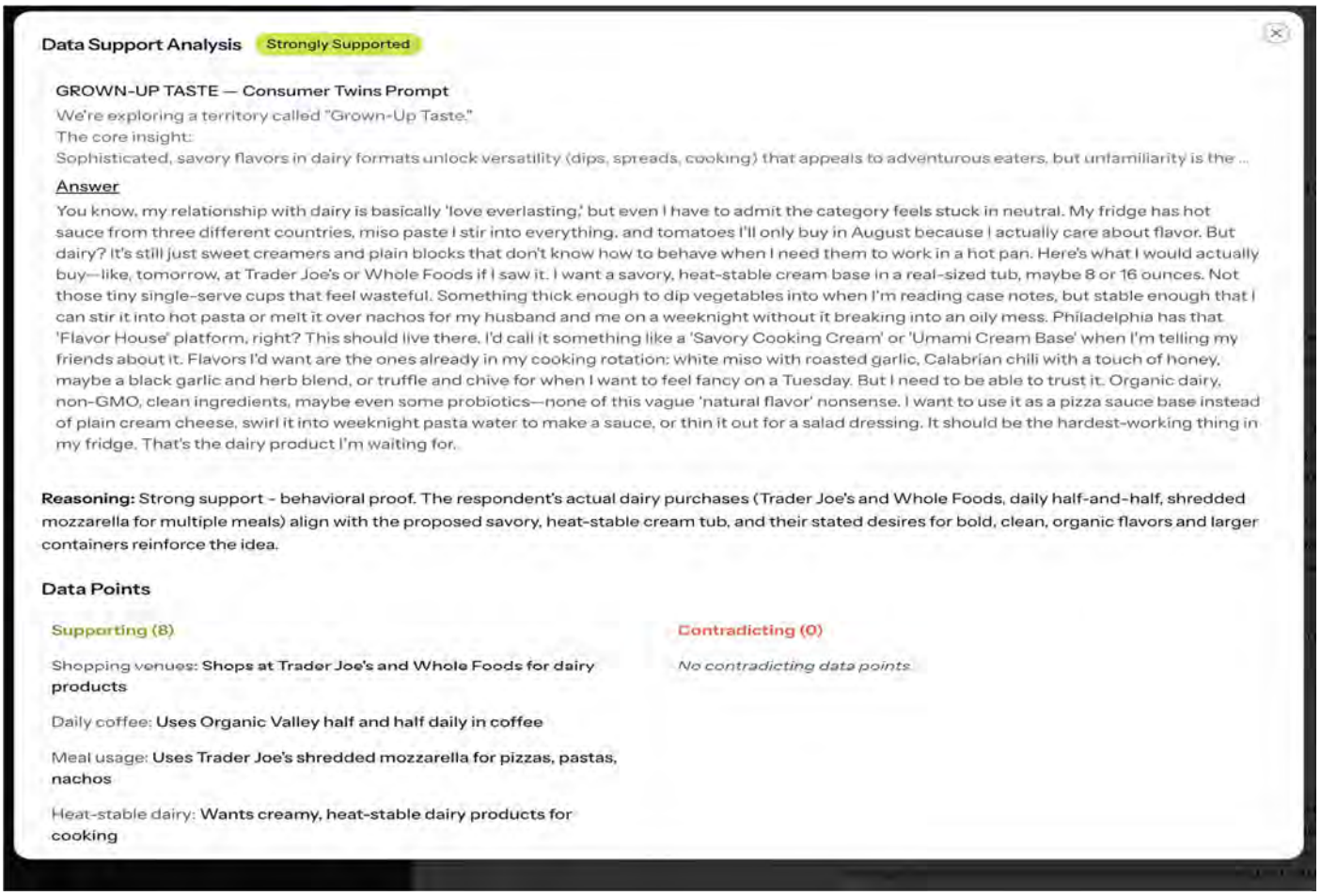



*Figure 3. This is a screenshot of how the digital twin's response is validated against source data.*

At this stage, we had arrived at 12–15 potential dig sites: the initial territories for innovation Philadelphia could credibly explore. Each carried its evidence underneath it – what years of consumer work had already shown, where the category was pulling toward it, what the cultural moment was doing alongside, what occasion it answered and where it sat against the company's own strategy. These were not yet recommendations; they were broad spaces, each with its case for being looked at.

Trust was driven through transparency.

Each response from the twins was tracked back to the source data, allowing us to track if there were any contradictions in the twin responses compared to their original data; these contradictions were also treated as insight rather than quality metric.

### *Step 4: Consumer Voices*

After the initial dig sites emerged from all the internal wisdom and past work, reasoned forward, we had very clear dig sites to go deeper into with consumers. The rationale for this step was clear: we needed to tease these ideas out with real people, based on their taste preferences, experiences, individual contexts and creative ideation.

Once we arrived at the generated list of ideas – around 50, built from the initial dig sites – we put these back into the model to test, validated against past wisdom and iterated to refine further.

The entire exercise merged strategic judgment with real participant creativity, a way to activate all the internal wisdom and the research archives, together woven to create the final list of ideas - All amplified through AI, but guided by a strategist at the helm.

### *What Emerged*

We ended up with the final set of ideas that had been fully validated against the brand strengths & equity, category, dynamics, trends, cultural movements, strategic goals and ambitions, along with consumer appetite and needs.

This gave us a very confident understanding of the potential impact of these ideas.

The layers built on each other. The archive surfaced the spaces; the strategist's judgment sharpened them; the twins filled them with concepts; real consumers teased out those concepts; the strategist made the final call. The effort went into extending what Kraft already knew. Work that would normally run for months ran in weeks, but not by cutting corners; by starting from what already existed.

And it did not end when the project closed. Because it was built as connected, living intelligence rather than a finished document, it stays active till date: no longer the record of one project but the start of a model the brand can keep questioning. It keeps compounding.

## 5. Implications for Research Industry and Clients

### *What This Asks of Us*

Three things follow from this case that every research leader should sit with.

- Treat your data as a living asset, not a record. Most insights teams do not have a data shortage. They have an activation shortage. Before commissioning the next study, the first question is what the existing archive – properly activated – could already answer. Most of the time, more than you think. New research does not disappear; it becomes sharper, leaner and more powerful when it adds to an activated archive rather than starting every conversation from scratch.
- Build twins around real people. If you use digital twins at all, the question to interrogate first is whose reality they are built from. A twin drawn from a generic population reasons like no one in particular – generic average. A twin drawn from your actual target audience, with their language, tensions and behaviour loaded in, will truly reflect your audience in a way that you can trust.
- The advantage has moved to activation. Every major brand has a research archive. Most are filing cabinets. The advantage is moving to whoever turns theirs into a living, queryable asset first, by feeding it continuously, combining it with fresh inputs and acting on it with conviction. The sourdough principle is not really about AI. It is about institutional knowledge, and a provocation for every insights leader to look at what is already in the fridge before ordering takeout.

## *A Few Principles to Carry Forward*

- Your archive is a fermenting asset, not a filing cabinet. Every enterprise with more than two years of primary research is sitting on a compounding intelligence resource. The barrier is not technology; it is the cultural assumption that old research is dead research. Polanyi (1966) made the point that we know more than we can tell: much of what an organisation understands never reaches a document. An archive holds the record; the judgment sits with the people who built it.
- Fresh primary research and the archive feed each other. Pure archive activation risks confirming stale assumptions; fresh research without archive context wastes time rediscovering known territory. The work is in feeding the starter.
- Speed-with-conviction beats thoroughness-without-direction. The three-week timeline kept strategic momentum high and prevented the analysis paralysis that sinks longer innovation programmes.
- The method is transferable today. Not a bespoke solution for one CPG brand. Any organisation with a research archive, a strategic question and the willingness to activate accumulated knowledge can apply it, from CPG to mid-market fintech to healthcare.

The industry has spent its history treating knowledge as perishable – used once, then thrown out for something fresher. What is changing is that knowledge no longer has to spoil. Kept alive and fed, it compounds. Reasoned forward, it starts to think alongside us instead of sitting behind us. That is a real gain, and it is worth being clear-eyed about the cost.

A model only knows the world it has been shown. The part of this work that cannot be automated – the part that decides whether the discipline gets wiser or just more confident – is the strategist at the helm, guiding the direction and the outcomes.

The starter was always there. We now have a way to keep it alive – and the right expert can bake it just right!

Copyright © **ESOMAR** 2026


## References

Argyle, L. P., Busby, E. C., Fulda, N., Gubler, J. R., Rytting, C., & Wingate, D. (2023). Out of One, Many: Using Language Models to Simulate Human Samples. *Political Analysis, 31*(3): 337–351. https://www.cambridge.org/core/journals/political-analysis/article/abs/out-of-one-many-using-language-models-to-simulate-human-samples/035D7C8A55B237942FB6DBAD7CAA4E49

Duncan, D. (2026). *How Do Workers Develop Good Judgment in the AI Era?* Harvard Business Review. https://hbr.org/2026/02/how-do-workers-develop-good-judgment-in-the-ai-era

Forrester Research. (2016). *Hadoop Is Data's Darling for a Reason*. https://www.forrester.com/blogs/hadoop-is-datas-darling-for-a-reason/

Gartner. *Dark Data*. Gartner IT Glossary. https://www.gartner.com/en/information-technology/glossary/dark-data

Levy, S. (1984). *A Spreadsheet Way of Knowledge*. Harper's Magazine (repr. Wired, 2014). https://www.wired.com/2014/10/a-spreadsheet-way-of-knowledge/

Polanyi, M. (1966). The Tacit Dimension. University of Chicago Press.https://press.uchicago.edu/ucp/books/book/chicago/T/bo6035368.html

Sisense. (2025). *2025 State of Analytics*. https://www.sisense.com/blog/from-digital-friction-to-embedded-insights/

STRAT7 with Dunnhumby. (2025). *Putting Synthetic Data to the Test: A Real-World Evaluation*. STRAT7. https://strat7.com/ebooks/putting-synthetic-data-to-the-test/

## About the Authors

Sushma Panchawati, CEO, The Qlue, India.
Kelli Johnson, Co-Founder, Pepper Strategies, USA.
Kristina Lenz, Co-Founder, Pepper Strategies, USA.
Scott Peiser, Associate Director, Consumer Insights (Innovation), Kraft Heinz, USA.

esomar

# When AI Sees but Doesn't Understand

## Evidence From 50+ Parallel Tests of AI Versus Human Eye-Tracking

*Patrick Young*

*Olivier Blanchet*

Esomar

Office address:

Burgemeester Stramanweg 105

1101 AA, Amsterdam

Phone: +31 20 664 21 41

Email: info@Esomar.org

Website: www.Esomar.org

Publication Date: September 2026

Esomar Publication Series Volume S424 Congress 2026

ISBN 92-831-0337-8

Copyright © **Esomar** 2026

Copyright © **Esomar** 2026

Congress 2026

# When AI Sees but Doesn't Understand

## Evidence From 50+ Parallel Tests of AI Versus Human Eye-Tracking

*Patrick Young*

*Olivier Blanchet*

## Introduction

AI-powered research tools are increasingly becoming embedded across the insights ecosystem. From synthetic respondents and automated qualitative analysis to predictive attention modelling, the industry is experiencing a rapid expansion of technologies promising faster, cheaper and more scalable research. Packaging evaluation has become one of the clearest frontiers for this transformation. AI-led eye-tracking platforms promise to estimate what shoppers will notice on store shelves and on packaging without the need for respondent recruitment or behavioural testing. For insights teams operating under increasing commercial pressure, the appeal is obvious: rapid turnaround times, lower costs and scalable evaluation. At the same time, packaging itself is becoming increasingly strategically important. Rising raw material costs continue to squeeze manufacturer margins, while cost-of-living pressures have increased retailer resistance to price increases. Simultaneously, the fragmentation of the media landscape has elevated retail media and packaging itself as increasingly important communication channels. Packaging is often the first place where shoppers encounter new products, variant launches, sustainability claims or promotional messaging.

In this context, understanding how shoppers visually process packaging is not simply a design question, it is a commercial one. However, despite growing adoption of AI-led attention modelling tools, there remains limited published evidence comparing AI predictions against real human visual behaviour in commercial FMCG environments. This creates a significant challenge for the insights industry. Moderate correlation between AI predictions and human behaviour may appear reassuring, but if divergence occurs in commercially critical moments, such as pack hierarchy, comprehension or brand recognition, then the resulting decision risk may be substantial. The purpose of this paper is therefore threefold:

- To evaluate the extent to which AI-led attention modelling aligns with shopper-led eye-tracking;
- To identify where and why divergence occurs;
- To propose a practical framework for responsible integration of AI attention modelling into packaging research.

## Background: AI and The Rise of Automated Attention Modelling

The emergence of AI-led eye-tracking tools reflects a broader shift toward automation across the research industry. In recent years, research has experienced increasing pressure to deliver faster and cheaper outputs. This pressure has accelerated investment into automation technologies capable of reducing human intervention in data collection, coding, analysis and reporting. Attention modelling tools sit within this wider trend. Most AI eye-tracking systems operate using computational salience modelling. These systems are typically trained on visual attention datasets and use variables such as colour contrast, size, orientation, texture, positioning and visual prominence. The underlying logic is largely derived from Gestalt theories of visual attention and salience. In principle, this creates a compelling proposition for packaging research. If visual attention can be predicted computationally, brands may no longer need to rely on slower and more expensive behavioural methods. However, a critical question remains unresolved: Does human visual processing operate primarily through salience, or does cognition fundamentally alter what shoppers actually attend to?

Packaging is not viewed in a cognitive vacuum. Shoppers approach shelves with a number of competing thoughts and attitudes: brand familiarity, category expectations, shopping mission, memory structures, habitual scanning behaviours, as well as existing associations. This distinction may

appear subtle, but it has major implications. A packaging design that appears visually salient to an algorithm may not necessarily align with how shoppers process meaning, hierarchy or trust cues in real shopping environments. This paper explores this distinction empirically.

### *Literature Review and Theoretical Foundations*

Visual attention has long been recognised as a critical precursor to information processing and decision-making in marketing contexts. Eye-tracking research has consistently demonstrated that what consumers notice influences subsequent comprehension, evaluation and choice (Wedel & Pieters, 2008). Consequently, understanding how attention is allocated across packaging and shelf environments has become a central concern for marketers and researchers alike. Traditional eye-tracking methods measure actual human visual behaviour, providing an objective record of where attention is directed and in what sequence information is processed. These methods have been widely used to evaluate advertising effectiveness, shelf stand-out, packaging hierarchy and shopper navigation (Pieters & Wedel, 2004; Wedel & Pieters, 2008).

### *The Rise of Computational Attention Models*

The emergence of AI-powered attention modelling tools is rooted in computational theories of visual salience. One of the most influential models was developed by Itti, Koch and Niebur (1998), who proposed that attention is driven by contrasts in colour, intensity, orientation and other low-level visual features. This work forms the foundation of many contemporary salience-based prediction systems. Subsequent developments in visual attention modelling have improved the sophistication of these approaches, enabling machine-learning systems to predict likely fixation patterns without requiring direct observation of human behaviour (Borji & Itti, 2013). For packaging research, the appeal is obvious. If attention can be predicted computationally, businesses can potentially evaluate designs more quickly, more cheaply and at greater scale than traditional behavioural approaches. However, a key question remains unresolved: does visual salience, in the form of colour, intensity and orientation, fully explain how shoppers process and understand packaging?

### *Attention Beyond Salience*

Research from psychology and consumer behaviour suggests that attention is not determined solely by visual prominence. Consumers rely extensively on memory structures, heuristics and prior knowledge when navigating choice environments. Brands often act as cognitive shortcuts, allowing shoppers to simplify decisions and reduce cognitive effort (Nedungadi, 1990; Gigerenzer & Gaissmaier, 2011). Kahneman's (2011) distinction between fast, intuitive processing and slower, deliberative thinking further highlights that attention is influenced not only by what stands out visually, but also by what is meaningful, familiar or relevant to the task at hand. This distinction is particularly relevant in FMCG packaging. Shoppers do not approach shelves as passive observers; they bring category expectations, brand familiarity, shopping goals and previous experience to the task of visual processing. Consequently, what captures attention may differ substantially from what is visually salient. The present study investigates this distinction directly by comparing AI-led attention modelling with shopper-led eye-tracking across more than 75 packaging evaluations.

## Methods

### *Study Design*

To evaluate the reliability of AI attention modelling, a structured parallel testing programme was conducted across 77 FMCG packaging stimuli. The study compared:

- AI-led eye-tracking using Dragonfly AI—a leading AI tool used by agency & client-side researchers;
- Shopper-led eye-tracking;
- Both shelf-level and pack-level evaluations were included.

The stimuli covered multiple FMCG categories, including alcohol, bottled water, confectionery, personal care, meal kits and tea. Each stimulus consisted of:

- High-resolution pack imagery;

Copyright © **Esomar** 2026


- Competitive shelf imagery.

The objective was to replicate realistic packaging and shelf decision environments. Three primary metrics were compared.

1. Shelf stand-out

This measured the extent to which packs visually emerged within competitive shelf environments. Comparison metrics were as follows:

Human eye-tracking: percent noticed;

AI eye-tracking: probability of perception (POP).

2. Pack element visibility

This measured whether individual elements on pack were visually attended to. Examples included:

- Brand name
- Product descriptor
- Benefit claims
- Imagery
- Variant cues
- Comparison metrics were as follows:
- Human eye-tracking: percent noticed;
- AI eye-tracking: probability of perception (POP).

3. Eye-flow / hierarchy

This measured the sequence in which shoppers visually processed key elements on pack. The first four fixations were analysed as these are particularly important in determining:

- Pack comprehension;
- Hierarchy effectiveness;
- Communication sequence;
- Analysis approach.

The analysis evaluated:

- Correlation between AI and shopper results;
- Absolute score bias;
- Variance across stimuli;
- Divergence in hierarchy interpretation;
- Potential impact on commercial recommendation.

The objective was not simply to assess statistical alignment, but to understand whether divergence would materially alter interpretation and decision-making.

## Results

### *Shelf Stand-Out*

Shelf stand-out showed moderate correlation between AI and shopper-led eye-tracking. The overall correlation measured R = 0.68.

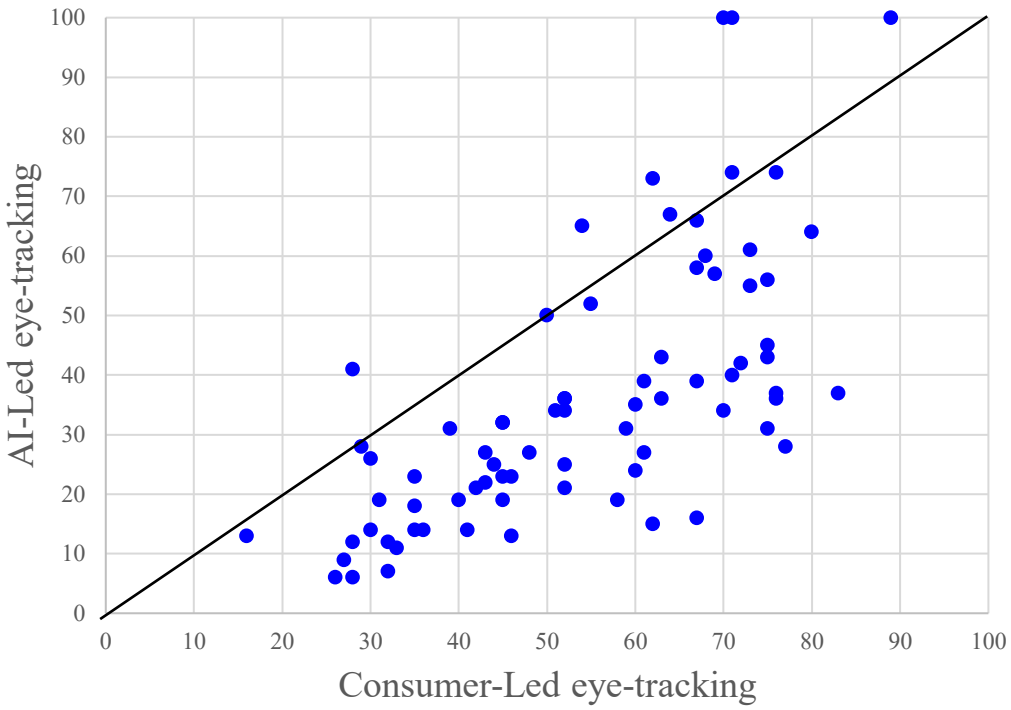



*Figure 1. Correlation of AI-led eye-tracking results and consumer-led eye-tracking results for shelf stand out.*

At first glance, this appears relatively reassuring (see Figure 1 and Appendix I for a comparison of AI-led and consumer-led shelf eye-tracking heat maps). However, a more detailed examination revealed a consistent systematic bias (see Figure 2). AI-led attention modelling consistently under-reported visibility compared with shopper-led testing, with an average score compression of approximately 18 percentage points.

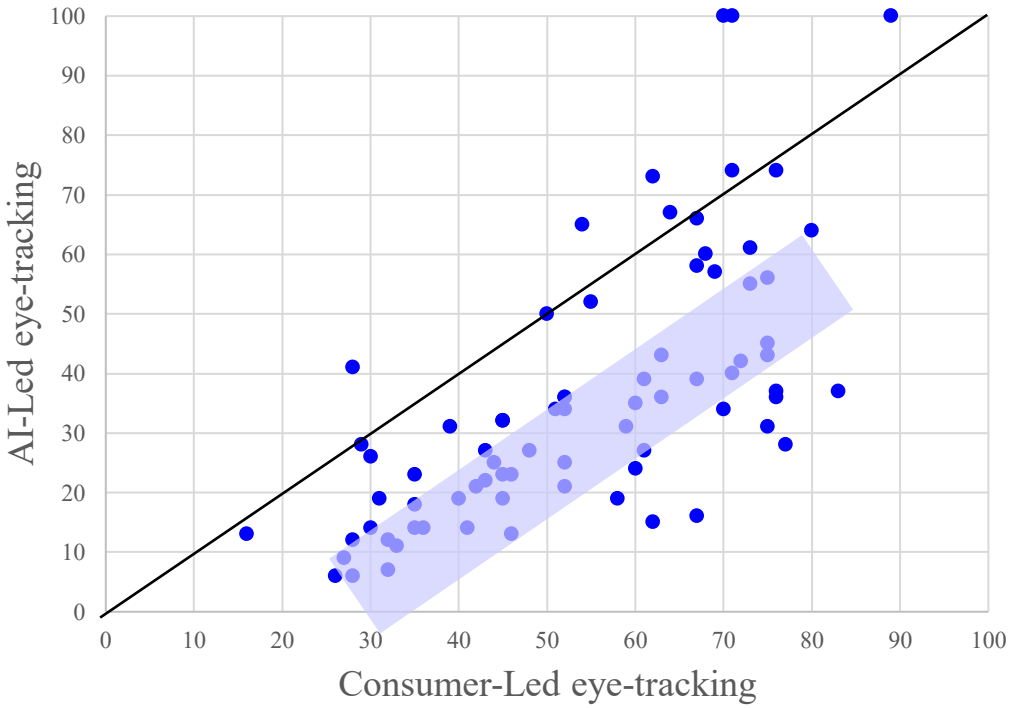



*Figure 2. Highlight of systematic under-reporting of visibility by AI-led eye-tracking.*

This compression effect is of importance because packaging performance is typically evaluated against normative benchmarks. Systematic underestimation of visibility can therefore:

- Misclassify strong packs as underperforming;
- Obscure meaningful competitive advantage;
- Reduce sensitivity between competing designs.

Despite this limitation, hierarchy between shelf stimuli was often directionally consistent. This suggests that AI attention modelling may have practical value in high-contrast shelf evaluation scenarios, particularly when calibrated against appropriate norms.

### *Package Element Visibility*

Package element visibility also demonstrated moderate overall correlation: R = 0.69. However, unlike shelf stand-out, variance across individual stimuli increased substantially. While some packs showed relatively strong alignment between AI and shopper attention, others showed large gaps in visibility prediction. This inconsistency reduced confidence in relying on AI outputs for detailed package interpretation. The issue became even more pronounced when examining eye-flow.

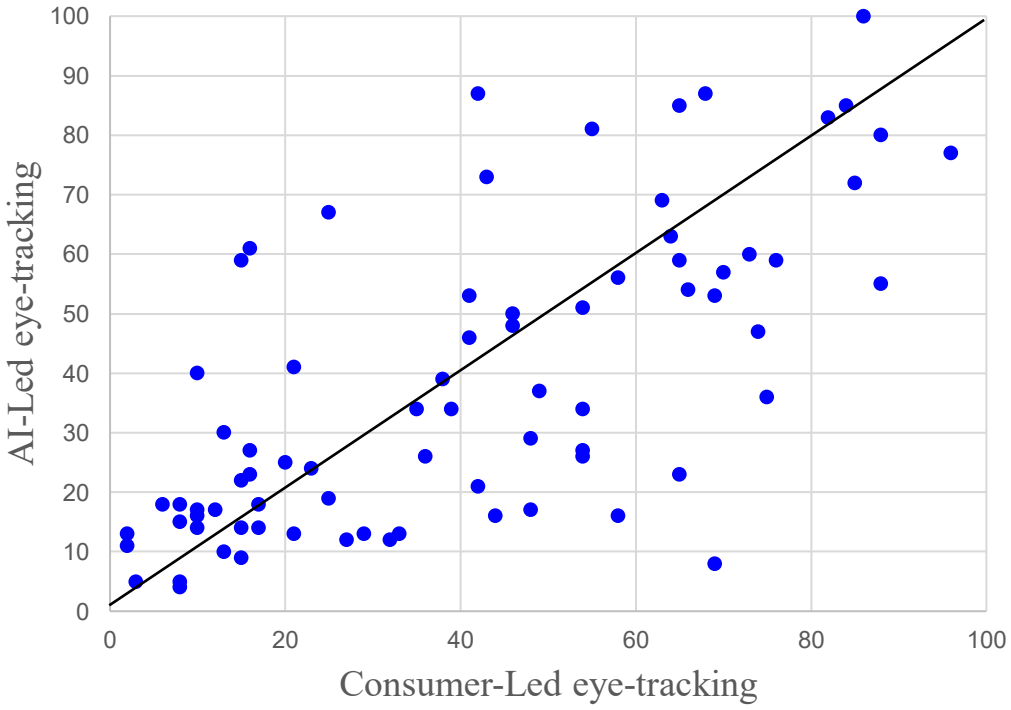



*Figure 3. Correlation of AI-led eye-tracking results and consumer-led eye-tracking results for package element visibility.*

### *Package Eye-Flow*

Eye-flow analysis produced the weakest alignment between AI and shopper behaviour. Correlation for the first four elements viewed was R = 0.55. This represented a significant reduction in predictive reliability.

Copyright © **Esomar** 2026



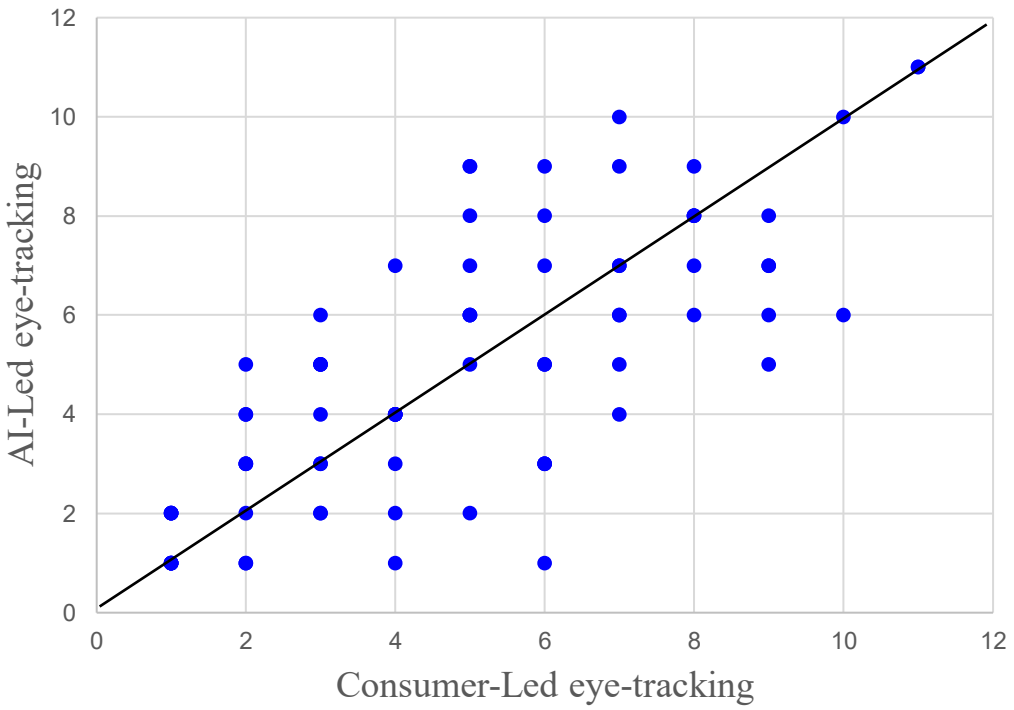



*Figure 4. Correlation of AI-led eye-tracking results and consumer-led eye-tracking results for package eye-flow.*

Importantly, divergence was not random. AI systems consistently prioritised:

- High-contrast claims;
- Large graphic elements;
- Salient visual features.
- Human shoppers, by contrast, frequently prioritised:
- Brand assets;
- Familiar category cues;
- Recognisable structures;
- Navigation signals.

This often led to materially different conclusions about whether packaging communicated effectively. Across the dataset, 67% of packaging evaluations would have resulted in different commercial recommendations if AI-led interpretation had been used instead of shopper-led eye-tracking. This finding is particularly significant because hierarchy and communication sequence are among the most commercially important dimensions of packaging performance. A packaging may achieve high visual salience while still failing to communicate the intended message structure.

*Figure 5. Comparison of eye-flow: Human-led eye-flow and AI-led eye-flow on Vittel bottle multipack.*

The example shown in Figure 5 highlights the stark divergence between shopper-led and AI-led eye-flow analysis. In the shopper-led eye-flow, the Vittel brand mark dominates initial attention, creating a natural top-to-bottom visual hierarchy. Attention is drawn first to the brand name, then to the "100% recyclable" claim, before moving to the provenance cue at the base of the packaging. In this interpretation, the packaging effectively communicates a clear and commercially intuitive sequence: brand first, supporting claim second. By contrast, the AI-led eye-flow produces a markedly different viewing sequence. Attention is initially drawn to the "100% recyclable" claim, before shifting to the brand mark and subsequently to the forest imagery. This creates a less intuitive right-to-left and bottom-up visual flow, implying that secondary visual assets dominate shopper attention more strongly than the core brand identity. The divergence between these outputs is commercially significant. Under the shopper-led interpretation, the design succeeds because the brand remains the primary navigational anchor while sustainability messaging acts as a supporting reinforcement. Under the AI-led interpretation, however, the sustainability claim appears to overpower the brand, potentially leading to recommendations to reduce claim prominence or rebalance hierarchy.

This example illustrates a broader pattern observed throughout the study: AI attention modelling appears highly sensitive to visual salience, while human shoppers process packaging through a combination of salience, familiarity, learned category structures and shopping intent (see Appendix II for an additional illustrative example). As a result, AI may correctly identify what is visually prominent, without accurately reflecting what is cognitively important.

## Discussion

### *The Difference Between Seeing and Understanding*

The findings suggest that AI attention modelling and human visual processing are not operating on identical principles. The findings of this study are broadly consistent with existing theories of visual attention while extending them into the emerging field of AI-led packaging evaluation. The relatively strong performance of AI attention modelling in shelf stand-out tasks supports the salience-based framework proposed by Itti Koch and Niebur (1998). Where visual contrast dominates attention allocation, computational prediction appears capable of approximating human behaviour. However, the substantial divergence observed in package hierarchy and eye-flow suggests that shoppers rely on more than salience alone. These findings align with research demonstrating the importance of memory structures, heuristics and prior knowledge in consumer decision-making (Nedungadi, 1990; Gigerenzer & Gaissmaier, 2011).

The implication is that AI systems may be highly effective at predicting what is visually prominent, but considerably less effective at predicting what is cognitively important. AI systems primarily model perceptual salience, whereas human shoppers integrate both perception and cognition. This critical distinction forms the basis of the Dual-Process Visibility Model proposed in this paper. Visual processing in FMCG environments is not merely a physiological exercise in noticing contrast. It is also shaped by:

- Brand memory structures;
- Prior shopping experience;
- Category expectations;
- Motivations;
- Meaning attribution.

For example, a shopper may visually prioritise a familiar brand logo despite lower visual contrast because the brand acts as a navigation shortcut within the category. Similarly, shoppers may ignore highly salient graphic elements if those elements are not relevant to their shopping goals. This helps explain why shelf stand-out performed more reliably than packaging hierarchy. Shelf stand-out is more heavily driven by perceptual salience. Packaging hierarchy increasingly involves cognition.

### *The Dual-Process Visibility Model*

To explain these findings, this paper proposes a Dual-Process Visibility Model, which distinguishes between two fundamentally different forms of visual processing in packaging research:

1. Perceptual processing: where attention is driven primarily by visual salience;
2. Cognitive processing: where attention is shaped by memory structures, intent, category understanding and meaning.

This distinction helps explain why AI attention modelling performs relatively well in some contexts while struggling in others. Figure 6 presents the Dual-Process Visibility Model as a 2x2 framework based on the interaction between visual salience and cognitive influence.

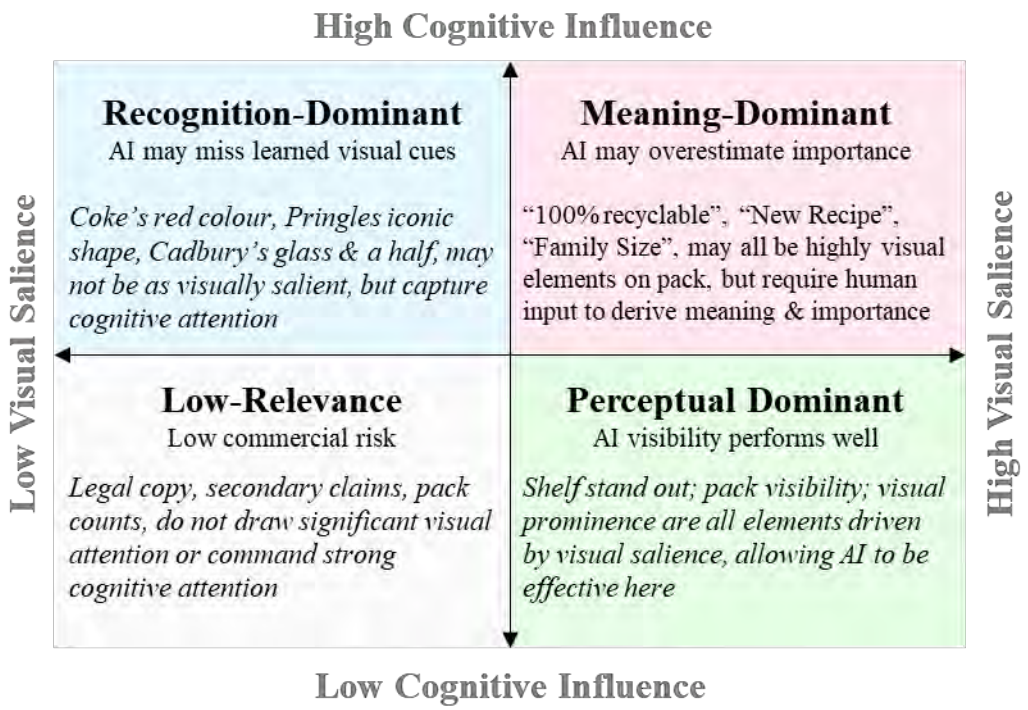



*Figure 6. Dual-Process Visibility Model of visual salience and cognitive influence.*

As attention shifts from perceptual processing toward recognition and meaning-based processing, AI attention modelling becomes a progressively less reliable proxy for shopper behaviour. Exploring this further, the model suggests that cognition influences attention in two

distinct ways. First, through recognition, where shoppers use familiar assets and learned category structures to navigate quickly. Second, through interpretation, where shoppers seek to understand the meaning and relevance of claims, benefits and messages. These represent different challenges for AI attention modelling and help explain the patterns observed in the study. We now explore the four quadrants of the Dual-Processing Visibility Model in more detail.

1. Perceptual-dominant quadrant

This quadrant comprises environments where visual attention is driven primarily by contrast and salience. Examples include:

- Shelf stand-out;
- Identifying packs in cluttered shelf environments;
- High-level visual emergence;
- Basic visual prominence testing.

In these contexts, AI attention modelling performs comparatively well because shoppers are relying heavily on physiological attention mechanisms, such as colour contrast, size, orientation and positioning. This aligns closely with the underlying architecture of most AI salience models. The shelf stand-out findings in this study sit largely within this quadrant. While systematic score compression existed, the relative hierarchy between shelf stimuli was generally directionally aligned between AI and shopper-led testing.

2. Meaning-dominant quadrant

This quadrant contains elements that are both visually prominent and cognitively rich. While AI systems can often identify these elements successfully, understanding their behavioural significance requires interpretation. Examples include:

- Sustainability claims;
- Promotional violators;
- Functional benefits;
- Nutritional claims;
- “Natural” or “healthy” design cues.

In these environments, AI systems may correctly identify visually prominent elements, but overestimate their importance within the overall communication hierarchy. For example, highly salient claims may attract visual attention computationally while being assigned less importance by shoppers who prioritise other information on packaging. As a result, AI may accurately predict visibility while misinterpreting relevance. Human interpretation remains essential to understand how these elements contribute to the overall meaning of the packaging.

3. Low-relevance quadrant

This quadrant contains low-salience and low-cognition elements. Examples include:

- Minor legal copy;
- Small secondary claims;
- Peripheral design details.

These elements are often relatively unimportant commercially and may receive limited attention from both AI systems and shoppers. This zone is comparatively low risk from a research perspective.

4. Recognition-dominant quadrant

This quadrant is the most commercially significant, and the area where the largest divergence between AI and shopper behaviour emerges. Examples include:

- Pack hierarchy;
- Pack navigation;
- Variant identification;
- Message sequencing;
- Premiumisation cues;

- Familiar category structures.

In these environments, shoppers rely heavily on learned behaviour, memory structures, category expectations and shopping intent. Attention is influenced not simply by what is visually prominent, but by what is familiar and useful for navigation. The eye-flow findings from this study sit primarily within this quadrant. For example, AI systems frequently prioritised highly salient claims or graphic elements, while shoppers instead prioritised familiar brand assets and category navigation structures.

### *Salience-Meaning Gap*

The findings suggest that AI attention modelling is highly effective at identifying signals, but less effective at interpreting meaning. Packaging is not simply a visual object; it is a communication system. Human shoppers do not merely respond to what is visually prominent, but to what is familiar, relevant and meaningful within the context of their shopping mission. As a result, visual prominence and communicative importance are not always the same thing. Put simply, AI attention models primarily evaluate the visual properties of a design: colour, size, contrast, positioning and prominence. Human shoppers, by contrast, interpret those visual cues through the lens of memory, category knowledge, brand familiarity and shopping intent. This distinction helps explain several of the findings observed in the study. In the Vittel example, the sustainability claim was visually more salient and therefore attracted greater attention in the AI model. Human shoppers, however, were drawn first to the Vittel brand mark, creating a markedly different eye-flow. This suggests that cognition does not merely shape the interpretation of packaging after attention has occurred; it helps shape attention itself. Through repeated category exposure and accumulated brand memories, shoppers learn to recognise and prioritise familiar assets that act as navigation shortcuts within the category. In this context, what attracts attention is determined not only by visual prominence, but also by recognition, relevance and meaning.

This often resulted in materially different conclusions about whether packaging communicated effectively. The implication is not that AI is ineffective. Rather, AI appears most reliable in perceptual-dominant environments and progressively less reliable as attention becomes increasingly influenced by recognition and meaning. As cognitive influence increases, human validation becomes increasingly important to ensure that visibility is interpreted within the context of shopper behaviour. This creates what might be described as a “salience–meaning gap”; i.e., AI systems appear highly capable of predicting what is likely to attract attention, but less capable of predicting what shoppers will interpret as meaningful. The larger this gap becomes, the greater the risk that visibility metrics alone may lead researchers to conclusions that differ from actual shopper behaviour.

### *Limitations*

Several limitations should be acknowledged. First, the analysis focused on FMCG packaging and may not generalise to other visual domains such as advertising, e-commerce or retail environments. Second, the study evaluated a single AI attention modelling platform. While the findings likely have broader relevance to salience-based systems, results may differ across alternative approaches. Finally, the analysis focused on attention rather than downstream behavioural outcomes such as purchase or sales impact. Future research should explore whether similar patterns emerge across other AI research applications and whether cognition-dominant tasks consistently produce greater divergence between machine prediction and human behaviour.

### *Implications for the Insights Industry*

The findings have important implications as to how AI should be integrated into packaging research. The conclusion of this paper is not that AI attention modelling should be rejected. AI clearly offers meaningful benefits including speed, scalability, rapid iteration and the option for early-stage screening. However, the findings suggest that AI should not currently be treated as a wholesale replacement for behavioural research in commercially critical packaging decisions. Moderate correlation alone may create false confidence. Even where average statistical alignment appears acceptable, divergence in hierarchy and interpretation can materially alter commercial recommendations. This suggests the need for

stronger governance around AI integration. The future of insights is unlikely to be fully automated. Instead, the findings support an augmentation model in which human and machine capabilities are combined. This perspective aligns with broader thinking around collaborative intelligence, where AI systems enhance rather than replace human expertise (Wilson & Daugherty, 2018). The objective is not to automate judgement, but to improve it.

The findings also reflect the broader principle described by Moravec (1988): tasks that appear effortless to humans often prove difficult for machines, while tasks that are computationally complex may be comparatively straightforward to automate. In the context of packaging research, AI appears well suited to salience-driven tasks, but struggles where meaning, familiarity and context shape attention. AI can accelerate screening, iteration, processing and pattern detection. Human researchers remain critical for interpretation, contextual understanding, strategic judgement and recommendation. The greatest value therefore lies not in replacing researchers, but in augmenting them.

## Conclusion

This paper evaluated more than 75 parallel tests comparing AI-led attention modelling with shopper-led eye-tracking in FMCG packaging contexts. While visibility metrics showed moderate overall correlation between AI and shopper behaviour, substantial divergence emerged in hierarchy and eye-flow measures. These divergences were not merely technical differences. In 67% of cases, they would have resulted in materially different commercial recommendations. The findings suggest that AI attention modelling primarily captures perceptual salience, while human shoppers integrate cognition, memory structures and category understanding into visual processing. To explain this distinction, the paper introduced a Dual-Process Visibility Model separating perceptual-dominant from cognition-dominant tasks. The practical implication is clear: AI attention modelling can provide meaningful value in specific contexts, particularly for rapid screening and perceptual-dominant tasks such as shelf stand-out. However, commercially important packaging decisions continue to benefit from human validation, particularly where recognition, meaning and interpretation play a significant role.

The future of packaging research is therefore unlikely to be defined by replacement. Instead, the greatest value is likely to come from augmentation: combining the speed and scalability of AI with the contextual understanding, judgement and strategic interpretation provided by human researchers.

## References

Borji, A., and Itti, L. (2013). State-of-the-art in visual attention modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 35*(1), 185–207.

Gigerenzer, G., and Gaissmaier, W. (2011). Heuristic decision making. *Annual Review of Psychology, 62*, 451–482.

Itti, L., Koch, C., and Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 20*(11), 1254–1259.

Kahneman, D. (2011). *Thinking, Fast and Slow*. Penguin.

Moravec, H. (1988). *Mind Children: The Future of Robot and Human Intelligence*. Harvard University Press.

Nedungadi, P. (1990). Recall and consumer consideration sets: Influencing choice without altering brand evaluations. *Journal of Consumer Research, 17*(3), 263–276.

Pieters, R., and Wedel, M. (2004). Attention capture and transfer in advertising: Brand, pictorial, and text-size effects. *Journal of Marketing, 68*(2), 36–50.

Wedel, M., and Pieters, R. (2008). A review of eye-tracking research in marketing. *Review of Marketing Research, 4*, 123–147.

Wilson, H.J., and Daugherty, P.R. (2018). Collaborative intelligence: Humans and AI are joining forces. *Harvard Business Review, 96*(4), 114–123.

## About the Authors

Copyright © **Esomar** 2026

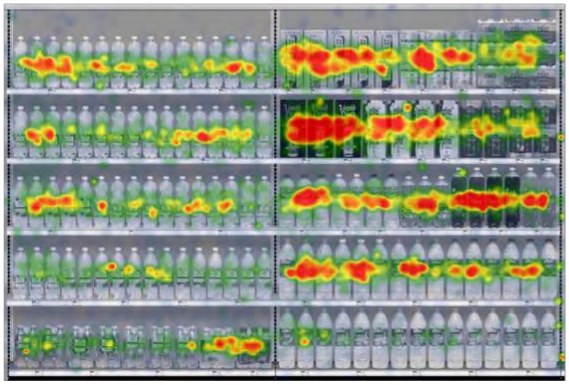
Congress 2026

Patrick Young, Managing Director, PRS IN VIVO, Ipsos, UK.

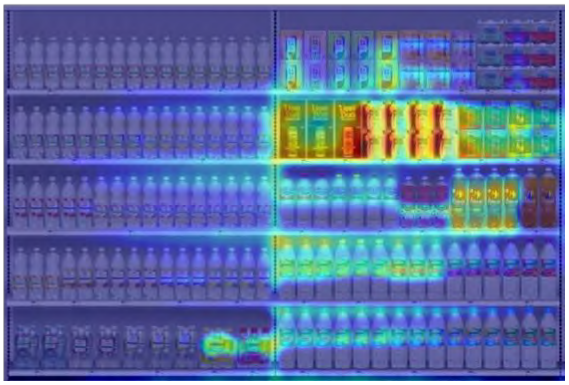
Olivier Blanchet, General Manager, PRS IN VIVO, Ipsos, France.

## Appendix I: Comparison of Human-Led and AI-Led Shelf Stand Out

Consumer-Led



AI-Led



## Appendix II: Comparison of Human-Led and AI-Led Pack Stand Out

Consumer-Led

AI-Led

# Smart Labels, Real Homes, True Usage

## Closing the Say-Do Gap in Product Testing with Passive Observations

*Nirmit Behal*

*Karolina Dobrzyńska*

*Bert Nijhuis*

*Susan Purcell*

*Omkar Shinde*

*Pratik Rane*

*Shreeraj Dhekane*

*Jinho Cho*

*Nikolai Reynolds*

Esomar

Office address:

Burgemeester Stramanweg 105

1101 AA, Amsterdam

Phone: +31 20 664 21 41

Email: info@Esomar.org

Website: www.Esomar.org

Publication Date: September 2026

Esomar Publication Series Volume S424 Congress 2026

ISBN 92-831-0337-8

Copyright © **Esomar** 2026

Copyright © **Esomar** 2026

Congress 2026

# Smart Labels, Real Homes, True Usage

## Closing the Say-Do Gap in Product Testing with Passive Observations

*Nirmit Behal*

*Karolina Dobrzyńska*

*Bert Nijhuis*

*Susan Purcell*

*Omkar Shinde*

*Pratik Rane*

*Shreeraj Dhekane*

*Jinho Cho*

*Nikolai Reynolds*

### Introduction: The Problem with Asking People to Remember

For consumer-goods companies, understanding how people really use products at home is essential. Yet, many product tests still depend on surveys and diaries. These methods are valuable, but they rely on people remembering small details from routines that receive very little conscious attention. Laundry is a good example. A person may claim to use half a cap of fabric conditioner every time, but the amount actually poured can change from one wash to the next. The bottle may be full or nearly empty. The washing-machine drawer may be awkward to reach. One hand may hold the cap while the other steadies the drawer. The participant may also be rushing or handling laundry at the same time. The say-do gap is therefore not evidence that consumers are dishonest. It is evidence that intention, memory and physical execution are different things. Unilever and Ipsos studied these differences by combining connected smart labels, online diaries, product ratings and videos analysed with product view AI.

*Figure 1. Consumers filmed themselves and consent was given by the consumers for displaying their faces.*

One of the key insights of this study was that consumers reported an average dose almost identical to the amount measured by smart labels. At first glance, traditional methods appeared accurate. However, this concealed substantial variation at the level of individual wash occasions. Only around 30% of claimed doses were within 20% of the measured amount. The average was right; the behaviour was not. The connected data showed that over- and under-dosing cancelled each other out. The top line looked correct, while the underlying behaviour was highly inconsistent.

| 1. Memory gap | 2. Behaviour gap | 3. Experience-attribution gap |
|---|---|---|

Copyright © **Esomar** 2026


| | | |
|---|---|---|
| **What consumers remember or report versus what happened. A familiar answer such as "half a cap" simplifies a routine that is difficult to recall precisely.** | What consumers intend to do versus what they physically do. The same honest diary answer can be followed by a different pour on each occasion. | What actually created satisfaction versus what consumers believe created it. A stronger experience may come partly from an unnoticed larger dose, but be credited to the product itself. |

*Table 1. The three gaps uncovered by the study.*

## A More Realistic Model of Product Experience

Traditional product testing often assumes a short and direct pathway:

**PRODUCT → USE → RATING**

The study showed a more detailed real-world pathway:

**PRODUCT → CONTEXT → BEHAVIOUR → ACTUAL DOSE → EXPERIENCE → RATING**

The package does not create the rating directly. The package is handled in a specific environment, under particular time and physical constraints. This shapes behaviour, behaviour creates the actual dose, the dose helps create the experience and the experience shapes the final rating. The connected method made this missing middle visible.

*Figure 2. Smart-label and a scale both measured the amount used.*

## How the Smart-Label with Physical AI Worked

Smart labels supplied by Adrich were attached to the product packs and connected to a mobile application and cloud system. Once installed, the labels recorded product-use events without requiring participants to enter each event manually. The system captured how much product was used, when and how often it was used, intervals between uses, duration, temperature and location signals, pack depletion and changes in dosing over the life of the pack. The key advantage was an occasion-level record. Researchers could compare what a participant reported in the diary with what physically happened during the same wash.

## How Product View AI Added Context

Smart labels showed what happened and when. Videos helped explain how and why it happened. Participants recorded opening the pack, dosing the product and unloading the laundry. Product view AI used six specialist perspectives: ethnography, product development, semiotics, anthropology, neuroscience and behavioural science. The observations were organised into nine dimensions: space, actor, activity, object, act, event, time, goal and feeling. Across 177 usable video records, the analysis focused on visible actions and context, such as whether the cap was

used, how the bottle was held, whether one hand steadied the drawer, whether liquid remained in the cap and whether washed fabric was brought close to the face for a sensory check.

*Figure 3. After the wash consumers tend to subconsciously smell the fragrance.*

## The UK Study

The UK study lasted 14 days and included 125 households across multiple product-use conditions. Traditional methods were run in parallel. Participants weighed products before and after dosing and completed a short online diary for each wash. The final linked dataset contained 272 usable product-use occasions from 79 respondents. Bases differ by analysis because not every occasion had complete diary, sensor and rating data.

## Results: What People Said and What They Did

### *People Were More Consistent in Intention Than in Action*

Actual dosing varied about three times more than the dose participants believed they were using. The diary captured a stable intention, but the smart label captured a much less stable physical action. A participant can provide the same honest diary answer on several occasions while pouring a different physical amount each time. This is the central behaviour gap. The remembered rule can stay the same while bottle weight, grip, pour angle, lighting, available hands and surrounding tasks change from wash to wash.

### *Consumers Were Not Dosing by Volume Alone*

The findings suggest that consumers did not behave like people measuring a laboratory volume. Instead, the dose emerged from a combination of cues and constraints:

- Visual cues: the stream, the apparent cap level and the visible level in the washing-machine drawer;
- Effort: the weight of the bottle, the need to carry the cap and the effort required to drain remaining liquid;
- Context: the shape of the laundry space, counter access, glare, clutter and the position of the machine;
- Time pressure: dosing was a brief step inside a larger task, not a separate measuring activity;
- Pack fullness: the balance and flow of the bottle changed as the pack became lighter;
- Habit: familiar expressions such as “half a cap” guided action even when the physical amount differed.

This explains why very different verbal cap descriptions often led to doses in a narrower practical range. Consumers were using an embodied “enough for this wash” judgement, not a precise volumetric calculation.

### *Cap Descriptions Simplified a More Complex Action*

Participants who described using small cap fractions generally poured considerably more than those fractions implied. In contrast, participants who described using a full cap generally poured somewhat less than a nominally full cap. The videos showed why. Participants often judged the stream by sight, poured at an angle, worked under a counter or in glare and stopped when the amount looked sufficient or safe to carry. “Half” and “full” often worked as labels for a routine rather than exact liquid levels.

### *Direct Pouring Made the Task Easier but Increased Usage*

Participants who poured directly from the bottle used considerably more product than participants who measured with the cap. Video evidence showed that direct pouring was a practical workflow choice. It removed the need to fill, carry, empty and clean the cap, and it worked better when only one hand was free. The trade-off was lower control: the amount depended on bottle angle, stream duration and judgement of the drawer level. Dosing style was flexible rather than fixed. Cap-based dosing appeared for most video respondents, direct pouring appeared for a substantial minority and several respondents used both methods across different occasions. People adapted the method to the load, space, time pressure and need for control.

### *The Cap Created Several Small Extra Tasks*

The cap acted as a closure, measuring tool, carrier and residue holder. Participants placed it on counters, drained it over the drawer, tapped it, rinsed it or continued holding it while managing laundry. These actions were rarely serious enough to appear in a questionnaire. However, each extra step created another opportunity for incomplete transfer, spills or variation. The absence of a reported problem did not mean the absence of friction.

### *Dosing Changed as the Pack Became Empty*

Participants used less product toward the end of the pack than at the beginning. Three mechanisms may have contributed: the balance and flow of a lighter bottle, learning to control an unfamiliar pack and a desire to make the remaining product last longer. Traditional testing often treats dose as a stable consumer habit. The passive data showed that dose can evolve during the pack journey.

### *Unnoticed Extra Dose was Linked with a Stronger Experience*

When the measured dose was higher than the amount implied by the participant’s cap claim, overall and fragrance ratings were consistently higher. This is an association and does not prove that the larger dose caused the better result. However, the result points to an experience-attribution gap. The product was dosed quickly at the machine, but judged later during unloading. By then, the earlier dose was no longer visible. A consumer may experience a stronger scent or fabric effect and credit the product, without realising that more product was used than intended. The videos supported this interpretation. Sensory-check behaviour appeared across multiple respondents, including bringing towels or fabrics close to the face and deliberately inspecting washed items.

## What Traditional Product Testing Would Have Missed

The combined method did more than make one measure more accurate. It changed the diagnosis of the product experience (see Table 2).

| Traditional result | What connected observation revealed | Why it matters |
|---|---|---|
| **The average dose is accurate.** | Large occasion-level errors cancel each other out. | Do not validate behavioural accuracy from the average alone. |

| | | |
|---|---|---|
| **A participant consistently says "half a cap".** | The same answer can lead to a different physical pour. | Treat diary answers as intention and memory, not exact measurement. |
| **A lower rating suggests a formula issue.** | Pack interaction or actual dose may be reducing performance. | Diagnose formula, package and behaviour together. |
| **No usability issue is reported.** | Handling, residue and one-handed compromises occur without complaint. | Observe the task, not only the final opinion. |
| **Dose is a stable habit.** | Dose changes with context, time pressure and pack fullness. | Study the full pack journey. |
| **A stronger rating reflects the formula.** | Unnoticed extra dose may partly create the stronger experience. | Separate product efficacy from the amount actually delivered. |

*Table 2. The connected observation added significant value to traditional survey means.*

## Business Impact: Better Diagnosis, Better Decisions

### *How the Finding Changes the Business Decision*

In a traditional product test, lower fragrance ratings could lead directly to reformulation work. This study shows why that decision may be too simple. Some of the variation in perceived performance may be linked to the amount of product actually delivered, rather than to the formula alone. The connected evidence allows teams to unpeel the result before deciding what to change. Is the issue the formula, the package, the dosing method, the clarity of the guidance or the context in which the product is used? The answer may lead to a different action. Instead of automatically reformulating, the team might improve flow control, dosing cues, cap design or pack handling. Reformulation remains an option, but only when the evidence points to the formula. The business challenge therefore shifts from improving formula performance in isolation to optimising the complete product-package-behaviour-experience system. The findings illustrate the types of diagnostic questions that connected observation can help teams answer:

- Usage execution: Is the product being used in the way consumers believe it is being used?
- Pack interaction: Are handling, pouring, residue or storage behaviours influencing the amount actually delivered?
- Dose interpretation: Are claimed measures such as "half a cap" being used as precise measures, or as simple language for a routine?
- Performance diagnosis: Could lower or higher ratings be partly explained by actual dose or usage context, rather than the product formulation alone?
- Pack journey: Does usage change as the pack becomes lighter, more familiar or closer to depletion?
- Claims and communication learning: Can measured consumption help teams better understand how consumers interpret usage guidance and expected results?
- Research efficiency: Can AI help organise and prioritise evidence, while retaining human review for interpretation?

### *Participant Privacy and Research Experience*

Passive observation creates detailed data, so privacy must be designed into the method. Participants were informed about the smart-label measurement and video tasks before taking part. Video recording was limited to specific product-use moments rather than continuous monitoring. Reporting focused on combined patterns or clips covered by the appropriate permission. The study should collect only the information needed, use secure storage and transfer, apply clear retention and deletion periods, restrict access and provide a straightforward way to withdraw data. Separate permission should be obtained when a video clip may be shown at a conference or in a client presentation. The smart label reduced the need to remember every dose and usage time. Diaries, weighing and video uploads still required effort. Future studies should measure respondent experience directly, including technical support needs, smartphone compatibility, upload effort and whether repeated recording changes behaviour.

Copyright © **Esomar** 2026


### *Beyond Laundry: A Category-Agnostic Framework*

The significance of these findings extends beyond laundry. Many FMCG categories depend on habitual, low-attention behaviours that consumers find difficult to recall precisely. People may estimate how much they dispense, how often they apply a product, how they prepare it or which steps they follow. When product performance depends on these physical actions, the same gaps can appear. The framework developed in this study can therefore be used across categories. Researchers should separate three questions: What does the consumer remember or report? What did the consumer physically do? What does the consumer believe created the final experience? These questions reveal the memory gap, the behaviour gap and the experience-attribution gap. The product and behaviour will differ by category, but the principle remains the same. Whenever performance depends on dose, frequency, preparation, sequence, handling, or application technique, connected observation can reveal the missing steps between product and rating.

### *The Future of Product Testing*

Connected measurement costs more and requires technical setup. These costs should be compared with the cost of solving the wrong problem. An unnecessary formula change, pack redesign or failed launch can be far more expensive than better diagnosis during testing. The greatest value comes from combining methods. The diary captures memory and intention. The smart label captures physical behaviour. Product view AI explains context and action. Product ratings capture the final experience. Together, these sources reveal the full pathway from package to context, behaviour, actual dose, experience and rating.

## Conclusion: Human and Machine are Stronger Together

The main lesson is not that consumers are unreliable. Consumer reports and physical actions answer different questions. A person can honestly report the same intended dose while pouring a different amount from wash to wash. Traditional product testing would have captured the intended dose and the final rating. It would often have missed the changing context and action between them. Smart labels and video observation made that missing middle visible. The future of product testing is therefore not passive technology alone. It is a combination of consumer voice, human interpretation, connected measurement and AI-assisted observation. Together, these methods show what people remember, what they intend, what they physically do and what they ultimately experience.

## About the Authors

Nirmit Behal, Consumer Insights Manager, Unilever, India.

Karolina Dobrzyńska, Consumer Technical Insights Manager, Unilever, UK.

Bert Nijhuis, Director Consumer Insights, Unilever, the Netherlands.

Susan Purcell, Senior Director Innovation, Ipsos, the UK.

Omkar Shinde, Associate Director Innovation, Ipsos, UK.

Pratik Rane, Research Manager Innovation, Ipsos, UK.

Shreeraj Dhekane, Research Manager Innovation, Ipsos, UK.

Jinho Cho, Director Innovation & Forecasting General Modeling Unit, Ipsos, South Korea.

Dr. Nikolai Reynolds, Global Head of Product Testing, Ipsos, Germany.

# Simulate the “What If” Before You Build

## Synthetic Populations as an Exploration-First Layer for Customer Insight, and the Governance that Keeps It Honest

*Mariana Camillo*

*Aline Souza*

*Leonardo Aragão*

*Fernanda Laurentys Airoldi*

*Claudia Furniel*

*Danny Wymmer*

## Esomar

Office address:

Burgemeester Stramanweg 105

1101 AA, Amsterdam

Phone: +31 20 664 21 41

Email: info@Esomar.org

Website: www.Esomar.org

Publication Date: September 2026

ESOMAR Publication Series Volume S424 Congress 2026

ISBN 92-831-0337-8

Congress 2026
Copyright © **ESOMAR** 2026

Copyright © **ESOMAR** 2026
Congress 2026

# Simulate the "What If" Before You Build

## Synthetic Populations as an Exploration-First Layer for Customer Insight, and the Governance that Keeps It Honest

*Mariana Camillo*

*Aline Souza*

*Leonardo Aragão*

*Fernanda Laurentys Airoldi*

*Claudia Furniel*

*Danny Wymmer*

### 1. Introduction

Insight teams have long worked validation-first: build the idea, then test it with customers, and often learn too late that the premise was wrong. This paper describes a metamorphosis in that sequence, not a break with research rigour but a continuity of it. By building synthetic populations from interviews with real customers, we created an exploration-first layer that lets teams simulate the "what if" of a concept before a single asset is built, and reserve human fieldwork for the decisions that truly need it.

The business case is the point, not a footnote to it. In an illustrative concept-exploration application, the cycle fell from about two weeks to a matter of hours, roughly a twenty-fold compression, while synthetic and real-customer results converged across key evaluation dimensions. That speed does not come from cutting research; it comes from moving it. With Simile as research partner, we report the validated convergence benchmark behind "strong alignment", about 86% distributional accuracy, or a Total Variation Distance of roughly 0.14, and ground the method in peer-reviewed Stanford research. We then set out the governance model we call Validation by Convergence, a four-day co-creation sprint that shows the approach in practice, an honest account of where the method is strong and where it must not be used, and a replicable blueprint other insight teams can adopt.

### 2. The Business Case: The Cost of Learning Too Late

Most product and communication decisions still follow a validation-first cadence. A team forms a hypothesis, invests in a concept or prototype, and only then exposes it to customers. When the idea fails, the organisation has already spent build time and budget on a premise customers were unlikely to accept. The learning arrives after the cost is sunk.

A second cost is signal loss. Research creates a rich reading of the customer, but the context is often weakened as work moves from insight to product, design and implementation. Every hand-off adds delay and strips away some of the reasoning that made the evidence useful in the first place.

At Itaú, this sequence reduced one illustrative concept-exploration cycle from approximately two weeks to hours, about a 20-fold compression, and supported a four-day co-creation sprint that had previously taken about five weeks. These are operational outcomes, not evidence that simulation is universally accurate. The methodological claim is narrower: a synthetic population can be useful when its population and question are covered by human grounding, its outputs are treated as probabilistic, and its performance is evaluated against human reference data.

Adding an exploration layer changes the economics: more alternatives can be considered while the decision is still open, weak assumptions are exposed earlier, and human research can concentrate on the shortlist where the stakes and residual uncertainty are highest.

The Congress theme, "metamorphosis with continuity", captures that change. The metamorphosis is the order of operations: explore first, validate second. The continuity is the discipline: every synthetic exploration is grounded in human evidence, bounded by its population and question, and checked before it carries decision weight. Figure 1 summarises the sequence.

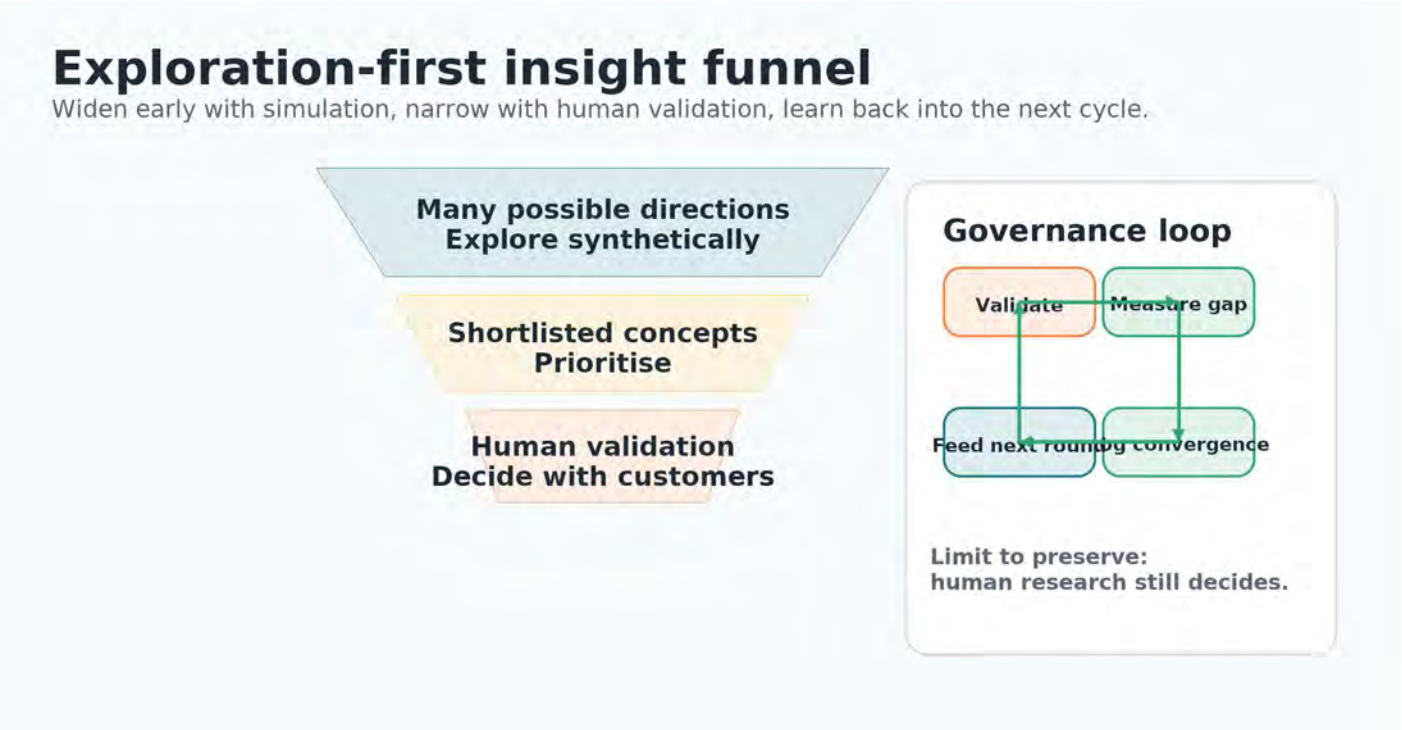



*Figure 1. Exploration widens the funnel at low cost; human validation narrows it on the decisions that matter, and the convergence record feeds the next round. The limit to preserve: human research still decides.*

## 3. The Shift: From Validation-First to Exploration-First

Exploration-first does not mean replacing customers with a model. It means inserting a fast, low-cost simulation layer before human validation. Teams can pressure-test alternative concepts, messages and service scenarios while they are still inexpensive to change. Human fieldwork then deepens and confirms the few directions that survive.

The relationship is co-evolving rather than sequential. Better interviews improve the simulated population; simulated scenarios expose gaps in the grounding and sharpen the next round of human questions. Researchers remain responsible for framing, interpretation, methodological judgement and the decision about what evidence is sufficient.

## 4. What a Synthetic Population Actually Is

A synthetic population is a model-based representation of a defined population, constructed from primary human research and other evidence appropriate to that scope. It can be queried at individual and aggregate levels to estimate how the represented population may respond to a new question or scenario. Its outputs are generated, not observed: they must be labelled as simulated and must never be reported as responses collected from customers.

Simulation is the scenario layer. A counterfactual asks what may happen if a condition changes, such as a message, benefit, journey step or product feature, and compares plausible alternatives that may not yet exist. The result is probabilistic and directional, not a deterministic statement about what every person will do.

The population is also bounded. It can speak only to people, topics and contexts represented in its construction and evaluation. Representativeness is therefore a design property of recruitment, quotas, evidence coverage, recency and question fit; it is not created automatically by generating more synthetic respondents. A population should be treated as silent, not authoritative, outside those boundaries.

Predicted Accuracy and Groundedness answer different questions. Predicted Accuracy is a model-derived estimate of how reliable a population-level prediction is expected to be for a closed question, calibrated from prior validations. Groundedness is the degree to which an individual simulated answer or quotation is supported by relevant evidence in the underlying human interviews. They are independent: high Groundedness does not compensate for low Predicted Accuracy, and neither is a guarantee of truth (Simile, 2026a; Simile, 2026b). Table 1 separates these signals from the human reference data used in evaluation.

| Signal | What it tells the researcher | What it does not establish |
|---|---|---|
| **Predicted Accuracy** | Expected reliability of the population-level prediction for a closed question, based on prior validation performance. | It is not a guarantee, an individual-level correctness rate or support for a quotation. |
| **Groundedness** | How directly a generated answer or quotation is supported by relevant evidence in the underlying human interview. | It is not population-level predictive accuracy and cannot repair missing population coverage. |
| **Human reference data (ground truth)** | Observed human responses or outcomes used to evaluate whether predicted distributions match human distributions. | It does not generalise automatically beyond the sampled population, question and context. |

*Table 1. Three different questions of evidence.*

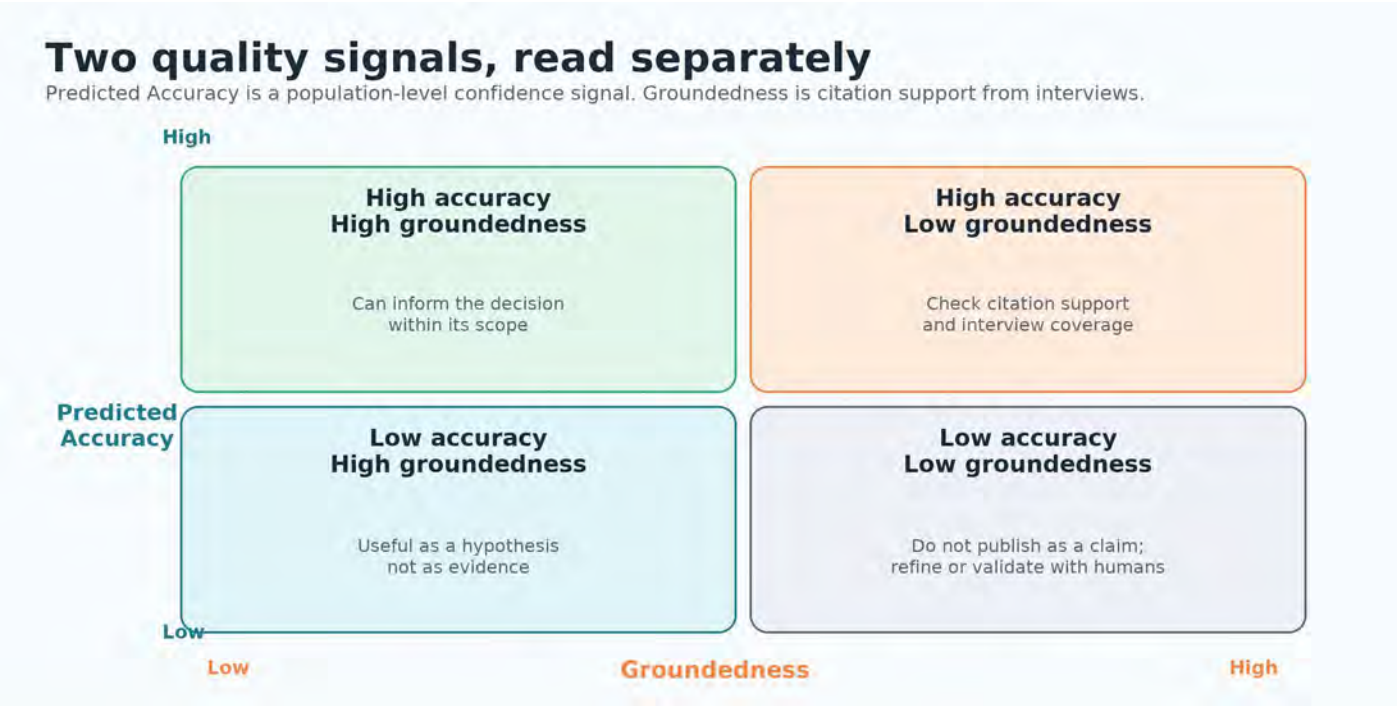



*Figure 2. Predicted Accuracy answers "how far can this population-level reading be trusted?"; Groundedness answers "how well does this individual citation reflect a real interview?" Read them separately.*

## 5. How It Works in the Insight Workflow

1. Ground. Collect or curate primary human interviews and profile evidence for the population and decision in scope. Document who is represented, what topics are covered and when the evidence was collected.
2. Explore. Query the population with concepts, messages, journey moments or counterfactual scenarios. Read the predicted distribution together with drivers, barriers, Predicted Accuracy and Groundedness where available.
3. Shortlist. Drop weak directions and identify the assumptions that most need human investigation. Simulation is most valuable here because the cost of considering another option is low.
4. Validate. Use human research to confirm the shortlist when the decision carries consequence, uncertainty or novelty. Human evidence remains the final gate for high-impact choices.
5. Compare and record. Compare simulated and observed human results using a labelled measure appropriate to the question, record agreement and divergence, and limit future trust to the validated class of population and decision.
6. Refresh. When new human evidence or contextual change makes earlier grounding less representative, rehydrate the population and re-evaluate before relying on the previous record.

## 6. Governance: Validation by Convergence

A simulation that is never checked against human evidence is a hypothesis generator, not a validation instrument. Validation by Convergence is our name for a decision process in which consequential synthetic outputs are compared with human reference data and the degree of agreement is recorded by population, question class and decision class. Strong performance in one context does not grant authority in another.

Convergence is not a single universal score. For a closed question it may be distributional fit; for concept exploration it may also include rank order, direction of preference and the principal drivers and barriers. Each measure must be named. Divergence is treated as a finding: it may reveal missing coverage, a poorly framed question, unstable human response, drift or an evaluation problem. It should change what the team does next.

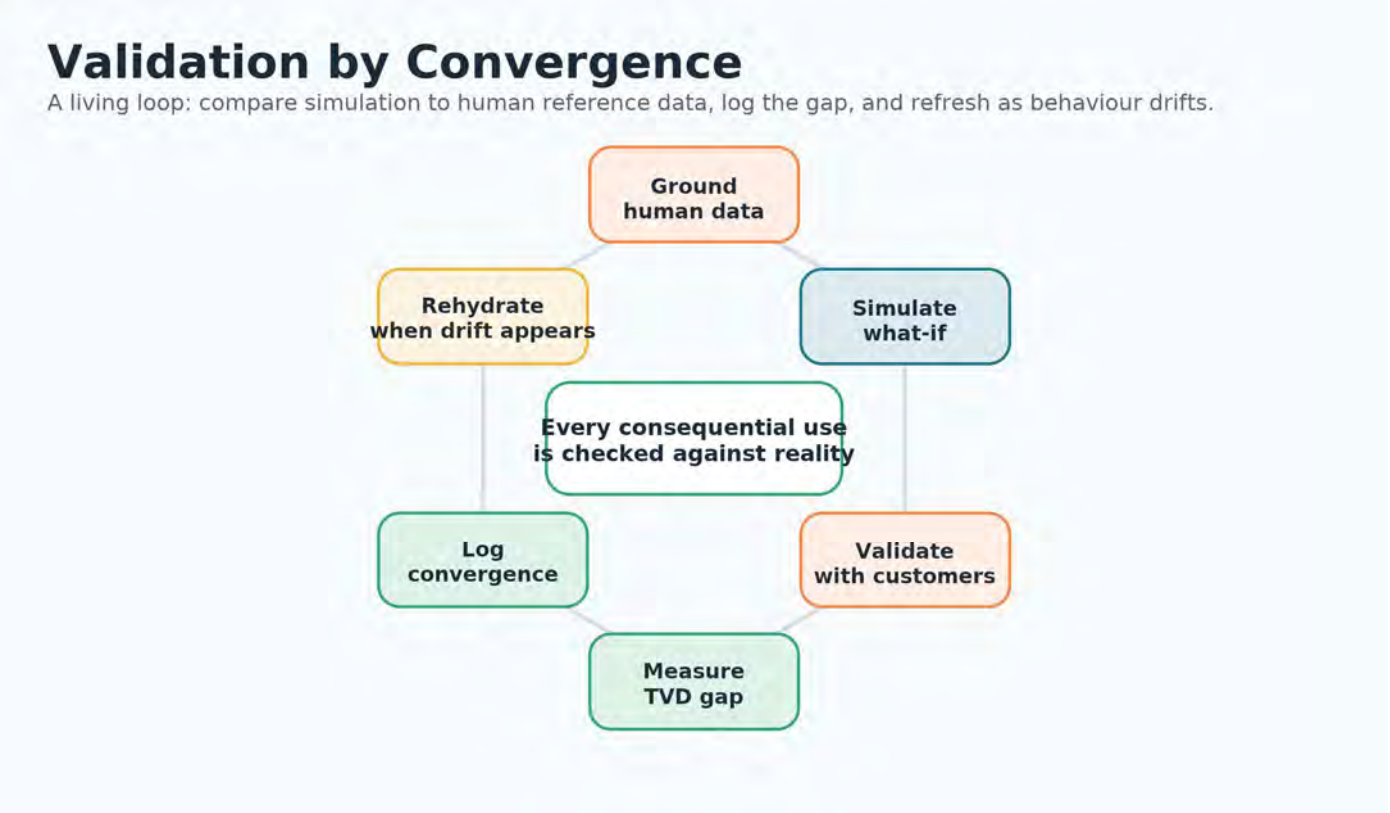



*Figure 3. Convergence is a loop, not a one-off check: every consequential use is measured against human reference data, the gap is logged, and drift monitoring triggers rehydration so the population stays representative over time.*

## *Hold-Out Evaluation: Test on What the Model Did Not See*

A clean evaluation begins before model construction. Human reference questions and answers are collected first; a subset is reserved as hold-out evaluation material; the population is built or calibrated on the remainder; the held-out questions are then asked through the simulation; and predicted distributions are compared with the human distributions. Questions that are materially equivalent to construction data must be detected and removed, not treated as independent evidence. A contamination audit should precede any published accuracy number.

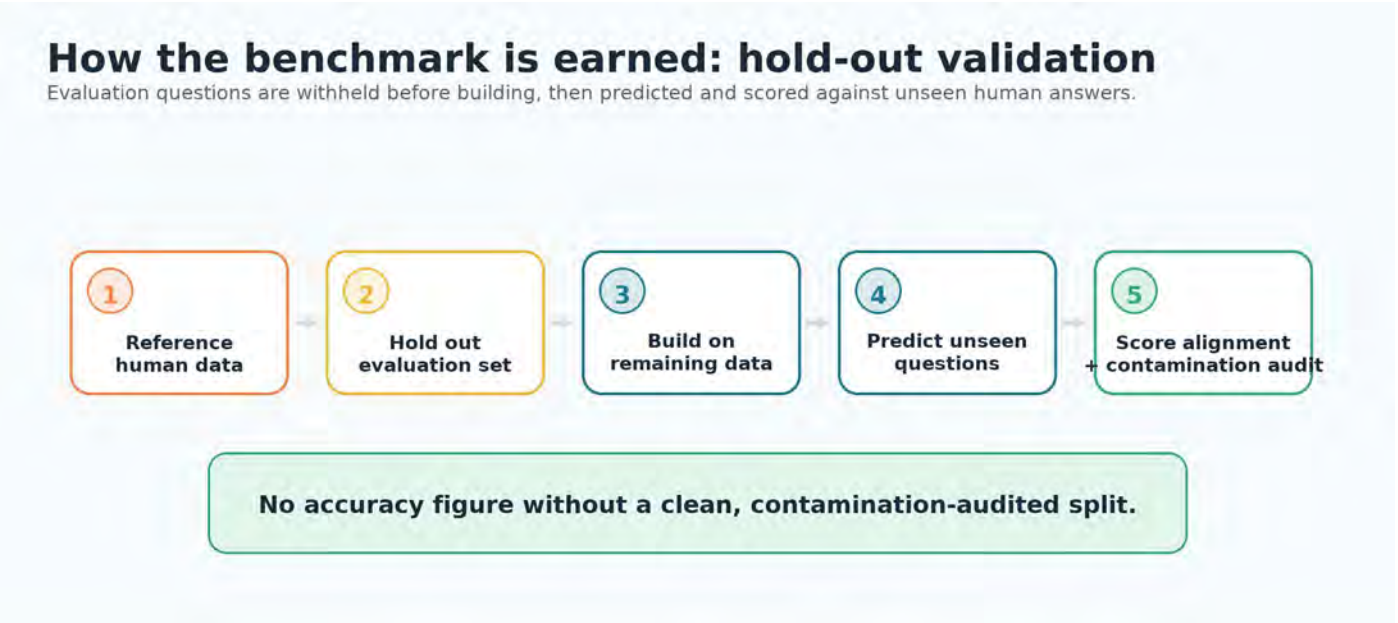



*Figure 4. The benchmark is earned, not asserted: evaluation questions are withheld before the population is built, then predicted and scored against the human answers the model never saw. The limit to preserve: no accuracy figure is reported without a clean, contamination-audited split.*

## *What "Strong Alignment" Means, in Numbers*

"Strong alignment" is a measured quantity, and it is worth being precise about which quantity. Simile's standard validation re-contacts participants and compares predicted distributions against real behaviour. In one such study, 500 participants from a Fortune 500 population were re-contacted and roughly 20,000 behavioural data points collected; the model reproduced the observed behavioural distributions with about 86% distributional accuracy, computed as one minus the Total Variation Distance (TVD ≈ 0.14). This is a population-level fit, not an individual-level prediction claim, and it provides an example of how convergence can be evaluated.

Distributional accuracy should not be read as individual-level prediction; the two are different measures, and conflating them overstates what the model does. At the individual level, Simile reports matching participants' own responses with about 85% accuracy across roughly 40,000 responses, with the important caveat that the remaining share largely reflects normal human inconsistency, since people answer the same question differently 15–30% of the time in reference surveys such as the General Social Survey (Hout & Hastings, 2016). The academic grounding sits alongside both: in Stanford research (Park et al., 2024), interview-grounded generative agents replicated individual survey responses at 0.85 normalised accuracy, against 0.71 for demographic-only agents and 0.33 for chance, with treatment effects correlated at r = 0.98.

Convergence statistics for individual Itaú studies are held to the same standard: reported internally and confirmed against the validated record before any external citation. The benchmarks above are Simile-reported and academic references against which internal results are read, not results imported from another population as if they were our own.

## *Interpreting Convergence with TVD: Making Distributional Convergence Visible*

For a closed question with response categories i, Total Variation Distance (TVD) compares the observed human distribution P with the simulated distribution Q: $TVD(P,Q) = 1/2 \sum_i |P_i - Q_i|$. TVD ranges from 0 to 1. A value of 0 means

the distributions match; a value of 1 means they do not overlap. Lower values therefore indicate closer distributional agreement. TVD describes the shape of population distributions, not whether each simulated individual gave the same answer as a particular participant.

The supplied Simile methodology classifies TVD below 0.20 as strong fit or high accuracy. We use that threshold only as the interpretation criterion declared in the referenced materials, not as a universal law. Its meaning depends on the population, question, reference data and evaluation design. Figure 5 makes that distinction visible.

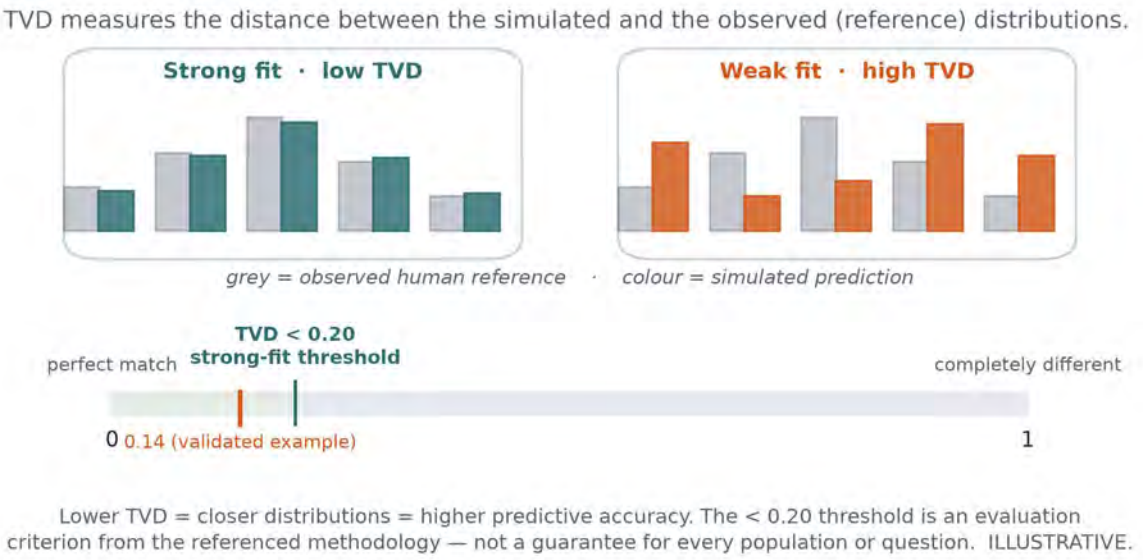



*Figure 5. Total Variation Distance is the distance between the simulated and observed distributions: 0 is a perfect match, and lower is better. Read the < 0.20 strong-fit band as an evaluation criterion from the referenced methodology, not as a guarantee that any population or question will reach it.*

## 7. Illustrative Application: Concept Exploration and Human Validation

To make the workflow concrete, consider an application run at Itaú: a concept exploration for a new automatic-payments experience, Pix Automático. A set of concept directions was first explored through the synthetic population, which returned adoption propensities and the drivers and barriers behind each direction. The team used this to drop the weakest concepts and refine the survivors before commissioning human validation on the shortlist.

The headline operational result was a reduction of the concept-test cycle from approximately two weeks to a matter of hours, on the order of a twenty-fold compression, with synthetic and real-customer results converging across key evaluation dimensions and the fit read on the TVD scale described above.

## 8. Simile in Practice: A Four-Day Co-Creation Sprint

The clearest demonstration of the exploration-first layer is a pilot in which an integrated Design and Research team redesigned a strategic co-creation initiative end to end. A cycle of this type traditionally takes about five weeks. The team compressed it to four business days, not by cutting corners, but by re-sequencing the work and assigning each task deliberately to a human or to AI.

### *The Sprint*

At the centre was a three-hour workshop with Product and Business leaders. Within that single session the team generated seven solution hypotheses and refined and stress-tested them in the room, in minutes, rather than across the weeks a traditional post-workshop cycle would require. The synthetic layer made it possible to interrogate each hypothesis against simulated customer reactions while leaders were still at the table, so the group left with a shared, evidence-informed shortlist instead of a list of things to test later.

### *What Made It Credible*

The speed rested on disciplined pre-work, all of it human-curated: a specialist curated 25 prior studies together with behavioural data from approximately 15,000 customers, synthesised into a structured reading of business and customer needs. From this, the team built a specialist agent, synthetic personas and a synthetic-population panel for in-session stress-testing. The division of labour was explicit: humans owned framing, curation, facilitation, interpretation and decision; AI was applied where scale and speed add value, such as structured synthesis, documentation and scenario exploration.

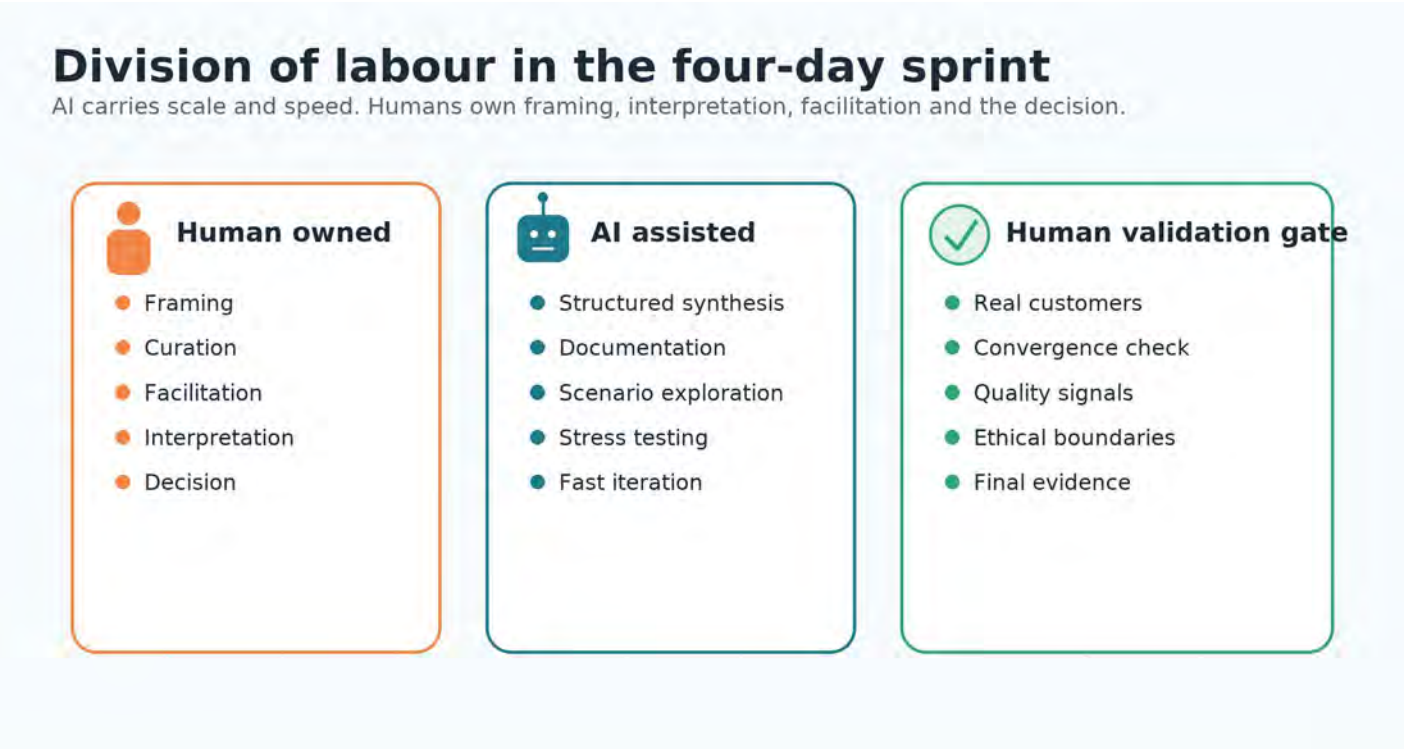



*Figure 6. The division of labour is deliberate: AI carries scale and speed, humans own framing, interpretation and the decision, and a human-validation gate spans both before anything consequential ships.*

Critically, human validation with real customers remained a non-negotiable gate. Personas and synthetic populations were treated as exploration instruments, not as final evidence. The sprint also produced reusable assets: a flow map, role definitions and prompt patterns that lower the cost of running the next sprint.

### *The Honest Limit*

The pilot does not claim to replace real-customer research. Its contribution is to move effort from the operational to the strategic, making a faster way of working viable without turning speed into a synonym for superficiality. The synthetic layer widened and accelerated exploration; human research still decided.

## 9. Findings and Outcomes

- Speed without abandoning rigour. Exploration that had taken about two weeks was compressed to hours, and a five-week co-creation cycle to four business days, while the human-validation gate remained.
- A wider exploration funnel. The lower marginal cost of comparing another direction allowed more alternatives to be considered before commitment.
- Interpretability changed the conversation. Drivers, barriers and interview-grounded explanations helped stakeholders interrogate a result instead of accepting a score at face value.
- Value in the room, not only in the deck. The layer could be queried while a decision was still open, allowing researchers and leaders to refine assumptions together rather than wait for a later read-out.
- Divergence became useful evidence. A mismatch between simulation and human research was logged as a signal to investigate coverage, question fit, drift or evaluation design, not hidden as a failed prediction.

## 10. Where It Is Strong, and Where It Must Not Be Used

Governance is the part of the method that makes the speed defensible. Before a simulation is run, the team should establish the decision consequence, the population and topic coverage, the expected reliability of the closed-question prediction, and the validation record for that class of use. Groundedness is reviewed separately for qualitative answers and quotations.

### *Strong Fit*

- Early-stage exploration and prioritisation of many concepts, messages or scenarios before build.
- Pressure-testing assumptions and generating hypotheses to take into human research.
- Reaching hard-to-recruit audiences for early exploration, where convergence has been demonstrated for that class of decision.
- Questions well represented in the grounding data, where Predicted Accuracy reads high and the population's coverage is dense.

### *Out of Scope: Do Not Use as a Substitute*

- Final go/no-go decisions of high consequence without human validation on the shortlist.
- Final calibration of pricing and elasticity, and decisions about under-represented populations not covered by the grounding interviews. The model should be treated as silent there, not authoritative.
- Any use that would imply synthetic respondents are real people, or that would report simulated figures as if they were collected from customers.

In practice, the boundary can be read off the model's own signals. Where Predicted Accuracy is high, a result can inform a decision; where it is medium or low, it is directional only and should be carried as a hypothesis; where it is very low, it is

exploratory and must not stand as evidence. Groundedness and population coverage act as a second gate: a thinly covered question, or one whose citations are weakly grounded, is routed back to human research regardless of how confident the population-level reading looks. Decision consequence sets the last line: the higher and less reversible the stakes, the more human validation must remain the gate, whatever the score.

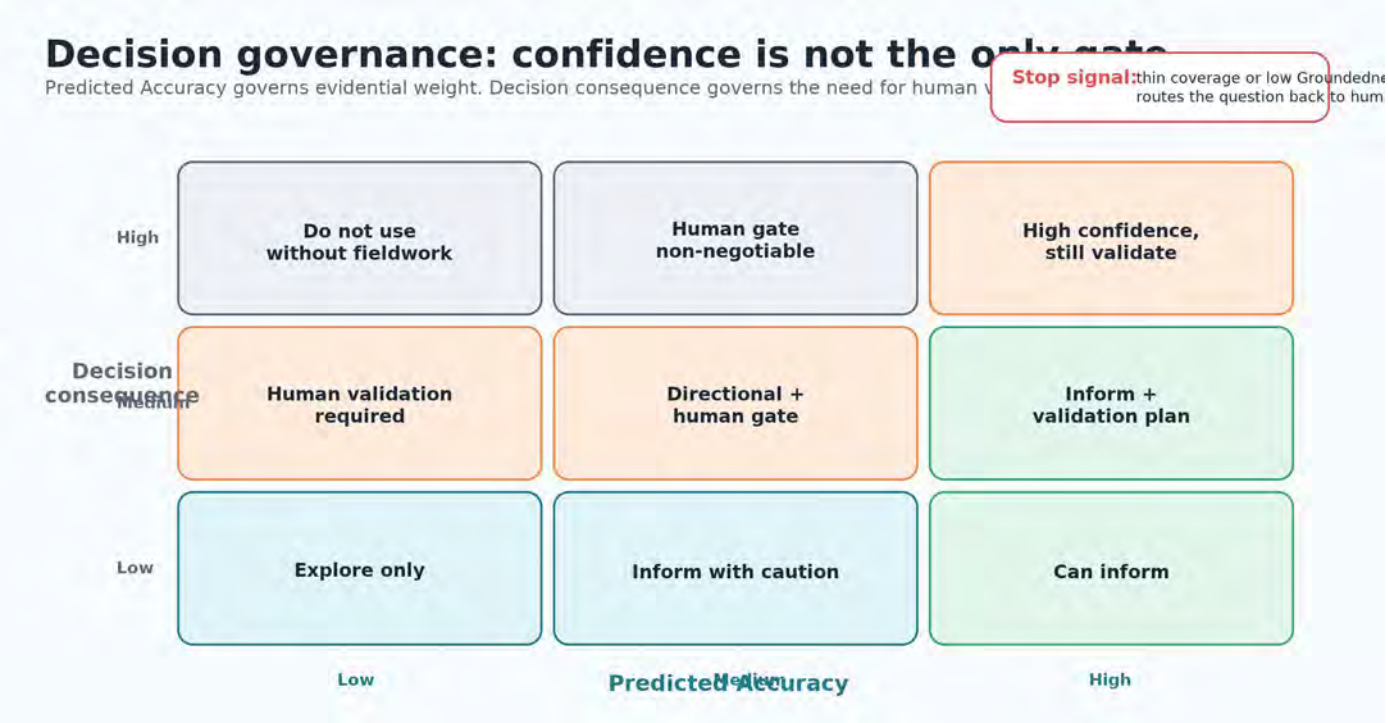



*Figure 7. Predicted Accuracy governs how much independent weight a result may carry; decision consequence governs whether a human gate is required. Thin coverage or low Groundedness is a stop signal on any consequential call.*

These boundaries are also an ethics safeguard. Synthetic data trained on synthetic data degrades, a risk documented in the literature on model collapse (Shumailov et al., 2024). Grounding every population in fresh human interviews, monitoring for drift and rehydrating when data ages are the practical defence against that drift.

## 11. Implications for the Industry

The exploration-first layer is not specific to banking or concept testing. It is a way to change where evidence enters a decision: before expensive commitment rather than after it. The most defensible transfers are uses in which alternatives can be compared cheaply and the final decision can still be validated with people.

- Marketing and communication. Compare message, claim and creative territories before production, then validate the shortlist with the intended audience.
- Customer experience and journeys. Stress-test a redesigned journey or service moment before build, focusing subsequent fieldwork on the moments with the greatest residual uncertainty.
- Product, proposition and service design. Explore feature and benefit combinations, identify trade-offs and narrow the portfolio before formal quantitative validation.
- Segmentation, strategy and innovation. Compare how represented segments may react to a move, generate hypotheses and widen the set of options considered at the front end.

For insight leaders, the shift reallocates human research rather than removes it. Budgets move from screening every early idea shallowly toward exploring more directions and validating the decisive few deeply. The transfer remains conditional: the population and question must be covered, synthetic outputs must remain clearly labelled, and divergence must change behaviour.

## 12. Governance in the Assurance Landscape

The approach is designed to sit inside the emerging assurance landscape rather than outside it. It aligns with the ICC/Esomar International Code on Market, Opinion and Social Research and Data Analytics (2016), and maps naturally onto the EU AI Act (Regulation (EU) 2024/1689), the AI-management-system standard ISO/IEC 42001:2023 and the NIST AI Risk Management Framework (2023). Each emphasises transparency, human oversight and documented validation. Validation by Convergence is the research-specific expression of those principles: quality signals on every answer for in-study transparency, a hold-out-validated convergence log for documented validation, continuous drift monitoring with rehydration to keep the evidence current, and a human-validation gate for oversight on consequential decisions.

## 13. A Replicable Blueprint

The approach is transferable. An insight team adopting it can start with five principles.

### *Five Principles*

- Ground in humans. Build every population from primary human research and evidence appropriate to the scope; do not use synthetic outputs as a substitute source of human truth.
- Explore first, validate second. Use simulation to widen the funnel; reserve human fieldwork for the decisive shortlist.
- Read the signals, not just the score. Keep Predicted Accuracy and Groundedness apart, so the team knows both how far to trust a reading and where it is unsure.
- Validate, log and refresh. Use clean hold-outs, record convergence and divergence, and rehydrate when new human evidence or context makes the prior grounding less representative.
- Stay inside the Code. Be transparent that respondents are synthetic, obtain client permission, and never present simulated figures as collected data.

### *Metamorphosis with Continuity: What Persisted After the Pilot*

The test of a metamorphosis is not a single fast cycle but whether the new way of working endures. Since the first applications, the exploration-first layer has moved from one-off pilot toward a repeatable practice: it has been applied across multiple proof-of-concept applications. Adoption has held because the discipline held: human validation remained the gate, convergence remained logged, and the continuity of rigour is precisely what allowed the change of sequence to stick.

## 14. Conclusion

The metamorphosis is a change of sequence, explore first and validate second, and the continuity is the rigour that makes it trustworthy. Synthetic populations, grounded in real customers, validated by convergence and kept current as

behaviour drifts, let insight teams simulate the "what if" before they build, learn earlier, and spend their human research where it counts. Used within its boundaries, the approach is not a replacement for the customer; it is a faster path back to them.

## References

Esomar (latest edition). *Guidance on the use of artificial intelligence and synthetic data in research*.

European Union. (2024) Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). *Official Journal of the European Union*.

Hout, M., & Hastings, O. P. (2016). Reliability of the Core Items in the General Social Survey: Estimates from the Three-Wave Panels, 2006–2014. *Sociological Science*, *3*, 971–1002.

ICC/Esomar (2016). *International Code on Market, Opinion and Social Research and Data Analytics*.

International Organization for Standardization (2023). ISO/IEC 42001:2023 — Information technology — Artificial intelligence — Management system.

National Institute of Standards and Technology (2023). *AI Risk Management Framework* (AI RMF 1.0). U.S. Department of Commerce.

Park, J. S., Bernstein, M. S., et al. (2024). *Generative Agent Simulations of 1,000 People*. arXiv:2411.10109.

Shumailov, I., Shumaylov, Z., Zhao, Y., Papernot, N., Anderson, R., & Gal, Y. (2024). AI models collapse when trained on recursively generated data. *Nature*, 631, 755–759.

## About the Authors

Mariana Camillo, Manager CX Research, Itaú Unibanco, Brazil.
Aline Souza, Business Agent, Itaú Unibanco, Brazil.
Leonardo Aragão, UX Researcher, Itaú Unibanco, Brazil.
Fernanda Laurentys Airoldi, Technical Lead CX Analytics & Research, Itaú Unibanco, Brazil.
Claudia Furniel, Head of CX Research, Customer & Market Insights, UX Research, Itaú Unibanco, Brazil.
Danny Wymmer, Itaú Unibanco, Brazil.

# Market Research from Evaluator to Architect

## How an Insight Team Re-Engineered the Casting Decision: A Vodafone Case Study

*Dirk Ziems*

*Christian Escher*

## Esomar

Office address:

Burgemeester Stramanweg 105

1101 AA, Amsterdam

Phone: +31 20 664 21 41

Email: info@Esomar.org

Website: www.Esomar.org

Publication Date: September 2026

ESOMAR Publication Series Volume S424 Congress 2026

ISBN 92-831-0337-8

Congress 2026
Copyright © **ESOMAR** 2026

Copyright © **ESOMAR** 2026
Congress 2026

# Market Research from Evaluator to Architect

## How an Insight Team Re-Engineered the Casting Decision: A Vodafone Case Study

*Dirk Ziems*

*Christian Escher*

### 1. Introduction: Re-engineering Casting in the Age of AI

The insight profession is being re-engineered in plain sight. Teams are reorganising, methods are blending, and workflows are being rebuilt to meet new expectations of speed, relevance and impact. Much of the public conversation about this change is dominated by artificial intelligence, and much of that conversation has curdled. After a period of inflated expectation – systems that would generate winning strategies and finished campaigns at the touch of a button – many marketing and insight leaders have entered a phase of disillusionment. Untrained or poorly trained language models produce generic copies; AI-generated visuals look interchangeable; agent workflows can often be fiddly and slower than expected. The reasonable scepticism around AI deployment is healthy, and this paper does not seek to dispel it purely with enthusiasm.

Our own framing, developed across several years of building psychologically grounded models, is deliberately modest about the technology. We treat AI not as artificial intelligence that replaces human expertise but as augmented intelligence that extends it – a co-intelligence that amplifies existing depth competence rather than substituting for it. This is what we call the "human in the loop" model. On that view, the interesting question is not what the tools can do unattended, but how they fundamentally reorganise workflows and most importantly, how they are able to add value to a human-centred processes. The difference between hype and durable value lies in the manner of use: delegate blindly to an untrained model and you produce arbitrariness; combine the model with depth-psychological grounding, based on real in-depth-interviews, and human oversight, and you produce a rigour and reliability (Ziems et al., 2025: 40).

This reframing matters because it relocates the problem. The challenge we describe is not, at root, a tooling problem. It is a workflow-design problem. Casting decisions go wrong not because creative teams lack talent or technology, but because the decision is structured badly: it happens too late, rests on intuition that is hard to interrogate, and sits outside the part of the process where consumer psychology is understood. If that is the problem, then better tools alone will not solve it. What changes the outcome is redesigning where insight sits in the process and what it is responsible for, transforming it into an iterative workflow that allows for optimal decisions to be reached on a consistent basis. That is the thread this paper follows: a real case in which an insight team re-engineered the casting workflow, and in doing so changed its own role from evaluator of finished creative to architect of the decision.

## 2. Why Casting is a Craft Problem the Discipline Under-Serves

Casting is consequential out of all proportion to the attention it usually receives from research. Small differences in who appears in an advertisement – their personality archetype, the way they hold the product, the outfit they wear – have a disproportionate effect on whether a brand communicates trust, aspiration or rejection. These are not marginal aesthetic details. In the Vodafone case they were the difference between a testimonial that reinforced brand meaning and one that eroded it.

This is easy to underestimate in the abstract. A facial expression meant to read as warm can instead read as manipulative; the way a person holds a product can signal possessiveness rather than generosity; a styling choice can tip a figure from credible into trend-led and undermine premium fit. Consumers experience these missteps as friction – a testimonial that feels too flippant, too aggressive, or too try-hard – without necessarily being able to name the cue responsible. The signal does real damage, but the cause stays buried beneath an overall impression.

Traditional insight is poorly positioned to catch this. Casting is normally evaluated post hoc: research is shown near-finished creative and asked whether it works, by which point the archetype has been cast, the shoot is booked or done, and the cost of change is high. The judgement that drives the original choice is intuitive and individual – a creative director's instinct about who feels right – and therefore difficult to scrutinise, document or defend. And it is made under time pressure, which compresses exactly the kind of careful, multi-layered reading that the decision actually requires. The result is a high-stakes decision evaluated late, subjectively and quickly: structurally under-served by the very discipline best equipped to inform it. The opportunity is to move insight upstream of the decision and give it a method in an iterative process.

## 3. The Re-Engineered Workflow: How the Work Has to Be Organised Differently

What distinguishes this case is less the inventory of tools than the way they were arranged. In the Casting Studio, this line was bent into a loop. Researchers, AI specialists, nonverbal expertise, visual creation and client decision-makers worked in repeated cycles, each iteration producing tangible options that were scored, interrogated and fed back in. Insight was not a stage in the chain; it was the organising hub around which the loop turned.

Four components were integrated into that loop. The first was a set of depth-psychology-based AI Consumer Twins: four psychologically differentiated twins, each representing a distinct target-group logic, built from real consumer depth interviews, concept m's telco lifestyle knowledge, and client input. These were not generic personas or ungrounded chatbots producing reflexive positivity. They played a deliberate dual role – audience simulators on one hand, running iterative individual interviews and focus-group-style sessions with a human or AI moderator, and creative optimisers on the other, reacting to candidate executions and helping shape the next iteration. Across the casting work a simulated sample on the order of 40 individual interviews and eight focus groups were generated. Crucially, they were treated as grounded simulation partners whose reactions were checked against known target-group logic, not as autonomous truth machines.

The second component was a specialist nonverbal communication expert system, informed by the Facial Action Coding System (Ekman & Friesen, 1978) and the principles of kinesics (Birdwhistell, 1970) and built on several thousand pages of domain-specific knowledge. Its job was to decode embodied meaning in visual material – to distinguish, for example, between a gesture that communicates offering and one that communicates pressure, distance, or authenticity. This is

precisely the layer that intuitive creative judgement struggles to articulate, and it functioned as a genuine second expert lens rather than a confirmation step.

The third component was generative photorealistic image AI, working alongside a specialist visual design collaborator, which translated insight hypotheses into testable casting alternatives – variations of type, body language and outfit that could be put back in front of the twins and the nonverbal system. The fourth was structured scoring. While the four depth twins drove the qualitative interviews and focus groups, scoring drew on a larger synthetic sample of n=200 AI Twins, each rating every iteration across four dimensions – Brand Fit, Product Fit, Customer Approach and Activation, on a 0–100 scale. This converted qualitative perception into comparable, actionable data, and reduced the risk that aesthetic novelty would be mistaken for persuasive effectiveness.

Conceptually, this maps onto a way of thinking about decision-making that concept m has set out elsewhere (Ziems et al., 2025: 40). Alongside the familiar distinction between intuitive and rational cognition, grounded synthetic consumer data introduces an additional input to the decision – a fast, psychologically informed source of consumer reaction that can be brought into the process continuously rather than at the end. The point of the architecture is to feed that input into a human-governed loop, moving deliberately from insight to impact to implementation rather than stopping at diagnosis.

## 4. The Vodafone Case in Depth: Three Layers, Worked Through

The casting work sat inside a broader pilot. Vodafone wanted to communicate a 30-day 5G trial offer through a customer flyer, and the qualitative response revealed that the offer was attractive but the communication generated resistance: the claim "Try for free" read as salesy rather than reassuring, the layout felt cluttered and overloaded, the printed activation steps looked like work, and the testimonials produced mixed reactions on trust and brand fit. The redesign followed an insight-to-impact-to-implementation arc – reframing the offer from a loud free trial into a more credible "thank-you gift", simplifying the information hierarchy, and replacing the printed step-by-step process with a single QR-led activation journey. Testimonial casting was one implementation stream within that wider redesign, and it is the stream we examine here because it shows the method at its most developed.

*Figure 1. The female testimonial before optimisation (rejected cast) and after (the winning "Doer" archetype).*

The central methodological claim emerged from the casting work: testimonial perception operates simultaneously across three interdependent layers, each capable of supporting or undermining the others. The first is personality type – whether the archetype aligns with the brand's premium promise. For the female testimonial, audiences wanted someone embodying competence, order and high-speed capability: an archetype we labelled "The Doer". Trend-led and influencer-adjacent types were rejected due to a lack of trustworthiness and seriousness, as were two named alternatives – "The Interchangeable Decision-Maker", who was generic and unmemorable, and "The Nice Collaborator", who was competent but too sweet, undermining the seriousness and authority the brand needed.

The second layer is body language and gesture – how posture, facial expression and object handling carry psychological meaning. Here the nonverbal system did its sharpest work. An open palm presenting the smartphone functioned as a gift signal; a Duchenne smile, the kind that reaches the eyes, communicated authentic joy rather than promotional performance; an open body orientation signalled trust and invitation. Together these produced what we came to call a "sharing psychology": the phone offered rather than displayed. The failure modes were just as specific. A static, deliberately presenting stance read as staged and inauthentic. A possessive grip communicated ownership rather than giving. A head tilt paired with a mischievous smile read as manipulative and promotional. None of these distinctions is visible to a quick intuitive read; all of them became legible through expert nonverbal analysis with trained AI.

The third layer is outfit and styling – how visual presentation reinforces or contradicts the type and body-language signals. The winning combination for The Doer paired a blue-and-white striped knit sweater (warmth and approachability), a loosely worn dark blazer (corporate credibility without stiffness), dark trousers (objectivity and confidence) and white leather sneakers (modernity and ease). It achieved a coherent brand signal across all four scoring dimensions, with the highest overall average of any option tested.

The scores carried out by n=200 AI Twins make the layering argument concrete. The winning outfit scored 85 on Brand Fit, 90 on Product Fit, 80 on Customer Approach and 85 on Activation. A rejected alternative scored 40, 30, 75 and 50 on the same dimensions – strong enough on customer approach to feel plausible at a glance, but collapsing on brand and product fit. Separately, the iterations surfaced a more instructive trap: a competing outfit carrying stronger explicit brand visual coding scored well on brand fit yet collapsed on product fit and activation. That pattern is the whole case in miniature. An option can win the layer you are looking at and lose the campaign, which is exactly why layer-by-layer evaluation matters. It took three outfit iterations before coherent fit across all four dimensions was achieved for the winning archetype.

The four-dimension scores for the female testimonial's outfit decision:

| Option | Brand Fit | Product Fit | Customer Approach | Activation |
|---|---|---|---|---|
| **Winning combination ("The Doer")** | **85** | **90** | **80** | **85** |
| **Rejected alternative** | 40 | 30 | 75 | 50 |

*Table 1. Scores on a 0–100 scale, applied by the AI Consumer Twins to each iteration. A male testimonial was optimised through the same three-layer process; the female testimonial is presented here as the fully worked example.*

The endpoint of this work was not a report. It was a casting decision that had been systematically stress-tested across personality type, body language and outfit, with explicit reasoning attached to every choice. That reasoning – not the imagery itself – is what makes the decision defensible, repeatable and transferable to future campaigns.

## 5. Assuring Quality in a Blended Workflow

A sceptical reader's first question about a workflow like this is the right one: if the audience is partly synthetic and the imagery is generated, how is validity protected? The honest answer is that validity was not a property of the tools; it was a property of how the workflow was disciplined. Several safeguards were built in deliberately, and they are the part of the case most worth copying.

The first safeguard was grounding. The credibility of the AI Consumer Twins came entirely from their foundation in real consumer depth interviews, telco lifestyle knowledge and client input, not from clever prompting. This human depth work was the prerequisite for everything that followed; without it, the twins would have produced plausible-sounding but ungrounded reactions. In our wider practice we treat depth interviews as the indispensable reference and control instrument for any synthetic model, and the same principle held here.

The second safeguard was treating nonverbal analysis as a genuine second expert layer rather than a rubber stamp. The FACS- and kinesics-informed system was there to challenge and decode the visual material, not to ratify decisions already taken.

The third was a refusal of single-pass acceptance: options were never adopted on the strength of one generation or one favourable reaction. Everything went round the loop, was scored and was compared.

The fourth was that twin reactions were continuously checked against known target-group logic, so that a result which did not cohere with what we understood about the audience was a flag for scrutiny, not a finding to be banked.

Underlying all of these is a principle we hold to firmly: the danger in this kind of work is the delegation trap – handing judgement to the system because it is fast and fluent. Augmentation, not delegation, is the operative idea. Oversight and steering stayed with the human researchers and the client throughout; the AI proposed, scored and accelerated, but did not decide. This is also where the most important learning sits. Speed and realism are not proxies for validity. A synthetic reaction can be instant and photorealistic and still be wrong, and a workflow that mistakes fluency for truth will fail confidently. The credibility of the output depended not on the speed of generation but on the rigour of the grounding and the discipline of the loop. Set against the wider concern about keeping people in control of AI, this is what responsible use actually looked like in practice: not a written policy, but a way of working with AI that is grounded in methodological rigour and kept human decision-making in the centre of it all.

## 6. The Role Shift: Researcher as Architect, Guardian and Translator

This is the heart of the argument. The project changed what insight was for. Research is no longer confined to evaluating finished material; it becomes the organising logic of the creative development process – generating hypotheses, structuring the decision, and translating psychological findings into design-ready options that could be tested and iterated before production. In the language of the theme, the insight team moves from being a post-hoc evaluator of finished creative to being the methodological architect of the casting decision itself.

That repositioning calls for three capabilities that are not the traditional core of the qualitative researcher's job. The first is workflow architect: someone has to design the loop – decide which disciplines enter where, what gets scored, how iterations are compared, and when a decision is mature enough to commit. The second is quality guardian: in a blended workflow the validity safeguards do not maintain themselves, and a named human has to own the grounding, resist single-pass acceptance, and keep the delegation trap at bay. The third is translation partner, working across research, design, strategy and AI to keep the disciplines speaking to one another.

Of these, translation is the genuinely new craft skill, and it is more specific than "AI literacy". Its essence is the ability to convert a psychological finding into a casting-ready hypothesis that a brief-writer, a casting director or an image model can actually act on. The instructive contrast is between a finding stated the old way and the same finding stated so that it can be built. "We need someone authentic" is not actionable; it is a feeling. "Open palm presenting the phone, Duchenne smile, no possessive grip, blazer worn loosely" is a specification. The first can only be evaluated after the fact; the second can be designed toward and tested before the shoot. That precision – turning psychology into buildable instructions – is what makes insight actionable in a production environment, and it is the skill on which the whole architect role turns.

There is a broader organisational reading of this, which is that the boundaries between functions soften. As insight takes ownership of decisions it used only to comment on, market researchers begin to operate as marketing consultants, and the people who own creative decisions become more like insight experts. The competence shifts from sitting in silos to spanning them. The Vodafone case is one concrete instance of that wider movement, and it suggests the shift is not aspirational but operationally available today.

## 7. Limitations and Open Questions

A few limitations are real and worth stating plainly.

First, the evidence base is narrow. This is a single brand, in a single market, on a single campaign. The three-layer framework and the workflow proved useful here, but we should not over-read one pilot into a general law of casting. Replication across brands, categories and markets is needed before the approach can be called robust rather than promising. The detailed example here is also limited to the female testimonial; a male testimonial went through the same process but is not presented in equal detail, and readers should weight the case accordingly.

Second, the validity of the twins is the central open question. They are grounded simulation partners, not autonomous sources of truth, and their reactions are only as good as the depth interviews and target-group logic beneath them. The continuous checking against known logic is a safeguard, not a proof; it can catch reactions that fail to cohere, but it cannot fully guarantee that a coherent synthetic reaction matches what real consumers would feel. How far twin-generated scores generalise beyond the conditions in which the twins were grounded remains genuinely uncertain, and we would not present the four-dimension scores as anything other than a structured aid to human judgement.

Third, and most importantly, human judgement remained indispensable throughout, and that is a limitation as much as a virtue. The method does not remove the need for expert interpretation; it concentrates it. Someone still has to ground the twins well, read the nonverbal analysis intelligently, decide when coherence across the layers has actually been achieved, and own the decision. The workflow disciplines and documents that judgement; it does not replace it. A team that adopted the tools without the discipline – or without the people able to exercise the judgement – would get the speed and lose the validity.

Copyright © **ESOMAR** 2026


## 8. Conclusion: What the Craft Looks Like Going Forward

The most useful thing this case offers the discipline is not a tool but an operating model. Casting was the example, but the structure generalises to many of the consequential creative decisions that insight has historically been invited to judge only after they were made. The pattern is: bring psychological grounding in before the brief is written rather than after the shortlist is approved; decompose a holistic creative judgement into the interdependent layers that actually drive it; introduce structured scoring not to replace human judgement but to discipline and document it; and run the whole thing as a human-governed loop in which AI proposes and accelerates while people decide.

What the Vodafone pilot produced beyond the casting decision was a reusable Casting Studio template – an adaptable workflow that can be applied across campaign formats, testimonial types and brand contexts, with reasoning attached to every choice. That reusability is the real deliverable, because it turns a one-off result into a repeatable practice.

If there is a single claim to carry away, it is this. The frontier of insight is not the arrival of AI; it is the redesign of how evidence is created and applied around it. In that redesign the researcher's role expands rather than contracts – from evaluator of finished work to architect of the decision, from commentator to quality guardian, from analyst to translator between psychology and production. The craft is not being automated away. It is being asked to take ownership of decisions it used to watch from the outside. That is a more demanding job than the one it replaces, and a more valuable one.

## References

Birdwhistell, R.L. (1970). *Kinesics and Context: Essays on Body Motion Communication.* University of Pennsylvania Press.

Ekman, P., & Friesen, W.V. (1978). *Facial Action Coding System: A Technique for the Measurement of Facial Movement.* Consulting Psychologists Press.

Ziems, Dirk, et al. (2025). Vom Hype zum Hebel für Exzellenz. *Markenartikel*, *10*, 40–43.

## About the Authors

Dirk Ziems, Co-Founder, concept m ai, Germany.
Christian Escher, Senior Manager Market Insights, Vodafone GmbH, Germany.

# Tuned to Fit

## Reporting Standards for LLM-Based Silicon Samples

*Jan Polišenský*

*Matej Koreň*

*Radek Hranický*

*Kamila Zahradníčková*

## Esomar

Office address:

Burgemeester Stramanweg 105

1101 AA, Amsterdam

Phone: +31 20 664 21 41

Email: info@Esomar.org

Website: www.Esomar.org

Publication Date: September 2026

ESOMAR Publication Series Volume S424 Congress 2026

ISBN 92-831-0337-8

Congress 2026
Copyright © **ESOMAR** 2026

Copyright © **ESOMAR** 2026
Congress 2026

# Tuned to Fit

## Reporting Standards for LLM-Based Silicon Samples

*Jan Polišenský*

*Matej Koreň*

*Radek Hranický*

*Kamila Zahradníčková*

### The Same Question, a Different Answer

Silicon samples are entering market research fast (Chintalapati & Pandey, 2022; Brand et al., 2023). Instead of recruiting people, waiting weeks for fieldwork, and managing panel quality, a researcher can prompt a large language model (LLM) to simulate responses from a defined segment, instantly, repeatedly and at a fraction of the cost (Argyle et al., 2023; Sarstedt et al., 2024; De Winter et al., 2024). The appeal is real. But speed and scale do not automatically produce validity.

Here is the uncomfortable part. The output of an LLM-based silicon sample is not determined only by the research question. It is also determined, often decisively, by a set of configuration choices that most researchers neither fully control nor consistently report: which model, what temperature, what prompt, how many personas. Change those settings and the same model, asked the same question, returns a meaningfully different distribution of answers. And as the experiments below show, with selective tuning you can make AI-generated data look like human data without that resemblance being robust, general, or honestly arrived at.

These are not edge cases. They are structural features of how LLM-based silicon samples work. This paper documents them with controlled experiments, explains why they create a governance gap, and ends with a concrete proposal: minimum reporting standards for AI-generated samples.

### A Hidden Layer of Decisions

In traditional survey research, transparency is a baseline expectation. Researchers report the sampling frame, the instrument, the fieldwork dates and the analytical procedure. Readers can then judge whether a finding is robust, replicate it or spot a source of bias. These norms exist because research decisions shape research outcomes, and trust depends on being able to scrutinise how a result was produced.

LLM-based silicon sampling introduces a new layer of decisions that sits before any analysis begins, at the data-generation stage: model selection, temperature, prompt architecture, persona construction, population size and aggregation rules. These are not neutral technical defaults. They are researcher degrees of freedom, the same kind of flexibility that the replication crisis taught empirical research to distrust in analysis (Simmons et al., 2011; Gelman & Loken, 2014), except that here they are usually invisible. When they go unreported, a reader cannot tell whether a finding reflects a genuine pattern or the particular settings under which the data were generated. Reproducibility becomes impossible.

Building on the observation that LLM performance is uneven and configuration-sensitive across tasks (Santurkar et al., 2023; Atari et al., 2023), we describe the resulting instability in generated respondent data as jagged samples: distributions whose shape shifts with the configuration rather than with the question. The term names the problem this paper sets out to make visible and auditable.

## What We Did

We used a controlled parametric design to isolate how configuration choices affect the distribution of synthetic responses. The logic is simple: hold the survey instrument, wording, and response format constant, then vary the configuration systematically and watch what happens to the distribution. The unit of analysis is the distribution of responses across a synthetic population, not the accuracy of any single simulated individual.

| Design element | Specification |
|---|---|
| **Models** | GPT-4.0 Turbo, LLaMA-3.1-8B-Instruct, DeepSeek-V3 (accessed via API; no fine-tuning or retrieval) |
| **Instruments** | Big Five Inventory (BFI-10; Costa Mastrascusa et al., 2023) and the Dark Triad (narcissism, Machiavellianism, psychopathy; Jones & Paulhus, 2014) chosen for well-documented human benchmarks |
| **Temperature grid** | 0.0 to 2.0 in steps of 0.1 (0.0 to 1.5 for DeepSeek-V3) — 21 conditions per model and battery |
| **Population size** | Synthetic populations of n = 10 and n = 100 personas |
| **Held constant** | Items presented verbatim; responses constrained to the 1–5 Likert scale; identical prompt structure across conditions |
| **Analysis** | Linear regression for temperature trends; independent-samples t-tests for model comparison; mean squared error (MSE) for benchmark alignment; $\alpha = .05$ |
| **Benchmarks** | German representative BFI study (n = 2,567; Rammstedt, 2007); Open-Source Psychometrics Project for the Dark Triad. Used for evaluation only. |

*Table 1. Experimental design.*

Each model impersonated every persona and completed each battery; responses were aggregated into trait-level means and standard deviations for each of the 21 temperature settings, for each model and battery, and stored separately by condition for traceability. All prompts, configurations, code and outputs are archived in a public repository. The study tested four straightforward hypotheses: that population size, temperature and model choice would each leave the response distributions unchanged, and that the synthetic distributions would not differ from human benchmarks. As we report below, three of the four did not survive contact with the data.

## What We Found

### *Population Size Barely Moved the Distributions, But It Can Still Mislead*

Comparing populations of 10 and 100 personas on the Big Five, only two standard deviations differed significantly, a negligible effect within this range. Population size was not a major driver of the distribution's shape. The caution is different and important: because conventional significance tests shrink standard errors as n grows, a large synthetic

sample can manufacture statistical confidence without improving validity (Andrade, 2020; Sarstedt et al., 2024). Size buys certainty in the arithmetic, not fidelity to the world.

### *Temperature Materially Changed Both the Mean and the Spread*

Temperature, the parameter governing randomness in generation, produced statistically significant shifts in the mean or the variance across most model–battery combinations. The effect was strongest in GPT-4.0, which across the four batteries showed twelve significant temperature-driven shifts in trait means and fourteen in variance. Because the temperature most commonly used in published work is 0.0, on the reasoning that it is "mostly deterministic", this sensitivity is rarely surfaced. Yet it means that two studies using the same model and prompt can report different distributions for no reason other than an unreported slider.

### *Model Choice Introduced Systematic Differences*

Holding the population and prompt fixed, the choice of model produced statistically significant differences in the response distributions, regardless of temperature. The largest gaps appeared between LLaMA-3B and DeepSeek-V3, concentrated in the standard deviations, which points to genuine differences in how much each model disagreed with itself. The Big Five was the least model-sensitive battery; the darker, more contested traits were more sensitive (Motoki et al., 2024). In short, "the LLM said so" is not a method. Which LLM matters, and it matters before any analysis begins.

### *You Can Tune It to Match Human Data – But the Match Does Not Generalise*

The most consequential finding concerns benchmark alignment. Apparent realism was achievable: at specific temperatures, synthetic distributions came very close to published human data, in some cases nearly copying them (mean squared errors as low as 0.03–0.08). But the temperature that produced the best match was not stable. It differed across models, and it differed across batteries within the same model — and it usually sat well above the common default of 0.0.

| Battery | GPT-4.0 | LLaMA-3B | DeepSeek-V3 |
|---|---|---|---|
| **Big Five (BFI-10)** | 1.1 (0.74) | 1.7 (0.99) | 1.2 (0.79) |
| **Machiavellianism** | 1.3 (0.14) | 1.9 (1.04) | 0.6 (0.06) |
| **Narcissism** | 1.1 (0.07) | 1.2 (0.32) | 0.7 (0.03) |
| **Psychopathy** | 1.5 (0.04) | 0.4 (0.63) | 1.1 (0.08) |

*Table 2. Temperature that best matched the human benchmark for each model and battery, with the mean squared error in parentheses (lower is closer). The best-fitting setting moves around, across models and across tests.*

Read across any row and the "best" setting jumps; read down any column and it jumps again. The implication is blunt: you can make AI-generated data look like human data, but that appearance is a function of how you configured the model, not evidence that the model reliably simulates human response patterns. Alignment is a conditional property of a specific configuration, not an intrinsic property of a model. A claim that a silicon sample "matches human data" is therefore only meaningful once the configuration, and how it was chosen, is on the table.

Copyright © **ESOMAR** 2026


## Why This is a Governance Problem, Not a Technical Footnote

Put together, the results describe a structural feature of LLM-based silicon sampling: configuration choices create degrees of freedom at the data-generation stage that have no clean equivalent in traditional survey research. When those choices are unreported, they reproduce exactly the failure mode that pre-registration and disclosure norms were designed to contain, only earlier in the pipeline, where they are harder to detect and impossible to correct after the fact. The risk is not bad faith. A researcher can tune temperature in good faith, find the setting that best matches a benchmark, report only that one, and unknowingly present a configuration-specific fit as a general result.

Two things make this urgent for the insights industry specifically. First, the tools are accessible, so these practices are scaling rapidly across teams that may not see configuration as a methodological decision at all. Second, the choices operate before analysis, which means they shape the data on which every later claim rests. The honest way to describe an LLM-based silicon sample is not as a neutral substitute for human respondents but as an instrument whose outputs encode model architecture and researcher decisions. Treated that way, the remedy is familiar: make the instrument's settings visible.

One scope note matters here. These findings concern LLM-based silicon samples specifically, approaches that rely on stochastic language generation. Other synthetic-sample architectures, including rule-based and hybrid systems, embed different assumptions and sources of variability, and the configuration sensitivities documented here may not transfer to them. The reporting standards below are written for the LLM case, where the degrees of freedom are both pervasive and consequential.

## A Practical Proposal: Minimum Reporting Standards

This paper does not end with a warning; it ends with a checklist. Drawing directly on the experiments, we propose a minimum set of disclosures for any study that uses an LLM to generate synthetic survey data, grouped into five clusters that track the five questions a reader should be able to answer: which model, what prompt, what settings (and what else you tried), why this configuration, and where to check it. Items marked with an asterisk (*) are specific to LLM-based silicon sampling and address degrees of freedom unique to model-driven data generation; the others are standard methodological practice applied to a new context.

| Cluster | What to report | Why it matters |
|---|---|---|
| **1. Model**<br>**Which model?** | Model identification – name and version, access method (API or local), and the date window of access. | Model choice produced significant distributional differences; results cannot be replicated without it. |
| **2. Prompt**<br>**What prompt?** | Population generation – how the population was built, its size (n), and whether it was generated once or per item.<br>Prompt & instruction * – the persona and task instructions used to elicit responses (verbatim or in an appendix). | In LLM-based sampling the prompt is the measurement instrument: it sets the sampling frame and the shape of the data. |

| | | |
|---|---|---|
| | Response constraints – allowed response format and scale range. | |
| **3. Settings**<br>**What settings and what else did you try?** | Configuration parameters – temperature value(s), with an explicit note when defaults were used.<br>Parameter exploration – the full grid tested: range, step size, number of configurations. | Temperature drove significant shifts in mean and variance; reporting a single setting hides the sensitivity revealed across the grid. |
| **4. Choice & evaluation**<br>**Why this configuration?** | Selection logic – the criterion for the reported configuration, and whether it was pre-specified or post hoc.<br>Unit of analysis – individual-level or distribution-level.<br>Evaluation & benchmarks – metrics, tests, thresholds, and the source and characteristics of any benchmark. | Alignment is achievable at specific settings, so without the selection criterion and the benchmark, post-hoc fitting cannot be ruled out. |
| **5. Repository**<br>**Where can I check it?** | Replication materials – prompts, configurations, code, and outputs at a persistent public link. | Given the sensitivity to configuration, replication materials are what make every item above verifiable. |

*Table 3. Minimum reporting standards for LLM-based silicon samples, grouped into five clusters. * marks items specific to LLM-based data generation.*

These are minimum standards, not a restrictive framework. They do not prescribe acceptable parameter values, prohibit exploratory analysis, or limit methodological innovation. Their purpose is narrower and more practical: to make the configuration choices embedded in AI-generated data visible and auditable, so that readers, reviewers, and clients can evaluate what they are actually looking at. The parallel is with the standards machine-learning research already expects around model and hyperparameter disclosure, LLMs are machine-learning models, and the same logic applies. Where full disclosure is impractical, this is the viable minimum: model, prompt, parameters, selection logic and a repository.

## Why This Matters Now

The insights industry is at an early and consequential point in its adoption of AI-generated data. The decisions made now, about what counts as acceptable practice, what disclosure is expected, what governs AI-enabled research, will shape the credibility of the field for years. Silicon samples are not inherently unreliable. But their reliability is architecture-dependent and configuration-sensitive in ways that are not yet widely understood, and the same flexibility that makes them powerful makes them easy to bend, intentionally or not, toward a desired conclusion.

The fix is not to slow innovation. It is to hold AI-generated data to the standard we already expect everywhere else: that the choices behind a number are visible enough for someone else to check them. When AI enters the research process as a data source, it should arrive with the same rigour and reproducibility as any other instrument. Trust in AI-enabled insight work depends on it – and getting the reporting conventions right while the field is still forming is far easier than retrofitting them once bad defaults have set.

Copyright © **ESOMAR** 2026


## References

Andrade, C. (2020). Sample size and its importance in research. *Indian Journal of Psychological Medicine, 42*(1), 102–103.

Argyle, L. P., Busby, E. C., Fulda, N., Gubler, J. R., Rytting, C., & Wingate, D. (2023). Out of one, many: using language models to simulate human samples. *Political Analysis, 31*(3), 337–351.

Atari, M., Xue, M. J., Park, P. S., Blasi, D., & Henrich, J. (2023). Which humans? [Preprint]. PsyArXiv.

Brand, J., Israeli, A., & Ngwe, D. (2023). Using LLMs for market research (Working Paper No. 23-062). Harvard Business School.

Chintalapati, S., & Pandey, S. K. (2022). Artificial intelligence in marketing: a systematic literature review. *International Journal of Market Research, 64*(1), 38–68.

Costa Mastrascusa, R., Loli, M., de Albuquerque, N. S., et al. (2023). Evaluating the complete, short and ultra-short versions of the Big Five Inventory (BFI) in the Brazilian population. *Scientific Reports*, 13, 7372.

De Winter, J. C. F., Driessen, T., & Dodou, D. (2024). The use of ChatGPT for personality research: administering questionnaires using generated personas. *Personality and Individual Differences*, 228, 112729.

Gelman, A., & Loken, E. (2014). The statistical crisis in science. *American Scientist, 102*(6), 460–465.

Jones, D. N., & Paulhus, D. L. (2014). Introducing the Short Dark Triad (SD3): a brief measure of dark personality traits. *Assessment, 21*(1), 28–41.

Motoki, F., Pinho Neto, V., & Rodrigues, V. (2024). More human than human: measuring ChatGPT political bias. *Public Choice*, 198, 3–23.

Rammstedt, B. (2007). The 10-item Big Five Inventory: norm values and investigation of sociodemographic effects based on a German population representative sample. *European Journal of Psychological Assessment*, *23*(3), 193–201.

Santurkar, S., Durmus, E., Ladhak, F., Lee, C., Liang, P., & Hashimoto, T. (2023). Whose opinions do language models reflect? *Proceedings of the 40th International Conference on Machine Learning*.

Sarstedt, M., Adler, S. J., Rau, L., & Schmitt, B. (2024). Using large language models to generate silicon samples in consumer and marketing research: challenges, opportunities, and guidelines. *Psychology & Marketing*, *41*(6), 1254–1270.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*(11), 1359–1366.

## About the Authors

Ing. Jan Polišenský, CTO, Lakmoos AI s.r.o., Brno, Czech Republic.
Ing. Matej Koreň, Student, Brno University of Technology, Brno, Czech Republic.
Ing. Radek Hranický, PhD, Assistant Professor, Brno University of Technology, Brno, Czech Republic.
Bc. et Bc. Kamila Zahradníčková, MSc, PhD Student, Department of Entrepreneurship, Prague University of Economics and Business, Prague, Czech Republic.

# Metamorphosis in Practice: From Static Insights to Creative Springboards

## How Culture-Led Insight Sprints Helped Unilever Build Briefs That Inspire Stronger Creative Outcomes Faster

*YeeMei Chan*

*Lanny Li*

Esomar

Office address:

Burgemeester Stramanweg 105

1101 AA, Amsterdam

Phone: +31 20 664 21 41

Email: info@Esomar.org

Website: www.Esomar.org

Publication Date: September 2026

ESOMAR Publication Series Volume S424 Congress 2026

ISBN 92-831-0337-8

Congress 2026
Copyright © **ESOMAR** 2026

Copyright © **ESOMAR** 2026


# Metamorphosis in Practice: From Static Insights to Creative Springboards

## How Culture-Led Insight Sprints Helped Unilever Build Briefs That Inspire Stronger Creative Outcomes Faster

*YeeMei Chan*

*Lanny Li*

### Rethinking the Role of Insight in Creative Development

Traditional models of creative development assume that the "insight" is complete before creative work begins. In practice, however, many briefs are built on generalised truths, category clichés or internal assumptions. This disconnect creates a structural flaw: creative teams are asked to build compelling ideas on unstable foundations.

This was evident across multiple Unilever brands. Despite strong strategic direction and high-quality executional capabilities, campaigns frequently failed to achieve strong pre-testing results. The issue was not with creative craft, but with the clarity and cultural relevance of the underlying insight.

Repeated cycles of development underscored the problem. Creative routes were explored, refined, and retested, yet performance remained inconsistent. This pattern indicated that execution alone could not compensate for a weak starting point.

Recognising this, the team reimagined the process entirely. Instead of a linear sequence: *insight → brief → creative → test*, the Insights Activator introduces a more iterative, collaborative sprint designed to co-create and validate insights before creative development begins.

### The Insights Activator: A Culture-Led Insight Sprint

The Insights Activator is a structured, two-to-three-week sprint that brings together brand teams, insights specialists and agency partners to develop and validate culturally grounded creative springboards.

The process consists of five interconnected stages:

### *1. Interrogating Existing Cultural Knowledge*

The sprint begins by consolidating and interrogating existing knowledge sources. These include youth insight studies, social tension reports and cultural learning programs across markets.

Rather than treating existing knowledge as fixed, this stage critically examines how culture is evolving, identifying tensions that are emerging, intensifying, or shifting in meaning. This ensures that insight development begins from a culturally informed baseline aligned with contemporary realities.

### *2. Community Immersion and Real-Life Engagement*

Rather than relying solely on desk research, the team immerses itself in consumers' real-world environments.

Teams engage directly with communities, such as Gen Z consumers, performers and grooming-focused individuals, through in-person conversations and contextual explorations. These interactions provide access to authentic language, lived tensions and coping mechanisms.

This stage builds deep empathy and surfaces nuances that structured research alone cannot capture. It shifts the insight from abstraction to lived experience.

### *3. Social Trend Scanning and Cultural Signal Mapping*

The third stage integrates real-time social listening and trend analysis. Teams analyse emerging narratives, expressions, and content formats across social platforms.

For example, in the CLEAR China case, trending youth expressions on social media such as 活人感 (feeling alive) and 活人微死 (a state of numbness within daily routine) captured a widespread emotional condition among young consumers. These signals help translate macro tensions into culturally fluent expressions that can inspire creative work.

### *4. Rapid Insight Articulation and Validation*

Instead of converging too early on a single insight, multiple articulations of tensions, motivations, and benefit statements are explored and tested with consumers.

This identifies:

- which emotional framings resonate most strongly
- which language feels authentic and compelling
- which benefit connections are credible

By validating these components upfront, ambiguity is removed before briefing begins.

### *5. Developing Validated Creative Springboards*

The output of the sprint is not a single insight statement, but a set of creative springboards – structured platforms combining cultural tensions, consumer motivations, and brand roles. These springboards provide:

Copyright © **ESOMAR** 2026


- clear tension territories
- validated language and expressions
- multiple routes for creative exploration

This fundamentally changes the nature of the brief – from a directive into a dynamic creative launchpad.

## From Insight to Brief: Embedding Springboards into Creative Development

A critical step in the Insights Activator process is the translation of validated insight platforms into the creative brief.

Rather than handing over a single insight, the outputs are structured into clear, actionable inputs that guide agency development from the outset. These include prioritised tension territories, validated language and defined roles for the brand within each creative route.

This ensures that the brief is no longer based on assumptions, but on consumer-validated insight articulation.

In practice, this changes how creative teams engage with the brief. Instead of interpreting a broad territory, they work from sharper, pre-tested directions, reducing ambiguity and accelerating ideation.

This step is critical in closing the gap between insight generation and creative execution – ensuring that cultural understanding is not lost in translation, but actively shapes the work that follows.

## Case Application: Reframing CLEAR Through Culture-Led Insight Activation Business Context

### *Rejuvenating CLEAR for a New Generation*

CLEAR was at a pivotal point in its brand journey. While it had strong heritage in anti-dandruff efficacy, its communication risked feeling overly functional and increasingly disconnected from the expectations of younger consumers.

The brand's ambition was to rejuvenate itself into a more modern, premium, and culturally relevant proposition – particularly among youth audiences navigating complex, high-pressure urban lifestyles. This required a deliberate shift away from traditional "clean & fresh" messaging toward a more elevated territory that could connect with how young people feel, not just what they need.

The concept of "scalp breathability" provided a compelling functional foundation – anchored in the idea that the scalp becomes suffocated by sweat, oil, and environmental stressors, and that CLEAR restores freshness and comfort. However, on its own, this benefit risked remaining technical and undifferentiated.

The key challenge, therefore, was not to define a new benefit but to reinterpret an existing one in a way that was culturally meaningful, emotionally resonant, and distinctively CLEAR. This meant finding ways to connect "breathability" to deeper human tensions around stress, pressure, identity, and the desire to feel alive again.

It is within this context that the Insights Activator Sprint was deployed.

## Application Across Markets: China and Indonesia

The Insights Activator was applied to CLEAR across multiple markets, most notably China and Indonesia, to test both its depth and scalability.

- China served as the primary deep-dive case, where the full cultural interrogation and insight development process was applied in detail.
- Indonesia provided validation of the model's transferability, demonstrating how the same approach could be adapted to a different cultural context while delivering similarly strong outcomes.

This dual-market application is critical: it illustrates not only how insights can be deeply rooted in local culture, but also how a consistent process can generate impact across diverse markets.

## China Deep Dive: From Macro Tensions to Cultural Insight

### *Understanding the Cultural Backdrop*

We start with cultural and macro tensions because they reveal the deeper forces shaping consumers' lives, not just their category behaviours. This gives us a sharper understanding of the emotional context in which the brand must play. By grounding the creative brief in real, contemporary tensions, we move beyond generic truths and create richer, more authentically resonant springboards for creative development.

The process began with interrogating and mining existing cultural learnings within the organisation network around Chinese youth. Several macro tensions emerged as defining forces shaping their daily lives:

- Intense competition driven by population density and limited opportunities
- Information overload fuelled by hyper-digitalisation
- Dislocation and instability linked to high mobility and a "floating" population
- Pressure from rigid societal expectations, particularly around success and social recognition
- A growing drive for individuality and self-expression

These tensions created a pervasive sense of pressure – both external and internal – shaping how young people navigate identity, ambition, and wellbeing.

*Figure 1. Grounding starts with understanding the macro forces and tensions facing our youth audience today.*

## *From Functional Benefit to Cultural Territory*

Within this high-pressure environment, the team explored how the idea of "breathability" could extend beyond its physical meaning.

Through in-situ community immersions – engaging directly with groups such as performers, athletes, and highly appearance-conscious youth – the team uncovered that "breathability" operates on two levels:

- Physical: Relief from scalp discomfort, oiliness, and suffocation
- Emotional: A symbolic release from pressure, stress, and the constraints of everyday life

This duality unlocked a crucial reframing: CLEAR's role could evolve from solving a hygiene problem to enabling a moment of release, reset, and regained vitality.

*Figure 2. Building intimacy with our target communities and bring to life tension stories and scenarios.*

## *Embedding Cultural Expression Through Social Signals*

Social trend analysis further sharpened this territory by identifying how these tensions were being expressed in contemporary culture.

For example, youth discourse increasingly reflected states of emotional numbness through phrases such as:

- "活人感"(feeling alive)
- "活人微死"(a subtle sense of lifelessness within routine)

These expressions captured a powerful insight: beneath the pressure and routine of daily life lies a deep desire to *reclaim sensitivity, vitality, and emotional presence*.

This led to a central insight articulation:

*In a culture of numbness, I fight to feel alive again by sharpening my senses and reclaiming the vitality that routine has drained.*

*Figure 3. Social scanning for trending youth cultural expressions and language for inspiration.*

## *From Insight to Creative Springboards*

Rather than landing on a single insight statement, the sprint generated a set of several validated insight platforms – each offering a distinct but coherent direction for creative development:

- Casting Off Constraints: Tuning out noise and breaking free from expectations
- Unapologetically Alive: Reclaiming vitality and embracing intensity
- The Calm Rebellion: Maintaining control within chaos
- Unnoticeable Comfort: Experiencing effortless confidence and ease

Each platform translated the idea of "breathability" into emotionally meaningful territories that could be expressed creatively in different ways – while remaining rooted in real consumer tensions.

| INSIGHT PLATFORMS | INSIGHT SPACES |
|---|---|
| A. CASTING OFF CONSTRAINTS | 1. BREATHE – LEAVE YOUR INHIBITIONS BEHIND<br>2. TUNE OUT THE NOISE |
| B. UNAPOLOGETICALLY ALIVE | 1. INHALE LIFE, EXHALE NUMBNESS<br>2. MY COMPETITIVE EDGE: SAVOURING THE MOMENT<br>3. INFINITE POSSIBILITIES |
| C. THE CALM REBELLION | 1. THE NEW ORDER OF CHAOS<br>2. JUST ONE BREATH AWAY |
| D. UNNOTICEABLE COMFORT | 1. OPTIMAL SCALPS GO UNNOTICED |

*Figure 4. Four potential platforms were identified, each underpinned with one to three insight spaces which are consumer-inspired narratives that can serve as creative springboards.*

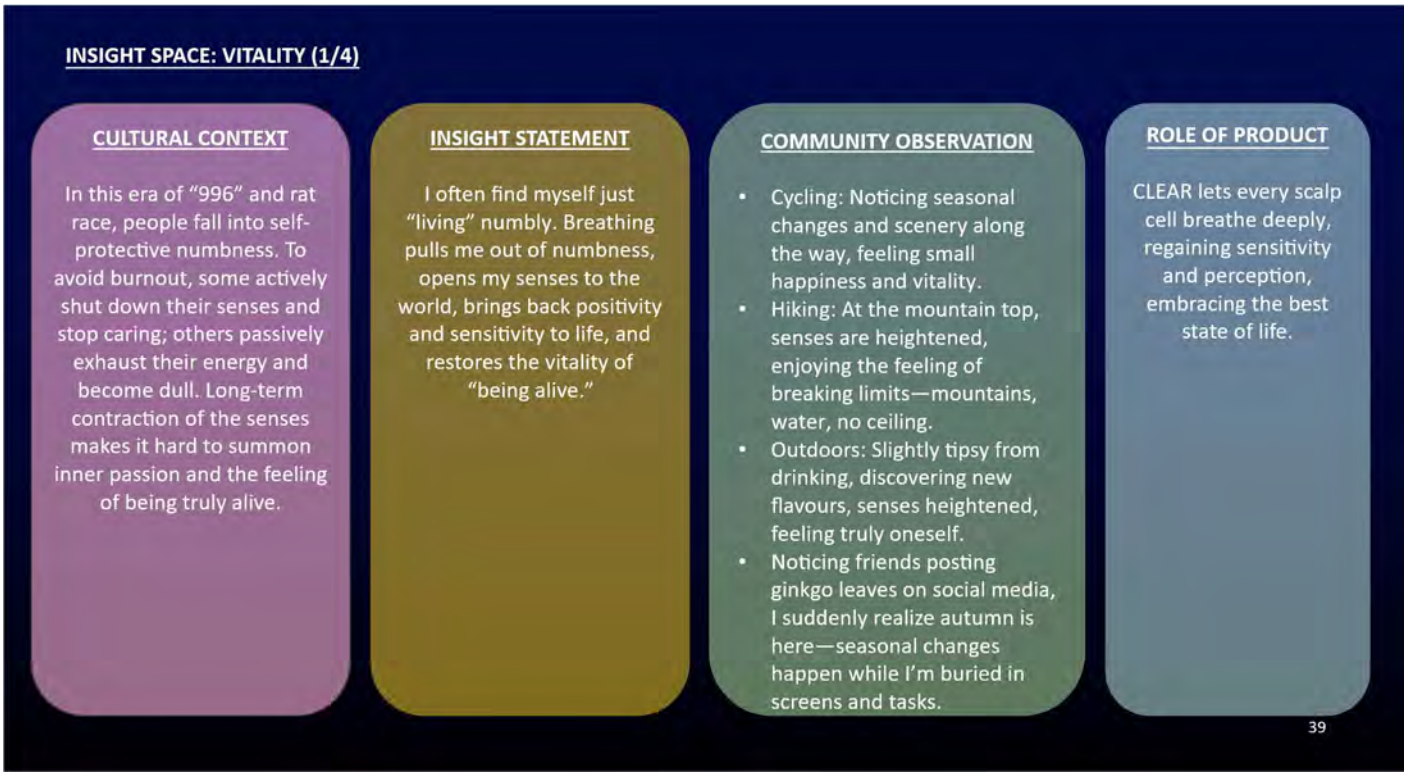



*Figure 5. An example of an insights space and role for CLEAR, validated with underlying cultural context and community stories heard.*

These outputs were then translated into the final creative communications brief, giving the agency sharper, validated springboards to inform ideation and development of the final creative works.

## Indonesia: Demonstrating Scalability and Cultural Adaptation

The same Insights Activator process was subsequently applied in Indonesia, validating its ability to scale across markets.

While cultural expressions differed, the underlying principle remained consistent: linking the functional benefit of breathability to locally relevant expressions of identity, resilience, and performance.

For example, insight articulation in Indonesia reframed the idea of strength and consistency:

Real Indonesian men cultivate deep roots for unshakable strength – valuing expertise, consistency and long-term impact over short-term gains.

This demonstrates how the same process yields culturally specific outputs, rather than forcing a one-size-fits-all global insight.

## Creative Impact and Business Outcomes

The most significant impact of the Insights Activator was observed at the point of creative execution.

Across both China and Indonesia:

- Creative ideas developed from validated springboards achieved strong pre-testing performance at first execution

- The need for multiple rounds of rework with agency partners was significantly reduced
- Time-to-market was accelerated, enabling faster deployment of campaigns

Consumer feedback further validated the strength of the approach:

- Ideas were perceived as highly relevant and reflective of real life
- The concept of "breathing from the scalp" was experienced as novel and emotionally engaging
- Messaging was seen as authentic, relatable, and grounded in real tension

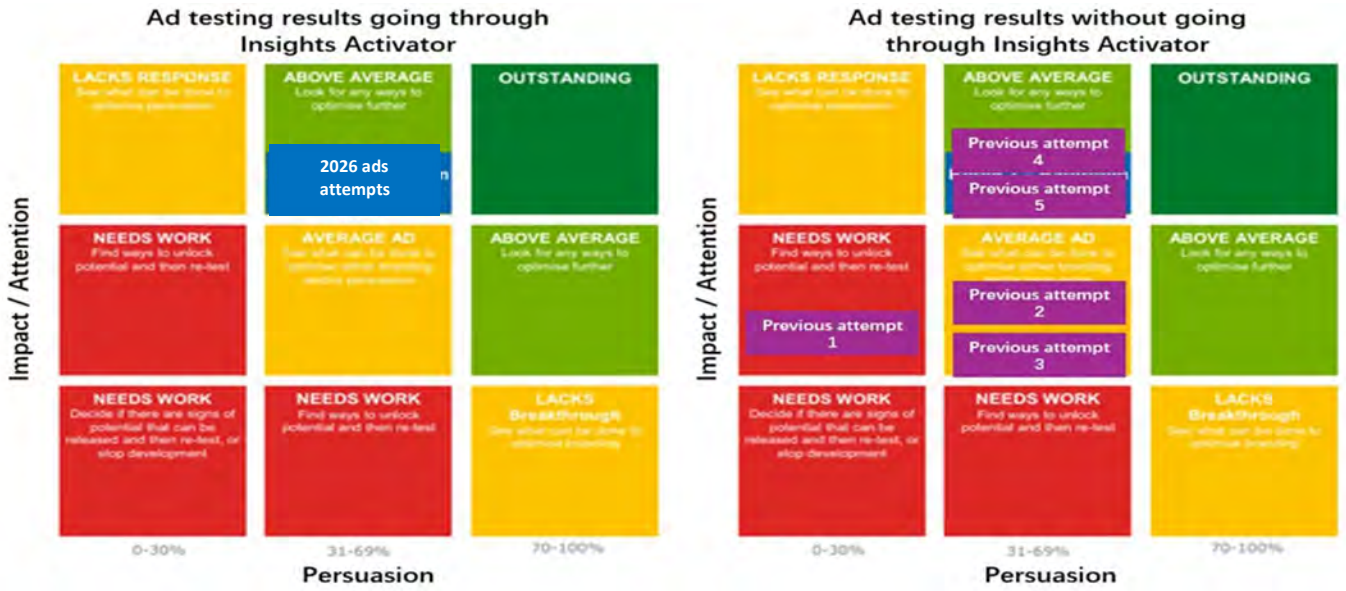



*Figure 6. Brand campaigns developed after the Insights Activator process did well in the pre-testing protocols at first try, compared to previous attempts where the outcomes are more mixed.*

Crucially, this marked a reversal of previous patterns where campaigns struggled to achieve strong results despite multiple iterations.

## What Makes This a True Metamorphosis

Insights Activator is not about a new tool – it represents a shift in how insights are created and used.

It transforms:

- Creation: from static to iterative and collaborative
- Role of culture: from background context to core driver
- Brief development: from assumption-based to validation-led
- Creative process: from trial-and-error to guided exploration
- Speed and efficiency: from repeated cycles to first-time success

## Key Learnings for the Insights Community

Several critical learnings emerge from this case:

Copyright © **ESOMAR** 2026
Congress 2026

1. Insights must be co-created, not assumed
2. Validate the articulation, not just the idea
3. Move from single insights to platforms
4. Design processes, not just outputs
5. Embed insights earlier in the creative journey

## Conclusion: Evolving the Craft of Insights

In today's fast-moving cultural landscape, traditional approaches to insight development are no longer sufficient.

Insights Activator represents an evolution in the craft – shifting insights from static inputs into active drivers of creative development. By connecting culture, consumer understanding, and validated articulation, the process enables more effective and consistent outcomes.

This is metamorphosis in practice: a transformation that elevates insights from informing decisions to shaping creative success.

## About the Authors

YeeMei Chan, Regional Director – Global Clients, Kantar, Singapore.
Lanny Li, Global Consumer Market Insights Lead for CLEAR, Unilever, Singapore.

# Decoding the Gold Standard

## From Niche Perspective to Sophisticated Balance: How AB InBev and Ipsos Decoded the Wellness Mainstream

*Marcela Ayres*

*Gustavo Araujo de Castro*

*Fernanda Wajchenberg*

*Tiago Fernandes*

## Esomar

Office address:

Burgemeester Stramanweg 105

1101 AA, Amsterdam

Phone: +31 20 664 21 41

Email: info@Esomar.org

Website: www.Esomar.org

Publication Date: September 2026

ESOMAR Publication Series Volume S424 Congress 2026

ISBN 92-831-0337-8

Congress 2026
Copyright © **ESOMAR** 2026

Copyright © **ESOMAR** 2026


# Decoding the Gold Standard

## From Niche Perspective to Sophisticated Balance:

## How AB InBev and Ipsos Decoded the Wellness Mainstream

*Marcela Ayres*

*Gustavo Araujo de Castro*

*Fernanda Wajchenberg*

*Tiago Fernandes*

### 1. Introduction: Permission Without Sacrifice

*Decoding the Brazilian Paradox of Wellness and Indulgence*

The global beverage industry is currently navigating one of its most profound metamorphoses. As consumers increasingly prioritise holistic wellbeing, traditionally indulgent categories, such as the beer category, face a complex challenge. The modern consumer demands "better-for-you" alternatives, yet they emphatically refuse to compromise on taste, heritage or the emotional reward of indulgence.

This macro-environmental shift is vividly captured by digital consumer behaviour. Social intelligence reveals a fundamental pivot in how consumers frame their beverage choices: while general search interest for "beer" in Brazil has experienced a 7% decline over the past 48 months, health-driven queries are surging exponentially. Driven by a more intentional and informed decision-making process, consumers are actively researching alternatives, resulting in a 46% increase in "healthy beer" queries. They are no longer passively accepting traditional choices; they are aggressively looking for functional benefits, evidenced by a 71% spike in searches for "beer calories", a 164% jump for "beer zero", and a staggering 306% surge for "beer gluten".

In Brazil, this challenge is amplified by a unique cultural duality. The country is a global leader in body optimisation and aesthetics, yet much of its social life still revolves around sharing meals, celebrating together and connecting over barbecue (the churrasco) and beer. In this context, beer is the ultimate "social lubricant". Asking a Brazilian consumer to drink a "clinical" or purely functional beer is not just a dietary choice; it risks feeling like an exclusion from the social ritual itself. While social listening confirms consumers are actively seeking "balance" between fitness goals and these social traditions, they face intense cultural headwinds. Digital conversations reveal a stark generational shift, with Gen Z and Millennials pulling away from traditional alcohol, while healthcare professional (HCP) influencers increasingly dominate online spaces by linking beer to negative fitness outcomes. This creates a subconscious friction and a pressing need for a product that bridges the gap between wellness and social indulgence.

This paper explores how Stella Artois, a 600-year-old brand synonymous with sophistication, navigated this exact socio-cultural tension. By pivoting away from linear research and adopting an agile, iterative learning ecosystem, AB InBev and Ipsos successfully transformed a struggling, niche line-extension into a mainstream wellness icon. With the new launch, we re-energised the brand, and innovation drove its growth, putting it back in the spotlight.

By continuously listening, learning and adapting, and by combining human-centric behavioural science, real-world social intelligence and advanced Artificial Intelligence (AI), this project set a new global benchmark for continuous product innovation.

## 2. From Functional Niche Trap to Premium Icon

*How Continuous Learning Pivoted a Failed Innovation into Brazil's Most Successful Wellness Beer*

To truly understand the metamorphosis of Stella Artois, we must first acknowledge a difficult reality and embrace vulnerability. In 2019, recognising the nascent wellness trend, AB InBev launched its first line extension for the brand in Brazil: Stella Artois Low Gluten. The initial premise was logical – offer a product that addressed the growing consumer desire for lighter alternatives. Early assessments revealed a highly differentiated proposition. Yet differentiation alone was not enough; the brand still needed to establish meaningful relevance in consumers' lives. However, the execution fell into what we define as the "Functional Niche Trap".

The brand attempted to impose a functional logic onto a market that historically purchased Stella Artois for aspiration and indulgence. Initial market monitoring and pilot testing in a test market revealed critical barriers. Consumers strongly associated the "Low Gluten" claim strictly with medical dietary restrictions or coeliac disease, failing to connect it with an aspirational lifestyle.

*Figure 1. Golden foil with the "Low Gluten" claim. Front label with zoom, highlighting "Pure Malt" and "Low Gluten" on a gold background.*

The metrics from this 2019 pilot phase were a harsh reality check:

- Lack of Mental Availability: A staggering 92% of consumers failed to recall point-of-sale marketing activations. The messaging simply did not penetrate the consumer's selective attention.
- Polarised Relevance: The product and packaging successfully signalled differentiation but struggled to broaden appeal. The prominent "Low Gluten" claim reinforced perceptions of a niche dietary product rather than an aspirational Stella Artois experience, contributing to low mental availability and weak consumer advocacy.
- Zero Advocacy: The Net Promoter Score (NPS) during the pilot phase was at an alarming 0.

As an AB InBev executive reflected: "We realised we were answering a question our mainstream consumer wasn't asking." The company faced a profound challenge: How could a legacy brand pivot a "functional exclusion" product into an aspirational "mainstream wellness" icon representing sophisticated balance?

## 3. Methodology: A Groundbreaking Ecosystem of Agile Intelligence

To solve this paradox, AB InBev and Ipsos discarded traditional, linear research models. In accordance with agile principles, we created a continuous feedback loop: launching in a test market, meticulously tracking consumer reactions, pivoting the idea and improving the offer.

This monumental dataset comprised exactly 10 research projects conducted over 8 years, completing 6,289 robust interviews. The research architecture was divided into three advanced analytical pillars.

### *3.1. Mapping the Wellness DNA via Choice-Based Conjoint Analysis*

Traditional surveys often fail to capture true consumer desires because they do not force trade-offs. To identify the authentic drivers of choice in the premium wellness segment, we utilised a comprehensive Choice-Based Conjoint Analysis. We tested 18 distinct claims across 6 different brands. The data yielded a counter-intuitive but foundational insight: the attribute "Gluten Free" was a weak driver of overall relevance. Conversely, attributes such as "Low Calorie", "Pure Malt" and "Natural Ingredients" significantly outperformed gluten-related claims.

### *3.2. Behavioural Science: The Power of Implicit Response Latency*

To validate the Conjoint findings and understand the subconscious friction of the original product, we applied Behavioural Science principles, focusing on System 1 (fast, intuitive) versus System 2 (slow, rational) cognitive processing. We measured not just what consumers chose, but how fast they made their choice, using Response Latency metrics.

Our thesis was simple: a fast win shows conviction in innovation adoption, while a slow win indicates hesitancy and cognitive dissonance. By measuring decision speed in milliseconds, we quantified the implicit friction caused by the "Low Gluten" positioning. The data proved that "Gluten Free" forced consumers into a slow, System 2 rationalisation, whereas "Low Calorie" combined with "Pure Malt" triggered immediate System 1 aspirational acceptance.

### *3.3. AI Sentiment Calibration with Ethical Governance (Human-in-the-Loop)*

To decode the underlying emotional barriers, our implicit behavioural data was cross-referenced with a sophisticated, AI-driven sentiment calibration engine. We utilised Natural Language Processing (NLP) to parse and analyse over 20,000 open-ended, spontaneous consumer verbatims across digital ecosystems.

In our sentiment model, while positive sentiments incrementally lift trial KPIs, negative sentiments disproportionately suppress them. Crucially, this AI application was governed by a strict "human-in-the-loop" protocol. To prevent the AI from generating "cultural hallucinations", the prompts and analytical frameworks were continuously calibrated by human experts deeply embedded in Brazilian culture.

The true power of this human-machine collaboration became evident when the AI uncovered a subtle psychological barrier that human researchers hadn't explicitly isolated. While human analysts initially focused on general product

rejection, the AI mapped a pervasive layer of social vulnerability in the verbatims. It revealed that the prominent "Low Gluten" claim triggered an immediate association with illness, medical restriction or physical frailty.

The concrete insight discovered by the AI was that ordering the product in public settings, such as a casual get-together or a weekend churrasco, subconsciously signalled to peers that the drinker was "sick" or managing a health condition. Instead of feeling like a modern lifestyle choice, the beverage felt like a medical prescription. This public association stripped the brand of its core premium asset: effortless sophistication.

This precise output proved mathematically that the 2019 product's relevance was trapped in a restrictive, clinical niche archetype. By converting vast text volume into this hyper-specific cultural friction point, the AI provided the exact strategic baseline needed to pivot the brand's positioning toward mainstream wellness.

## 4. The Strategic Pivot: "Flavour Without Compromise"

Armed with these predictive insights, the strategy fundamentally shifted from a message of restriction to a message of aspirational choice.

- The Rebranding: The product was relaunched as Stella Artois Pure Gold. The name itself shifted the focus from a missing ingredient to the premium quality of the liquid.
- The New Core Message: The winning concept paired brand sophistication with the attribute of "Balance". The central thesis became: Flavour without compromise, which was translated into the claim "All the flavour of Stella, with fewer calories".
- Conversion Claims Hierarchy: We restructured the claims, connecting quality endorsers ("Pure Malt") directly to the high-conversion driver identified in our research: Low Calories.
- Repositioning Gluten: The low-gluten aspect was deliberately moved to a secondary benefit, ensuring the product appealed to its original niche base while aggressively expanding into the mass audience.
- Building Aspirational Associations: Beyond product innovation, Stella Artois strengthened the new positioning through strategic partnerships with tennis (a sport strongly associated with success, discipline, balance and premium lifestyles). This helped shift consumer perceptions from a product designed for dietary restrictions to one aligned with aspirational consumers seeking moderation without sacrificing enjoyment.
- Amplifying Through Communication: Initially focused on functional benefits like reduced calories and sophisticated packaging, the communication strategy for this innovation ultimately succeeded in driving its "sophisticated balance" positioning through celebrity partnerships and powerful storytelling. By anchoring our messaging in consumer insights and capitalising on the innovation's core strengths, we successfully nurtured the essence of the concept. This outcome underscores the value of adopting a more linear, integrated approach to our innovation and launch processes.

*Figure 2. Stella Artois Pure Gold label*

*Figure 3. Ipsos framework for innovation and creative process.*

## 5. Results and Outcomes: Transformative Business Impact

The relaunch represented a structural metamorphosis in brand performance, proving that agile pivoting based on decoded human needs drives massive commercial success.

- Advocacy Turnaround: The strategic pivot solved the relevance polarisation. The product achieved an outstanding Net Promoter Score (NPS) of 64 – a monumental leap from the 0% NPS recorded during the 2019 pilot, with 95% of trialists indicating a clear intent to repurchase.
- Explosive Volume Growth: In the first quarter following the strategic consolidation, Stella Pure Gold registered a staggering 160% increase in volume, driving a 20% advance in the company's overall premium segment.
- Brand Equity Elevation: The innovation reinforced Stella's premium positioning by driving substantial growth in both core Power and Meaningful metrics after the launch. This revamp significantly amplified consumer perceptions across key brand attributes – elevating the product's association with High Quality and transforming it into a personal badge that Highlights Myself. Ultimately, by generating a powerful lift in making Occasions More Special, the product successfully evolved from a functional alternative into an aspirational anchor for modern social rituals.

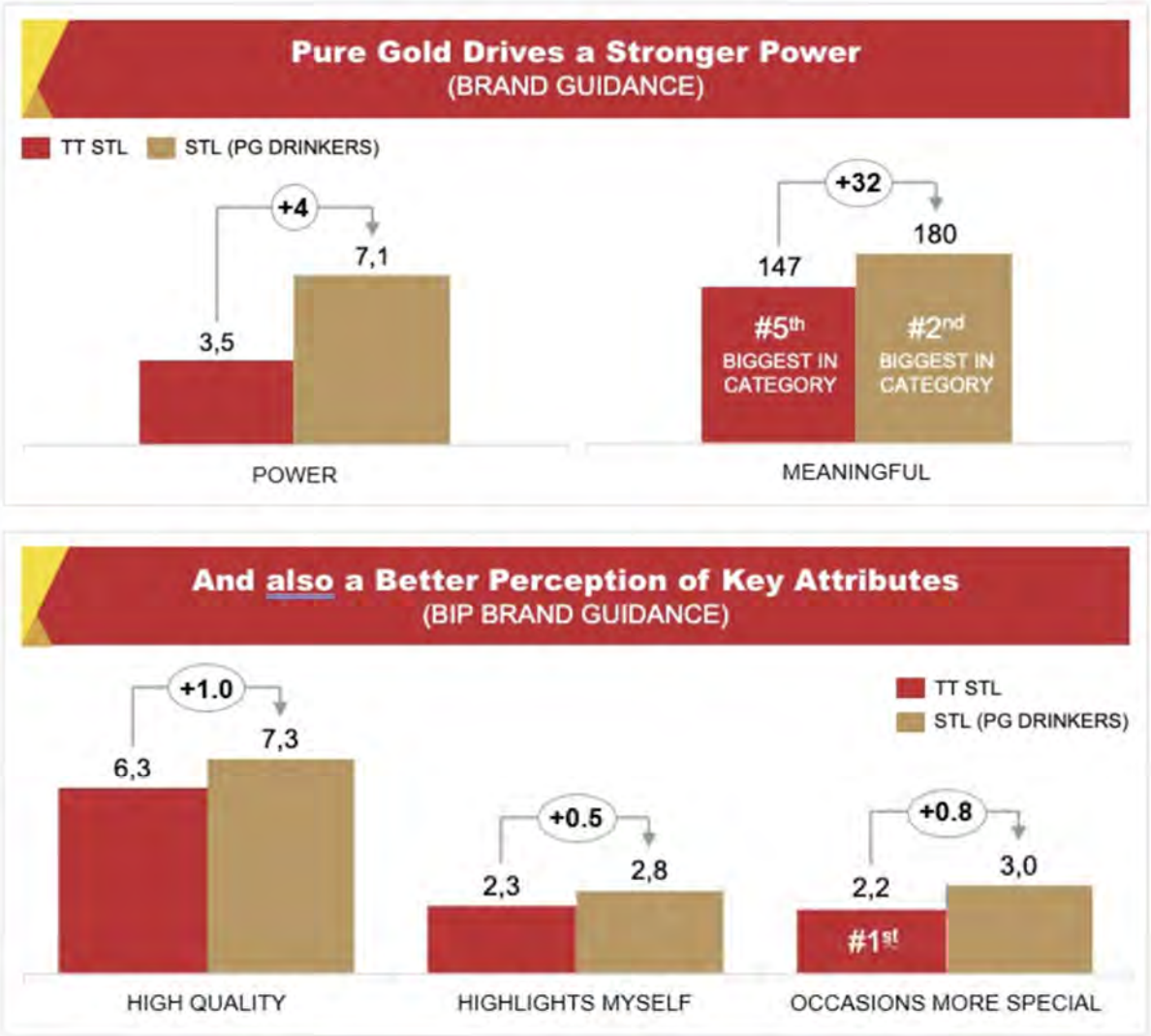



*Figure 4. Kantar Insights Brand Tracking Analysis H1'25 (Shared by ABI).*

- High volume and strong engagement across social listening platforms demonstrate how Stella Artois Pure Gold has successfully repositioned beer consumption into a choice without trade-offs. The innovation stepped in to

solve a universal human need: the desire to enjoy a premium beer during meaningful social moments without consequences. By seamlessly combining lightness, wellbeing and social status, Pure Gold has completely redefined what it means to "drink well". This cultural shift is vividly captured in real-world consumer verbatims:

- The Wellness Shift: "I'm obsessed with Stella Pure Gold. Now it's the only one I drink. I don't get a hangover at all." – Post on X
- The Premium Lifestyle Badge: "Inspired by Stella Pure Gold... fashion capsule collection" – Post on Instagram
- The Ultimate Paradigm Shift: "I really think Stella Pure Gold is the Coca-Cola Zero of beers." – Post on X

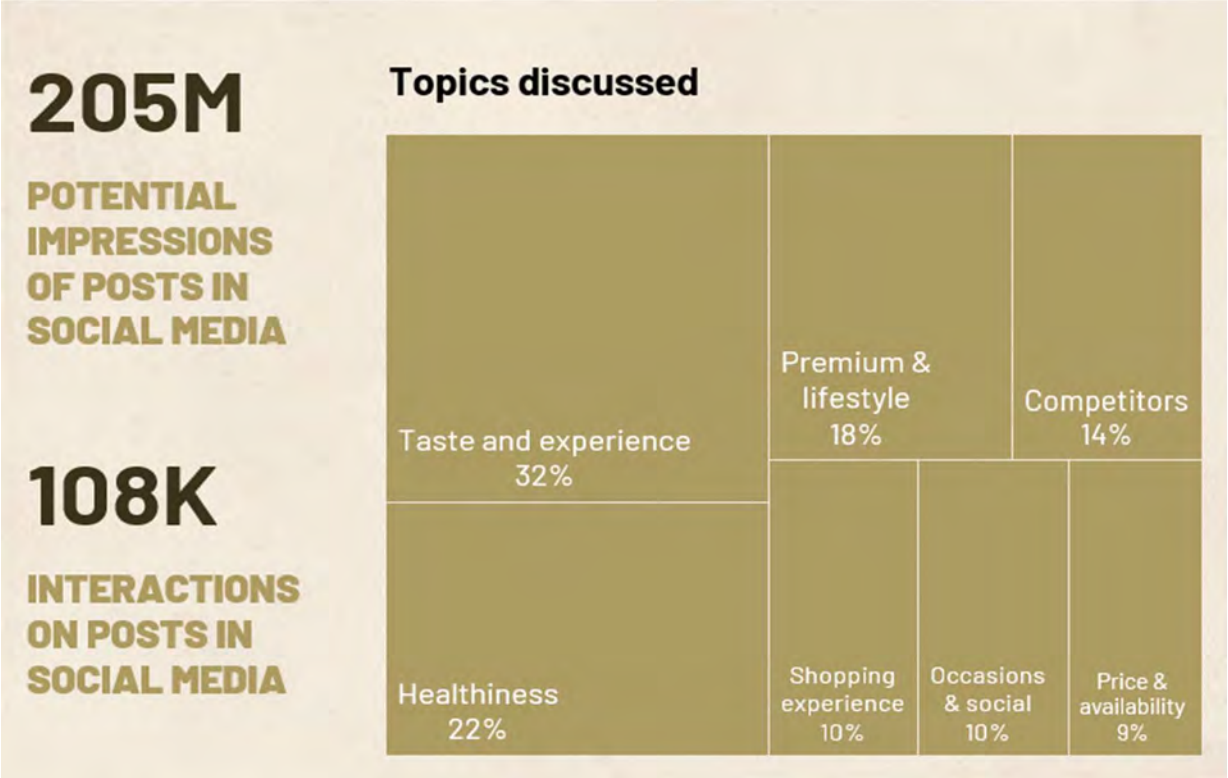



*Figure 5. Ipsos Social Listening Analysis (24-Month Trend Data)*

- Market Domination: The InfoMoney report highlights that Ambev achieved 17% growth in its premium brands during Q4 2025, capturing 100% of the entire premium segment's growth in Brazil for that period. Crucially, the financial markets explicitly highlighted Stella Artois Pure Gold alongside Spaten and Corona as the core engines driving this historic 19-quarter premium growth streak.

## 6. Beyond Stella Pure Gold: The Rise of Continuous Innovation

To consolidate the learnings generated over nearly a decade of continuous research, the project has now entered its most advanced phase. We have developed a national monitoring programme designed to understand not just the success of Stella Pure Gold, but the broader cultural transformation of wellness beer consumption. This predictive intelligence platform focuses on:

- Real ROI Measurement: Quantify the incremental value created by Stella Pure Gold through an integrated analysis of consumer, sales and penetration data. This provides a clear view of the brand's contribution to portfolio growth, and whether performance is exceeding the investment required to support the extension.

- Blue Ocean Mapping: Understanding if the brand has reached its awareness ceiling or if there are new, untapped consumer segments.
- Occasion and Volume Dynamics: Identifying exactly where, when and why consumption happens, pinpointing which competitor brands or occasions are being actively replaced.
- Real-World Social Listening: Investigating organic consumer conversations to determine how moderation, healthiness and premium consumption are being embedded into daily life, proving this is a consolidated cultural habit rather than a fleeting trend.

The Pinnacle of Innovation: The journey does not end with launch. This journey marks a major shift in how we work: research is no longer just used to check if an idea works, but is now actively creating the ideas. This change has brought to life a "continuous innovation ecosystem".

The ultimate step is the incorporation of a Concept Generation Agent.

That will be fed by AI-based Insights from accumulated knowledge of the 10 studies, 6,200+ interviews and proprietary Ipsos frameworks; this agent acts as an active copilot. Its role is to continuously refine Stella Pure Gold's value proposition in real time, and to autonomously generate new innovation territories and line extensions aligned with the exact principles that drove this brand's historic success.

## 7. Implications for the Profession: Humanity at Scale

This case study serves as a roadmap for the future of the market research industry. The profound learning here is that the innovation funnel is no longer a linear path; it is a living ecosystem that continuously feeds back on itself. We began with unmet consumer needs and a claim that seemed powerful but was trapped in a niche; by listening, pivoting and repackaging with the right aspirational communication, we unlocked exponential growth.

The ultimate implication is that AI does not replace the human researcher. Rather, by automating the complex science of data parsing and predictive generation, it liberates our industry to focus on the art of strategic storytelling and cultural empathy. Research has evolved from a tool that simply measures the market into an active force that builds the next frontiers of growth. The success of Stella Artois Pure Gold was not the result of a better liquid; it was the result of a deeper understanding of what balance means in modern life. The future of insights does not lie in predicting what consumers will buy next, but in continuously decoding what they need before they can even articulate it themselves.

### *Three Core Lessons for the Insights Industry*

#### *Lesson 1: The Trap of the Wrong Frame*

The biggest market opportunities are often hidden right in front of us, masked by the wrong framing. Success requires the vulnerability to recognise when an insight is trapped in a functional niche and the agility to reposition it toward an aspirational lifestyle.

Copyright © **ESOMAR** 2026


### *Lesson 2: From Linear to Iterative*

Innovation is no longer a linear waterfall process of testing and validation. To win in fast-changing cultural landscapes, brands must shift toward an iterative, continuous innovation ecosystem where research actively drives the pipeline in real-time and connects the different stages.

### *Lesson 3: Scaling Observation vs. Creating Meaning*

AI scales observation, but humans create meaning. While advanced algorithms and NLP can compress weeks of data parsing into hours and uncover hidden semantic patterns, it takes human empathy and cultural context to turn those data points into a powerful brand strategy.

## About the Authors

Marcela Ayres, Account Director in Innovation, Ipsos, Brazil.
Gustavo Araujo de Castro, Strategy & Insights Director, AB InBev, Brazil.
Fernanda Wajchenberg, Head of Innovation in LATAM, Ipsos, Brazil.
Tiago Fernandes, Senior Client Officer, Ipsos, Brazil.

# From Tool to Teammate

## Embedding AI into the Everyday Workflow of Insights and Marketing Teams at Philips

*Mathilde Beljaarts*

*Fenny Léautier*

## Esomar

Office address:

Burgemeester Stramanweg 105

1101 AA, Amsterdam

Phone: +31 20 664 21 41

Email: info@Esomar.org

Website: www.Esomar.org

Publication Date: September 2026

ESOMAR Publication Series Volume S424 Congress 2026

ISBN 92-831-0337-8

Congress 2026
Copyright © **ESOMAR** 2026

Copyright © **ESOMAR** 2026


# From Tool to Teammate

## Embedding AI into the Everyday Workflow of Insights and Marketing Teams at Philips

*Mathilde Beljaarts*

*Fenny Léautier*

### 1. Introduction

Artificial intelligence is transforming the insights industry. Yet despite unprecedented investment in generative AI, most organisations remain stuck in experimentation mode. The challenge is not access to technology. AI tools are now widely available. The challenge is embedding AI into everyday decision-making in a way that people trust, adopt and sustain.

At Philips Personal Health, we discovered that successful AI transformation is not primarily a technology challenge. It is an organisational one.

Our starting question was simple:

How do we move AI from being an interesting tool to becoming a trusted teammate within insights and marketing teams?

The answer became the foundation of a new operating model built around what we call the Human–AI Adoption Triangle: Platform, Process and People. The model recognises that sustainable AI adoption requires AI, data and technology; workflows, governance and decision-making; and human capability, coaching and culture to evolve together. While many organisations focus primarily on deploying technology, our experience shows that value is created only when all three dimensions are intentionally developed in parallel.

To enable this shift, we developed illuminAI, a proprietary AI platform that supports insights generation, AI-assisted ideation, and concept testing and optimisation through synthetic consumer personas built on proprietary Philips data.

But the real innovation was not the tool itself. The breakthrough came from embedding AI into the way teams work.

We introduced a new operating model built around three principles: structured AI-assisted workflows, human-in-the-loop governance, and large-scale capability building. Through bootcamps, Super Clinics, capability building through the Philips Marketing Academy and integration into processes, teams learned how to use AI platforms & tools, to challenge outputs, co-create ideas, and translate results into consumer centric innovations and actionable marketing plans.

This shift turned AI from an experimental technology into an everyday collaborator in the innovation and marketing planning process. Instead of replacing researchers or marketers, AI now acts as an extra team member, helping teams move from consumer insights to innovation concepts faster and more consumer-centric and building stronger consumer-focused marketing plans.

The results are significant. We see up to 50% faster time-to-insight & concept and 20–30% reductions in research costs, with improving quality of marketing briefs and innovation concepts.

This paper argues that the future of insights is not about adopting AI tools but about designing organisations where humans and machines learn and work together. The Philips experience offers a practical blueprint for insights leaders seeking to operationalise AI responsibly and effectively.

The following presents Philips Personal Health's journey from AI experimentation to an operational human–AI insights workflow. We illustrate how technology, capability building, and governance were combined to embed AI into everyday insight and innovation practices at a global scale.

## 2. The AI Adoption Gap in Consumer Insights

Generative AI has triggered unprecedented excitement across the insights industry. Organisations now have access to tools capable of synthesising research, generating ideas, simulating consumer responses, and producing innovation concepts.

Yet many companies are discovering that deploying AI tools does not automatically create value.

According to Bain and McKinsey, the majority of organisations experimenting with generative AI remain stuck in pilot mode, with only a minority achieving meaningful productivity gains (2026). The problem is not the technology itself, but the lack of clear operating models that integrate AI into everyday work.

For insights teams, this challenge is particularly acute. Traditional research workflows are built around sequential processes: research design, fieldwork, analysis and reporting. AI fundamentally changes this dynamic by enabling continuous iteration between insights, ideation and testing.

The question therefore becomes:

How do we redesign insights workflows so humans and AI can collaborate effectively?

## 3.The Human–AI Adoption Triangle

As we moved from experimentation to adoption, we realised that technology alone could not drive the transformation. Sustainable AI adoption required three dimensions to evolve simultaneously: Platform, Process and People.

While many AI adoption frameworks focus primarily on technology deployment, the Human–AI Adoption Triangle places equal emphasis on workflow redesign and capability development. Our experience suggests that lasting business impact emerges only when all three dimensions are intentionally developed in parallel.

### *Platform: illuminAI, the technology foundation.*

This includes proprietary data, purpose-trained language models, synthetic personas, and the illuminAI platform itself. The objective is to provide teams with trustworthy AI capabilities that are relevant to their business context.

*Process: the workflow foundation.*

AI was embedded into existing insights, innovation and marketing processes, from insight generation and ideation to concept testing and marketing planning. Human-in-the-loop governance, standardised briefing templates, testing frameworks, and decision rights ensured that AI became part of how work gets done rather than an isolated tool.

*People: the capability foundation.*

Bootcamps, Super Clinics, coaching, and Marketing Academy integration ensured teams developed the skills and confidence required to collaborate effectively with AI. New roles emerged, including AI champions, coaches, and intelligence orchestrators who help teams translate AI outputs into business decisions.

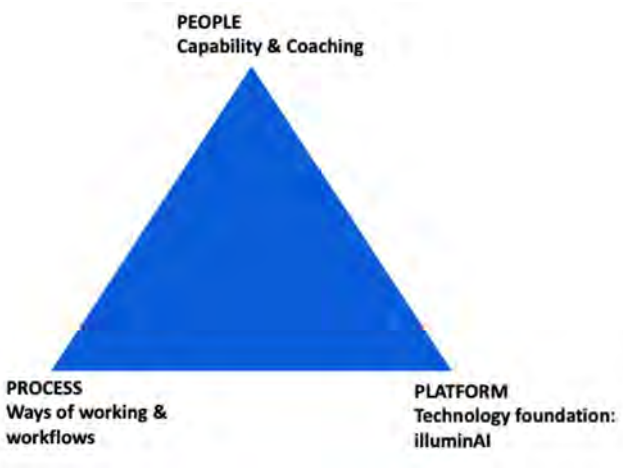



*Figure 1. The Human–AI Adoption Triangle.*

The central insight is simple:

Sustainable AI adoption happens when Platform, Process, and People evolve together. Overinvesting in one dimension while neglecting the others limits impact. The Philips approach intentionally balanced all three.

## 4. Our Transformation Journey

While the Human–AI Adoption Triangle became the framework that guided our transformation, the journey itself evolved through several distinct phases. Each phase generated important learnings that shaped the next step, gradually moving Philips from isolated AI experiments to a scalable human–AI operating model.

| Phase | Objective | Key Learning |
|---|---|---|
| **Experimentation (2024)** | Explore the potential of generative AI across insights and marketing use cases | AI generated curiosity, but usage remained fragmented and inconsistent |
| **Early Pilots (H2 2024)** | Test external AI solutions and identify business applications | Access to AI alone does not drive adoption |

| | | |
|---|---|---|
| **Human–AI Adoption Triangle (2024–2025)** | Develop a scalable adoption framework | Sustainable transformation requires Platform, Process and People to evolve together |
| **Capability Building (Q4 2025–2026)** | Build confidence through bootcamps, Super Clinics and Marketing Academy integration | Trust and capability are as important as technology |
| **Embedded Workflow (Q4 2025–2026)** | Integrate AI into insight generation, ideation and concept development processes | AI adoption accelerates when embedded into everyday work |
| **Scale-Up (March 2026 onwards)** | Global launch at our yearly Global Innovate to Win event and deployment across Business Units and Regions | Human–AI collaboration can be scaled through structured ways of working |

*Table 1. The Philips Human–AI Transformation Journey.*

The transformation from AI experimentation to large-scale adoption did not happen overnight. It evolved through a series of deliberate steps that gradually moved AI from a promising technology to an embedded way of working.

Importantly, each phase informed the next, reinforcing the need to evolve Platform, Process and People together rather than sequentially.

## 5. Rethinking the Role of Insights

Historically, insights teams have played a relatively defined role within organisations: synthesising and resighting previous research, designing new research, generating insights, and delivering reports that inform business decisions.

AI disrupts this model in two important ways.

1. First, it dramatically lowers the barrier to accessing information. Marketing teams can now generate hypotheses and explore consumer questions independently.
2. Second, AI expands the creative possibilities of insights work by enabling co-creation between humans and machines as AI can be a powerful source of inspiration

This shift requires insights professionals to move from being producers of insights to orchestrators of intelligence and ideas, combining multiple sources of knowledge, including AI-generated insights, ideation, synthetic persona conversations, and traditional research.

At Philips Personal Health, this realisation led us to rethink how insights and marketing teams collaborate.

Rather than introducing AI as a standalone tool, we redesigned the workflow itself.

## 6. Platform: Building the Technology Foundation

IlluminAI was not designed to replace research. It was designed to accelerate the journey from consumer understanding to action. IlluminAI supports three critical moments in the innovation process:

1. AI insight generation
   Synthesising existing knowledge from thousands of consumer studies, trend reports, social and R&R conversations, and historical learning. Drawing on trusted proprietary datasets and curated sources to build a structured understanding of consumer needs, pain points, tensions, and trends. IlluminAI has over 6,000 consumer research reports, more than 5,000 trends reports, 250,000 Reddit Posts and more than 3,000,000 Ratings & Reviews to draw from and sources are continuously being added.

2. AI-assisted ideation
   Transforming insights into benefits, ideas, claims and innovation concepts through structured AI-assisted ideation. A guided process that helps teams translate consumer insights into benefits, ideas, headlines and full concepts. This structured ideation model ensures that creativity remains grounded in real consumer needs rather than intuition alone.

3. Challenging with synthetic personas
   Stress-testing and optimising concepts through conversations with synthetic consumer personas grounded and trained on our proprietary Philips data on our target consumers. There are currently 84 personas across Business Units and Regions, and we are adding where there is a need.

By leveraging curated data sources, built-in guardrails, and specialised models, illuminAI helps teams create better consumer-centric innovations faster, improve decision-making, and reduce the risk of AI hallucinations while maintaining a strong focus on real consumer needs and behaviours. Together, these three capabilities enable teams to move continuously between understanding, creating and challenging ideas—rather than treating these as separate stages.

## 7. People: Strengthening the Capability Foundation

Many users initially viewed AI as either intimidating or irrelevant to their daily work. Capability-building therefore focused as much on confidence and trust as on technical skills.

The cornerstone of the Philips approach is a human–AI workflow that integrates insights, ideation, and testing into a continuous cycle. However, the most important aspect of this workflow is that humans remain firmly in control.

AI generates possibilities. Humans interpret, challenge, regenerate, edit and decide.

AI capability-building is important to make sure everyone is not only able to use the tools, but, more importantly, to use the tools in the right way: how to judge and check the outcomes, how to value generated ideas, how to optimise further and what to take forward. We needed to create the mind shift from tool to teammate!

Three initiatives proved critical to enhancing these capabilities across the global organisation:

### *Bootcamps*

Hands-on workshops where teams learn how to use the platform and, more importantly, how to integrate AI into real business challenges, from insight generation to concept creation to immediate optimisation.

### *Marketing Academy Integration*

IlluminAI training was embedded into Philips' internal Marketing Academy to ensure long-term capability development. Dedicated AI-applied sessions show how illuminAI can support in developing marketing objectives, help write insights, identify benefits, ideas, headlines and full concepts, and is trained on concrete business challenges and ambitions on a global category level, as well as applied in category-level marketing planning in our key markets.

### *Super Clinics*

Live coaching sessions enable subject-matter experts to guide teams through the human–AI workflow while working on real projects. This hands-on approach significantly accelerates the adoption of illuminAI, helping teams experience firsthand how AI tools can enhance project outcomes while keeping consumer-centricity at the core.

Super Clinics can focus on innovation challenges as well as go-to-market approaches, illuminAI has the (AI) capabilities to support both objectives.

## 8. Process: From Tool to Way of Working

Introducing and gaining some initial hands-on experience with illuminAI was only the first step. We quickly discovered that adoption would not happen automatically. Many teams were curious about AI but unsure how to use it effectively within their existing work(flows). They were willing to try it out, but without clear guidance on how and when to use it, it was hard for some to integrate it in their daily way of working.

To address this challenge, we focused on embedding AI into daily routines, product development and decision-making processes.

Integration of AI in the Consumer Testing Framework and VPC Process

The people-centered testing framework used within Philips PH provides guidelines for marketing & consumer insights teams on when to test what and how. This framework is continuously reviewed and updated to reflect the latest tools & thinking and to make sure there is a good balance of AI tools/initiatives and traditional consumer research and real consumer connection along the product development cycle.

Throughout the product development process, we have indicated where AI tools & initiatives can be used. To stimulate adoption, together with the BUs we review the global consumer research programs on a quarterly basis to understand where AI has been used, where not and why and how to drive further adoption by also indicating the benefits and showing progress across teams. This helps us understand experiences, trigger adoption and when needed the framework can be adjusted.

### *Integration of AI in the Value Proposition Creation (VPC) Process*

The capabilities of illuminAI expand far beyond just synthesising information. The platform can play a key role throughout the VPC process, for insights, ideation and optimisation of ideas and concepts. Using illuminAI can give a kick-start to a VPC process, accelerate the process and improve the quality of outcomes.

Together with our largest Business Unit, we defined how illuminAI can be the best teammate along the different stages of the process and how it can be complemented by other AI tools and agents available within Philips.

These initiatives helped transform AI from a novelty into a practical, repeatable way of working.

## 9. Creating a Consumer Concept with illuminAI

One example came from an already launched Philips Personal Health product, where the team wanted to explore how the existing value proposition could be sharpened for future communication and innovation opportunities. The challenge was to move quickly from existing consumer knowledge to stronger benefit territories, claims and concept routes, while ensuring that any new direction remained grounded in real consumer needs and product experience.

Traditionally, this would have required the team to manually review previous consumer studies, identify the most relevant insights, align internally on priority consumer tensions, develop benefit territories and then translate these into early concept ideas. This process is valuable, but it can be time-consuming and dependent on the availability of researchers, marketers and historical knowledge owners. It also risks recreating work that already exists across different studies, markets or business units.

Using illuminAI, the team started by synthesising existing proprietary knowledge around the target consumer, category pain points, product experience and relevant trends. The platform then supported the translation of these insights into potential benefits, claims, headlines and concept routes. AI-assisted ideation helped expand the range of possible directions, while the structured workflow kept the ideas anchored in consumer understanding rather than purely creative intuition.

The team then used synthetic personas to challenge the relevance, clarity and distinctiveness of the emerging concepts from the perspective of priority consumer segments. This enabled a faster first read on which ideas felt most motivating, which language needed refinement and which consumer tensions were most worth building on.

Crucially, AI did not make the decision. The Insights and Marketing teams reviewed the outputs, challenged generic or less credible ideas, regenerated directions where needed and refined the language based on their category expertise and business context. Human judgement remained central in deciding which concepts were meaningful, ownable and commercially relevant.

The result was a faster and more structured concept-development process that helped the team move from existing knowledge to consumer-centric activation opportunities in a fraction of the time. IlluminAI accelerated the creation and refinement of concept routes, while human expertise ensured that the final outputs were relevant, distinctive and actionable. This example illustrated the core principle behind the Philips approach: AI generates possibilities, but humans interpret, challenge and decide.

What would traditionally have taken several weeks of desk research and alignment was compressed into a few structured working sessions.

Copyright © **ESOMAR** 2026


## 10. Lessons from Early Experiments

Like many organisations, our first assumption was that providing access to AI would naturally drive adoption. It did not. Early pilots revealed several barriers:

Teams were curious but uncertain how to integrate AI into existing workflows. Outputs were sometimes accepted too quickly without sufficient critical review. Different teams used different prompting approaches, resulting in inconsistent outcomes. Some users viewed AI as a threat to expertise rather than an enabler of expertise.

These early experiences reinforced an important lesson: AI adoption is not a technology challenge alone. Trust, governance and capability building are equally important. This insight directly informed the development of the Human–AI Adoption Triangle and the subsequent design of our operating model.

Ultimately, curiosity generated experimentation. Capability and workflow integration generated adoption.

## 11. Impact: Speed, Quality and Collaboration

Embedding AI into ways of working has produced measurable results. Results were measured across the first three months following global deployment in March 2026 at our yearly Global Philips Innovate to Win event and compared against traditional insight-generation and concept-development workflows used previously within Philips Personal Health.

In the first three months since its official launch, we have seen:

- up to 50% faster time-to-insight
- 20–30% reductions in research costs
- 53% repeat users for illuminAI
- questions asked: >6450
- >100 consumer-centric innovation concepts generated and refined through the illuminAI workflow
- improved collaboration between Consumer Insights and Marketing

Importantly, the platform is deployed globally across Philips Personal Health, supporting teams across all Business Units and Regions.

This global adoption ensures that the new human–AI workflow can scale across markets while remaining locally relevant.

## 12. Key Learnings

AI creates sustainable business value when organisations simultaneously invest in technology, redesign workflows, and build human capabilities – turning AI from a tool into an everyday collaborator.

- AI only delivers impact when it is embedded into new ways of working, not just new tools.
- Successful AI transformation requires three elements: Platform, Process and People.
- Embedding AI into daily workflows drives adoption.
- A human-in-the-loop model ensures AI accelerates insight generation while maintaining trust, quality and accountability.

- Capability building is critical, not just the tool but also the way of working.
- Proprietary data enhances AI value and creates a competitive advantage.
- AI can deliver measurable business impact: 50% faster, 20-30% lower cost.
- Insights teams must evolve from research providers to orchestrators of human and machine intelligence.

## 13. The Future Role of Insights Professionals

Perhaps the most significant impact of AI is the transformation of the insights profession itself.

Rather than replacing insights professionals, AI is expanding their role.

Insights leaders increasingly act as designers of intelligence systems, responsible for:

- combining multiple sources of data and knowledge
- guiding AI-assisted ideation and innovation
- ensuring ethical and responsible use of AI
- maintaining consumer understanding as the foundation of business decisions.

In this new model, the competitive advantage of insights teams lies not in access to tools, but in their ability to orchestrate human and machine intelligence effectively.

### *Implications for the Insights Industry*

Five lessons for organisations adopting AI in insights.

1. Tools are not enough. AI must be embedded into workflows.
2. Capability-building is critical. People must learn how to collaborate with AI.
3. Human oversight remains essential. AI should expand judgment, not replace it.
4. The consumer insights industry will continue moving from a research-centric model to an intelligence-centric model.
5. This creates an opportunity for the insights function to become a central orchestrator of human–AI decision making across innovation and marketing, rather than remaining a specialist research function.

The future of insights will not be defined by artificial intelligence alone. It will be defined by how effectively organisations combine human judgement, organisational processes and AI capability.

Our experience suggests that the most successful AI systems will not be those with the most advanced models, but those that become embedded in everyday ways of working. For insights professionals, this represents a profound shift. The role is evolving from research provider to orchestrator of intelligence—bringing together human expertise, proprietary knowledge and machine capability to accelerate better decisions.

The future is therefore not human or machine. It is human and machine, working together.

Copyright © **ESOMAR** 2026


## References

Bain Capital Ventures (January 2026). *Enterprise AI Adoption: from Pilots to real business value.* *https://baincapitalventures.com/insight/enterprise-ai-adoption-from-pilots-to-real-business-value/*

Databricks.com (November 2025). *AI Transformation: A Complete Strategy Guide for 2025.* *https://www.databricks.com/blog/ai-transformation-complete-strategy-guide-2025*

McKinsey (November 2025). *The state of AI in 2025.* *https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai*

World Economic Forum (June 2026). *The 5 faces of human readiness for AI adoption – and how to work with them.* *https://www.weforum.org/stories/artificial-intelligence/ai-workplace-adoption-readiness/*

## About the Authors

Mathilde Beljaarts, Superior Proposition Insights Leader, Consumer Experience Excellence (CXE), Philips Personal Health, Amsterdam.
Fenny Léautier, Consumer Insights & PH Academy Leader, Consumer Experience Excellence (CXE), Philips Personal Health, Amsterdam.

# How Estée Lauder is Transforming Historical Survey Data into New Insights Using Semantic Similarities

*Douglas Ethé-Sayers*

*Kristina O'Keeffe*

*István Hidegkuti PhD*

Esomar

Office address:

Burgemeester Stramanweg 105

1101 AA, Amsterdam

Phone: +31 20 664 21 41

Email: info@Esomar.org

Website: www.Esomar.org

Publication Date: September 2026

Esomar Publication Series Volume S424 Congress 2026

ISBN 92-831-0337-8

Copyright © **Esomar** 2026

Copyright © **Esomar** 2026

Congress 2026

# How Estée Lauder is Transforming Historical Survey Data into New Insights Using Semantic Similarities

*Douglas Ethé-Sayers*

*Kristina O'Keeffe*

*István Hidegkuti PhD*

## Introduction

Consumer test-based product benefit language is typically treated as subjective, requiring repeated testing if desired claim wording changes. While this ensures rigour, it can constrain teams to specific phrasing—even when alternative expressions may carry equivalent meaning. Over time, this has resulted in repeated testing of similar claims and the accumulation of fragmented data. This work explores how historical consumer research can unlock that value. By analysing consumer studies on 25 moisturizer products across the US and China, we test whether different benefit expressions behave similarly enough to be considered interchangeable. Combining statistical analysis with AI-enabled semantic modelling, we show that some descriptors generate comparable consumer responses and meaning. This suggests a new possibility: benefit language can be validated using existing data, enabling reduced redundancy, improved comparability across studies and more agile decision-making—while maintaining substantiation rigour.

## Background

At Estée Lauder Companies (ELC), product benefit statements are developed at the intersection of product performance, consumer perception, benefit language and legal substantiation. Because this work often begins one to two years before launch, teams must make decisions early about how a product benefit will be expressed, tested and ultimately supported. Today, benefit language is typically tested and vetted on a product-by-product basis, which has created a substantial body of consumer research data across brands, categories, markets and time. However, because each study is designed to support a specific product and a specific set of claims, the resulting evidence is often tied closely to the exact verbiage tested in that particular research. Once the product reaches market, consumer language, brand strategy, channel needs or cultural context may evolve, creating a need to articulate the same benefit in a new way. A benefit originally tested as "radiance" for example, may later need to be expressed as "glow", "luminosity" or "healthy-looking skin". This creates an important business, research and legal challenge: how can ELC preserve the rigour required for claims support, while allowing greater flexibility in how benefits are communicated?

## Research Question

*Can existing data determine whether benefit expressions can be used interchangeably, even if not tested together?*

Objectives:

- Test behavioural equivalency using historical survey data;
- Quantify linguistic similarity using semantic models;
- Identify interchangeable or standardised language candidates;

Copyright © **Esomar** 2026


- Develop a scalable framework for reuse across categories.

## Methodology

Data was drawn from home-use tests on 25 moisturiser products (13 US; 12 China), covering three benefit areas: radiance, moisturisation and lines/wrinkles. Benefit statements were measured via five-point efficacy scales and five-point Likert agreement scales at various time points throughout the usage period—immediately after first use, after one week of use and after four weeks of use. Benefit language and time points differed by test, with only some overlapping phrasing. Other common metrics across studies included hedonics, sensory performance, just about right (JAR) questions and open-ended comments. Table 1 shows the number of statements evaluated per benefit area per country, across time points.

| | US | China |
|---|---|---|
| **Radiance** | 77 | 66 |
| **Moisturisation** | 132 | 118 |
| **Lines/wrinkles** | 66 | 60 |

*Table 1*

The foundational assumption when testing the strategic interchangeability of different benefits is that if two benefits are truly interchangeable, they should exhibit identical or nearly identical behaviour while remaining close in meaning. Benefit language was evaluated using a three-pronged approach. Structural and distributional similarity were measured to examine benefit language behaviour at each time point. Linguistic similarity was measured via semantic modelling.

### *Structural Similarity*

Since most benefit statements were not present in the same historical tests, it was necessary to evaluate their structural similarity relative to another variable which was common to all tests. We evaluated how similarly the benefit statements behaved relative to this reference variable. Using a log-linear analysis (Agresti, 2013), we identified which statements behaved similarly to each other relative to this common reference variable.

### *Distributional Similarity*

We evaluated the internal properties of the benefit statements by testing whether their underlying response distributions were statistically homogeneous. If the statements share the same perception profile, the pattern of responses across the scale steps should remain consistent between them. We applied a Pearson's chi-square test of homogeneity (Agresti, 2013) to determine whether the benefit statements exhibited similar distributional characteristics.

### *Linguistic Similarity*

Our third test was to determine whether two benefit expressions are linguistically equivalent in meaning, beyond observable behavioral similarity. To quantify this, all possible statement pairs were evaluated across five theoretically motivated dimensions derived from feature-based semantic theory (Smith, Shoben & Rips, 1974):

- Perceptual meaning: Whether the two expressions describe the same core consumer outcome or skin benefit;
- Intensity and emphasis: Whether the implied strength or magnitude of the benefit is equivalent;
- Specificity: Whether the level of precision or granularity of the claim is aligned;
- Temporal framing: Whether the expressions imply similar timing of results (e.g., immediate versus gradual effects);

- Emotional and sensory impact: Whether the expressions carry comparable affective tone or experiential qualities.

For a given pair of benefit expressions, we generated structured semantic comparisons across these dimensions using a large language model (LLM) as a reasoning engine (Reimers & Gurevych, 2019). Interpretation of results follows a threshold-based framework:

- Fully interchangeable ≥ 0.90;
- Mostly interchangeable ≥ 0.80;
- Somewhat interchangeable ≥ 0.70;
- Related but distinct ≥ 0.50;
- Not interchangeable <.50.

By integrating this linguistic analysis with the prior statistical evaluations of relationship consistency and distributional homogeneity, we established a comprehensive, three-part framework for assessing benefit interchangeability. This final component ensures that even when two expressions behave similarly in observed data, they are also validated as conceptually aligned in meaning—thereby supporting both empirical and linguistic equivalence. By triangulating across all three, this framework minimises false positives and balances innovation with rigour—making it suitable for real world insights decision making rather than purely academic exploration.

## Key Findings

The analysis revealed consistent evidence that some benefit expressions—frequently treated as distinct—are interpreted similarly by consumers. Four broad patterns of interchangeability emerged.

### *Interchangeable Expressions*

Some statements showed high semantic similarity and no significant behavioural differences. These can be treated as interchangeable without additional testing, and represent immediate opportunities for reuse. For example, “skin looking brighter” could potentially be substituted with “skin looking more luminous” and “skin looking more radiant” in certain markets.

### *Behaviourally Similar, Semantically Moderate*

Some expressions behaved similarly but differed somewhat in meaning. These require case-by-case evaluation and may be candidates for conditional reuse. Regardless of whether they can be interchanged directly, they offer insight into the relationship between the expressions from the perception of the consumers. For example, “leaving skin with a fresh appearance” and “leaving skin looking more beautiful” are related, but not interchangeable. We see that freshness is related to beauty, which can be used to inform marketing or communications.

### *Behaviourally Similar, Semantically Different*

A third group showed strong behavioural similarity, but low semantic alignment. These are not interchangeable, but may reveal underlying perceptual relationships between claims. “Leaving skin looking less stressed/fatigued” and “leaving makeup looking fresh” are an example of such statements. The statements may not be interchangeable, but the concept of looking fresh is directly related to looking less stressed or fatigued.

### *Behaviourally Different*

The remaining statement pairs do not meet the criteria associated with interchangeability, nor do they have a close enough behavioural relationship to make inferences about the statements. Importantly, relationships between expressions evolved over time, with some phrases that were initially similar diverging in consumer response after extended use—highlighting the role of temporal context in claim equivalency. Overall, the results demonstrate that interchangeability is not binary, but exists along a spectrum and can be systematically identified.

Copyright © **Esomar** 2026


## Implications

This work demonstrates that benefit language equivalency can be validated using existing data, challenging longstanding assumptions about how claims must be tested and managed. It reframes language from a testing burden into a strategic, reusable asset. In practice, this approach has the potential to optimise claims tested, reduce redundant claim testing and/or accelerate decision-making timelines. The framework enables a more flexible and efficient approach to benefit language across the lifecycle:

- Early development: Identify language more likely to resonate and remain supportable;
- Launch preparation: Distinguish between benefit language requiring new research and those supported by existing evidence;
- Post-launch: Validate that revised language remains consistent with substantiated benefits.

This supports more adaptive communication while maintaining rigour. The goal is not universal standardisation, but clarity when consistency applies. Similar phrases may perform differently across categories, markets or product experiences. With connected data and semantic analytics, these patterns become measurable rather than assumed. This approach enables a shift from isolated studies to a connected data system. By linking language, consumer perception and context, organisations can reduce redundant testing, improve consistency through shared language frameworks and make better decisions about when differences are meaningful vs. cosmetic. For the insights community more broadly, this reinforces that language is a measurable, dynamic expression of value, not a fixed input.

## Limitations and Responsible Use

This approach is intended to enhance—not replace—primary research. Its value lies in informing when validation is needed.
Key limitations include:

- Not all expressions are equivalent—differences must be validated case-by-case;
- Findings depend on the quality and consistency of historical data;
- Semantic similarity does not guarantee behavioural equivalence;
- Context (category, market, usage) remains critical.

The framework should therefore be applied as a diagnostic tool, supporting evidence-based decisions, rather than replacing them.

## Next Steps

This research represents an important first step in our journey to understand semantic relationships between benefit statements within benefit areas. It opens the door to new research and further understanding in the topic area. Future development includes:

- Additional validation through new studies;
- Deeper exploration of partially aligned statements;
- Expansion into new categories and markets;
- Integration into tools and workflows.

## Conclusion

This study shows that benefit language—long treated as subjective—can be measured, validated and systematically managed using existing data. By combining statistical rigor with semantic analysis, organisations can identify where language is interchangeable and where distinction matters. The implications are significant: rather than re-testing similar claims, insights teams can extract new value from existing data—improving efficiency, consistency and decision-making. More broadly, this work reflects a shift in how insights systems operate. Innovation is not only about new data, but about rethinking how existing data is used. When connected and analysed effectively, historical datasets become active assets—supporting more agile, scalable and evidence-driven decision-making. Ultimately, this approach positions language not as a constraint, but as a scalable, evidence-based system for driving faster, more consistent and more consumer-relevant innovation.

Copyright © **Esomar** 2026



## References

Agresti, A. (2013, 3rd ed.). *Categorical data analysis*. Wiley.

Reimers, N., and Gurevych, I. (2019). *Sentence-BERT: Sentence embeddings using Siamese BERT networks*. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP).

Smith, E. E., Shoben, E. J., and Rips, L. J. (1974). Structure and process in semantic memory. *Psychological Review, 81*(3), 214–241.

## About the Authors

Douglas Ethé-Sayers, Managing Director, Market Probe International, US.

Kristina O'Keeffe, Vice President Global Clinical and Consumer Sciences, The Estée Lauder Companies, US.

István Hidegkuti PhD, Senior Associate Advanced Analytics, DataExpert, Hungary.

# Digital Twins: From Theory to Practice

## An Exploration of the Capabilities of Digital Twin and LLM-Generated Insights for Commercial Research

*Jon Puleston*

*David Priestley*

*Maciek Ozorowski*

*Davide Baldo*

Esomar

Office address:

Burgemeester Stramanweg 105

1101 AA, Amsterdam

Phone: +31 20 664 21 41

Email: info@Esomar.org

Website: www.Esomar.org

Publication Date: September 2026

ESOMAR Publication Series Volume S424 Congress 2026

ISBN 92-831-0337-8

Congress 2026
Copyright © **ESOMAR** 2026

Copyright © **ESOMAR** 2026
Congress 2026

# Digital Twins: From Theory to Practice

## An Exploration of the Capabilities of Digital Twin and LLM-Generated Insights for Commercial Research

*Jon Puleston*

*David Priestley*

*Maciek Ozorowski*

*Davide Baldo*

### Background

In late 2024, a research team centred at Stanford University released a study that caught the attention of the research industry. The team recruited just over 1,000 US adults, conducted two-hour deep-profiling interviews, and used the transcripts to construct AI-based generative agents, now known as "digital twins". These agents were designed to emulate how those same individuals might answer new questions.

The results were striking. Benchmarked against humans' own test-retest inconsistency, the twins reproduced responses to a broad set of social attitude questions at around 86% directional consistency (Park et al., 2024). For anyone working in research, the implication was hard to ignore. If AI could simulate real people with that level of accuracy, research could potentially become faster, cheaper and more scalable.

It sparked a wave of commercial interest, with technology providers and research companies exploring how digital twins, synthetic respondents and LLM-generated insight might be applied in practice.

But the study also raised important questions. Although the Stanford benchmarks included behavioural games and adjusted for differences in response consistency, much of the evaluation centred on broad social attitudes. Commercial research, often asks harder-to-predict questions rooted in specific brands, experiences, memories and immediate context. Would the same approach work as well in these settings? And beyond directional accuracy, how closely could digital twins reproduce the distribution of real human answers?

#### *From Academic Promise to Commercial Reality: The Ipsos Politics and Social Change Lab Collaboration*

To explore these questions, Ipsos teamed up with several researchers from Stanford's Politics and Social Change Lab, including Robb Willer, a member of the original research team, to examine whether the digital twin approach could be reproduced, adapted and stress-tested in a commercial research environment.

The collaboration gave Ipsos the opportunity to apply the original Park et al. methodology to a comparable sample of around 1,000 respondents from the Ipsos US Random Probability Panel, broadly matching the scale of the original study while extending the testing into commercial research applications.

Our aim was not simply to reproduce the original research with a commercial sample. It was to understand how far the methodology could travel beyond the academic setting.

Could digital twins reliably predict outcomes across the kinds of research tasks the insights industry relies on every day, including political opinion polling, advertising evaluation, brand tracking, ideation and concept testing?

To answer this, Ipsos assembled research experts from across specialist areas of the business to design test questions that would deliberately stretch the twins across different domains and levels of difficulty. The programme was coordinated through Ipsos' Synthetic Research Innovation team, combining expertise in AI, research methods and research technology.

The research programme was structured around three core areas.

First, replicability. Could the broad pattern of the Stanford-centred research team's result be reproduced using a shorter and more commercially practical profiling approach?

Second, optimisation. Could changes to profiling design, prompting strategy and the LLMs model orchestration improve predictive performance?

Third, commercial transferability. Could digital twins move beyond stable attitudinal prediction into more applied research tasks involving brands, advertising, product choice and consumer behaviour?

This final question proved especially important.

Commercial research often deals with questions that are inherently harder to predict than political or social attitudes. Brand choice can be habitual. Advertising response can be emotional and highly context-dependent. Product evaluation may involve price sensitivity, usage occasion, memory and comparison with familiar alternatives. These are often less stable, less consciously articulated and more situational than general beliefs or values.

This made the programme not just a replication exercise, but a broader stress test of how digital twins perform under real commercial research conditions.

## The Ipsos Politics and Social Change Lab Experiment: Adapting Digital Twins for Commercial Research

### *Part 1: Replicability*

*Question: Could we replicate the original Stanford-centred research team's results using a shorter evaluation instrument?*

The first challenge was to see if we could design a deep profiler that could provide enough information to prime the digital twins effectively, without placing unrealistic demands on respondents.

The original Stanford-centred research team's approach used around two hours of open-ended questions. In piloting, it quickly became clear that this would not be practical for a commercial research panel, especially once we added the more consumer-focused questions needed for market research.

To shorten the profiler, we undertook a detailed analysis of the pilot data. As part of this, in capturing the twins predictions, we asked the LLM to explain the rationale behind each predicted answer, then traced back which profiling questions appeared to be contributing the most useful collective insight. This allowed us to reduce the Stanford-style open-ended section to a core 20-minute set.

We then supplemented this with a range of structured questions covering financial circumstances, lifestyle, consumer attitudes, shopping behaviour and cognitive profiling. These were selected through separate experimentation for their ability to differentiate respondents and improve prediction across commercial research tasks.

*Answer: Just about*

So, could this shorter, more optimised Twins profiler perform as well as the original Stanford-centred team's deep profiling process? At first sight, the answer appeared to be no. Our initial results delivered around 73% raw directional accuracy against the benchmark questions used for comparison.

However, there was an important methodological detail behind the original Stanford-centred team's results.

As part of their evaluation, the Stanford team re-tested benchmark questions with real respondents and found that around one in five answers changed between test and retest. In other words, human responses themselves were not perfectly stable. The 86% headline figure was therefore not raw prediction accuracy against a fixed truth. It was a calibrated estimate that adjusted for the natural inconsistency in human answering behaviour.

This distinction matters.

If respondents themselves sometimes change their answers, then a twin can appear "wrong" when it is simply reproducing an alternative answer the same respondent might plausibly have given on another occasion.

Once we applied the same recalibration approach to the Ipsos data, our estimated accuracy rose to 84%, broadly in line with the original Stanford-centred team's result.

So, broadly, yes: by applying a comparable methodology and evaluation standard, we were able to achieve a similar level of predictive accuracy, using a much shorter and more commercially viable profiling process.

This was an important first finding. Achieving the original Stanford-centred research team's levels of accuracy did not require a two-hour interview. With careful optimisation, a shorter and more commercially practical profiler could deliver similar performance.

This also highlighted a wider challenge for digital twin research generally. Their accuracy is constrained not only by the model or volume of priming data, but by the reliability of the human data used to benchmark it. If real respondents are inconsistent, then a twin can appear to be wrong. Any serious evaluation of digital twins therefore needs to take account of human test-retest reliability, rather than treating the original survey answer as an unquestionable ground truth.

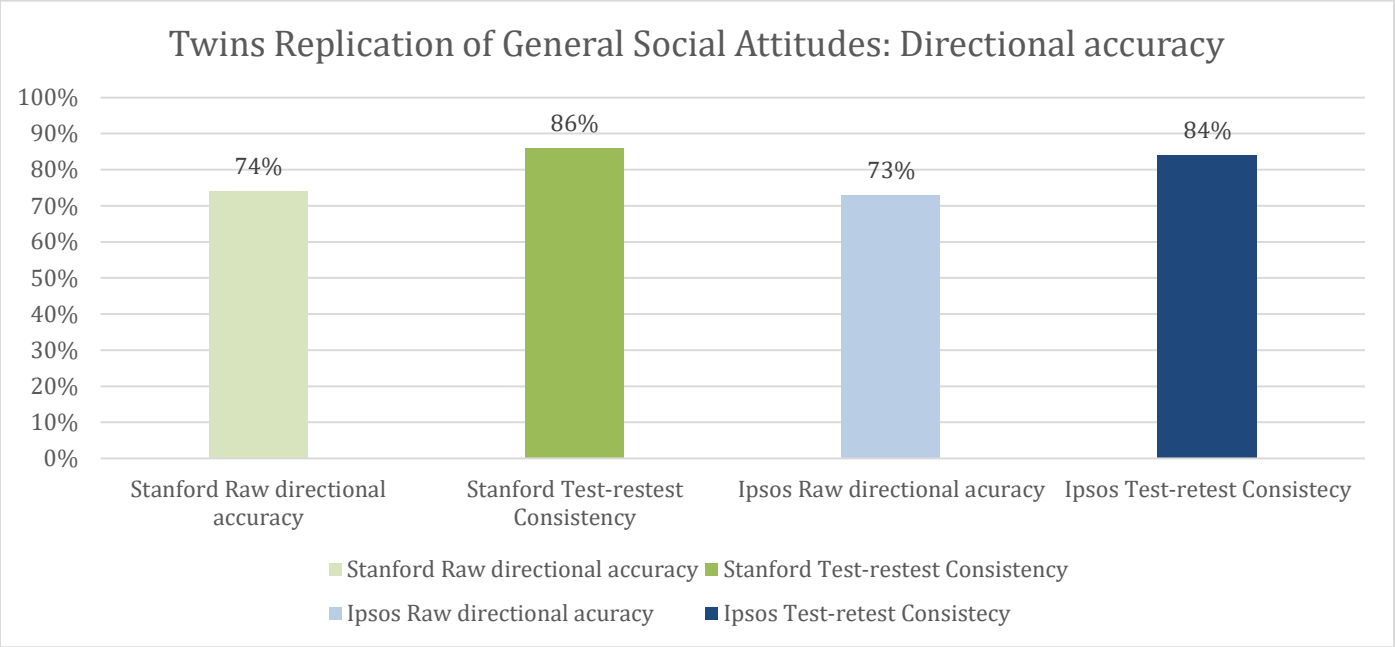



*Figure 1.*

*Question: So how much data is needed to prime a digital twin?*

It would be a natural assumption that the more information we give an LLM about a person, the better its predictions should become. And to a point, that appears to be true. A basic demographic profile is better than no profile, and a deeper profile can improve the model's ability to infer likely attitudes, preferences and behaviours.

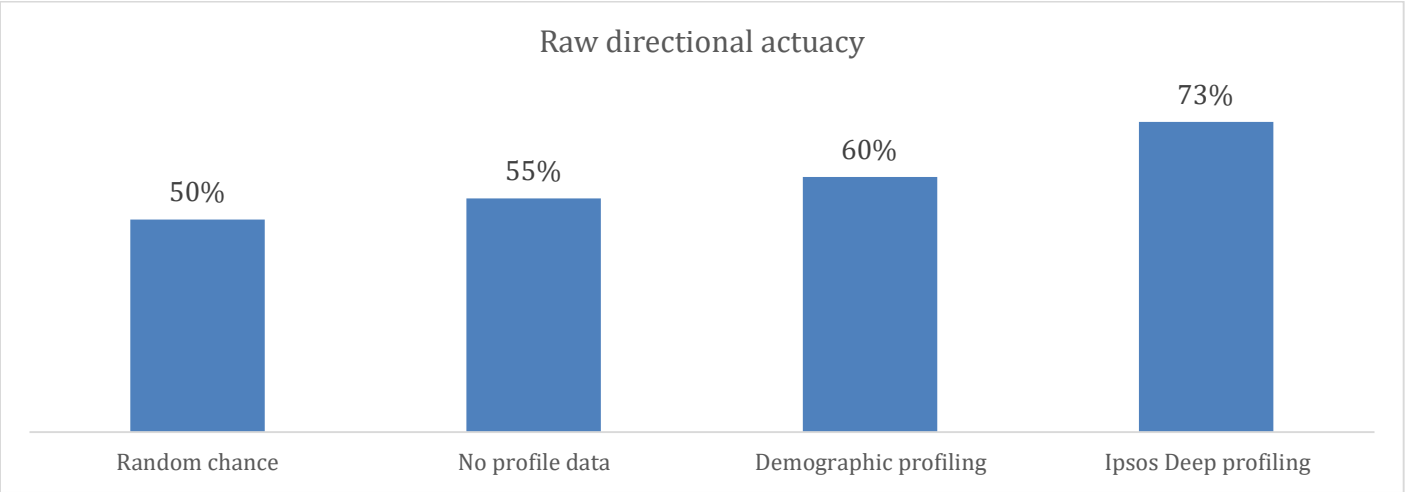



*Figure 2.*

## *Finding: Diminishing Returns on Additional Profiling Information*

To optimise the profiling process, the team tested how much prior information the twins needed to predict a set of benchmark questions reliably. Profiling questions were divided into smaller banks and added progressively, with prediction accuracy assessed at each stage.

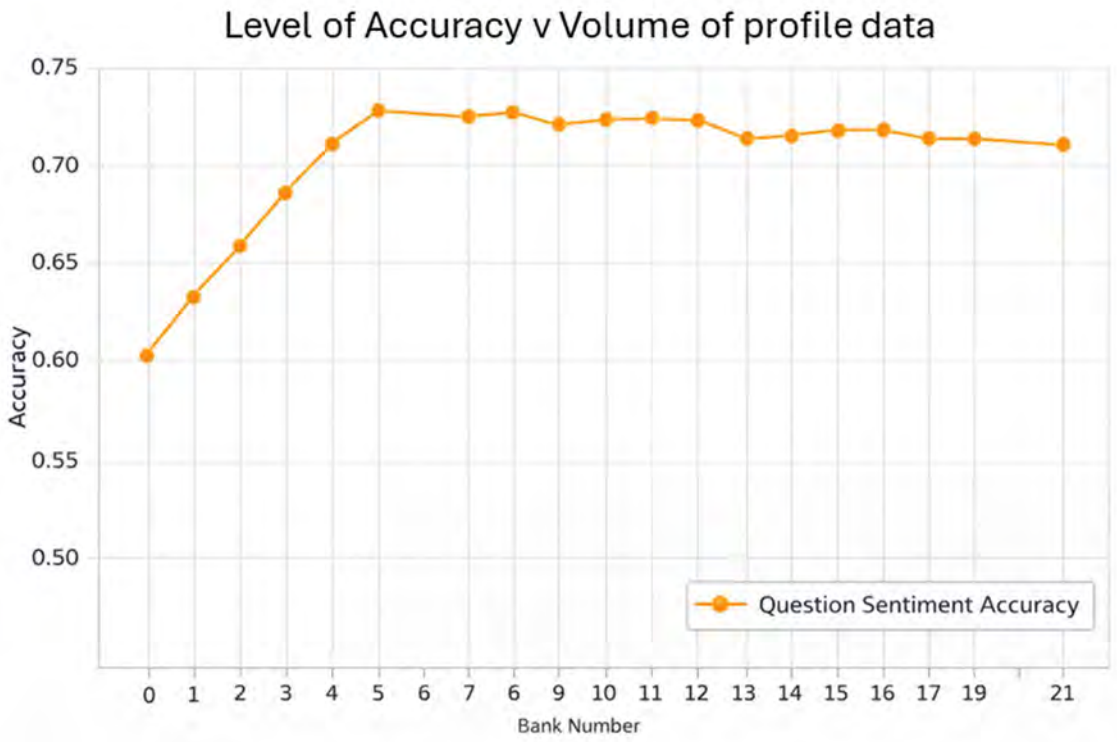



*Figure 3.*

As the chart above illustrate from one of these sequencies, prediction accuracy improved quickly at first, but then reached a plateau. Consistently it was observed that after a certain point, each additional bank of questions added less value and, in some cases, began to introduce noise which started to reduce rather than improve accuracy.

This should not be interpreted as a pure test of information volume. The later banks also contained information that earlier analysis suggested was less relevant to the predictions. The results therefore show that adding lower-value profiling information produced diminishing returns and could sometimes introduce noise.

One possible explanation lies in how LLMs process the information provided in a prompt. Position biases, including primacy and recency effects, can influence which information receives the greatest weight.* A language model does not "know" a respondent in the way another person might know them, or build a stable psychological portrait through lived interaction. Instead, it makes probabilistic judgements based on patterns in its training data and the information presented to it.

When given a profile, the model must infer which details are relevant, which are incidental and how different pieces of information should be weighted. This is easier when the signals are strong. A relatively short profile may contain highly predictive information, such as age, location, household structure, occupation, political leaning or category usage, providing clear anchors from which the model can infer likely opinions or behaviours.

As the profile expands, however, relevant signals may become harder to distinguish. Longer profiles can include details that are weakly predictive, ambiguous or context-dependent. Some details reinforce one another, while others conflict or distract from the information most relevant to the prediction task.

This is an important practical finding. Effective digital twin design may not depend on collecting the maximum amount of information possible, but on identifying the smallest set of information that supports reliable prediction for the task at hand. What information is useful will depend on the context in which the twin is being used.

The implication is that digital twin profiling should be approached with the same discipline as questionnaire design itself: not asking everything that could be asked, but focusing on what is most relevant and useful.

## *Part 2: Refining the LLM Prediction Protocol*

*Question: Could we refine the LLM prediction process?*

Our next step was to see if we could improve twins performance not simply by adding more priming information, but by changing how the LLM prediction process used that information.

The original Stanford-centred research team's study used GPT-4o to generate predictions and a fairly simple prompting process. By the time of the Ipsos experiments, model capability had advanced, but so had our understanding of how to prompt, structure and evaluate LLM-based predictions.

First, we tested whether using a more advanced model would improve performance while keeping the prediction process broadly comparable. It did, but only modestly: accuracy rose from around 85% to 88% on the original benchmark questions.

The larger gain came from changing the way the LLM's were used to predict the answers.

Ipsos' Synthetic Research Unit developed a multi-agent approach designed to make better use of the respondent priming data. Instead of asking one model to generate a single prediction, multiple LLM agents independently assessed the same respondent and question, each applying a slightly different lens and drawing on different parts of the profile and ordering the priming information in different ways.

A separate evaluation agent then reviewed the competing answers and rationales before selecting the most likely response.

*Answer: Yes*

This refined approach lifted absolute prediction performance from around 88% to 93% on the same benchmark questions.

That finding was instructive. Better models helped, but better orchestration helped more. The limiting factor was not simply model intelligence. It was how effectively the system could extract, weigh and reconcile the relevant evidence in the respondent profile. In other words, digital twin accuracy depends not just on which LLM is used, but on the quality of the priming data and the discipline of the prediction protocol built around it.

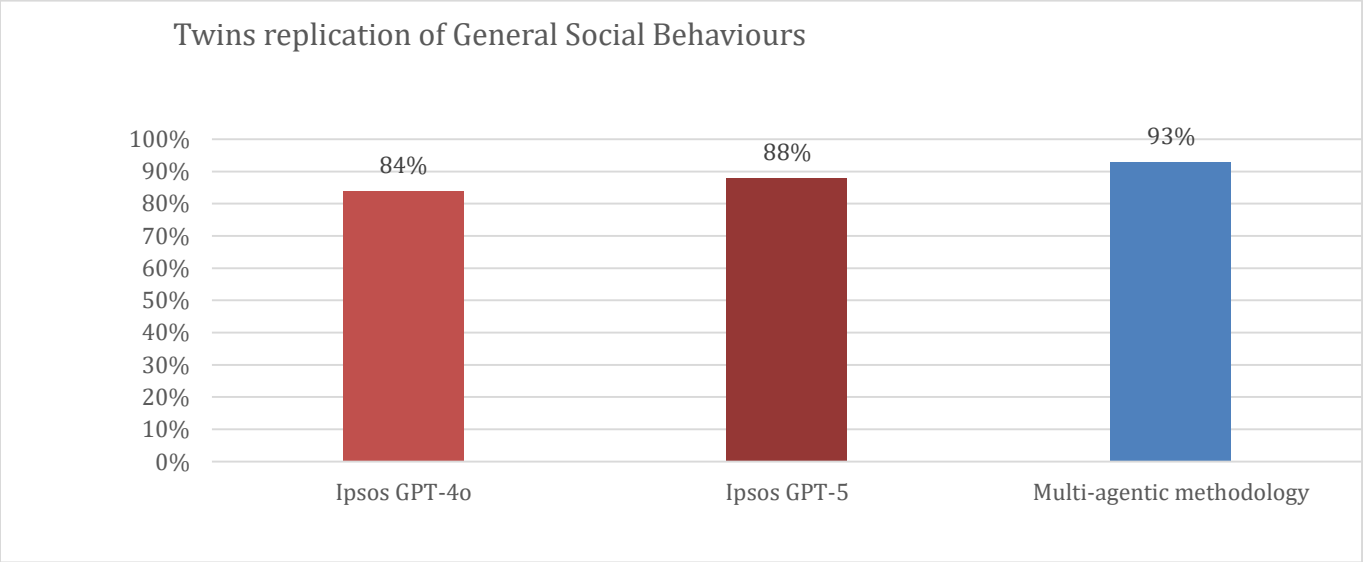



*Figure 4.*

## *Interpreting the Headline Accuracy*

At first sight, these prediction accuracy figures look strong. But they must be interpreted carefully.

First, these results were based only on the standard social behaviour and attitudes benchmark questions. It is also important to reinforce that this was a calibrated estimate based on consistency with real responses, not a raw directional accuracy figure.

Second, these figures only assessed the accuracy of each participant's answers, and in commercial research it is rarely enough simply to get individual answer directions right; the overall distribution balance is the key. A model might correctly predict which party wins an election, but there is a world of difference between predicting a 51% vote share and a 99% vote share. In market research, a few percentage points can materially change the interpretation.

The same applies to incidence and behaviour questions. It is easy to predict that most people do not own a fish. But a market researcher is more likely to need to know how many people actually do, within a small margin of error. A model that gets the majority answer right, while failing to estimate the minority accurately, may look strong on directional accuracy but still be unfit for the decision at hand. Respondent-level directional accuracy tells us little about whether the model can reproduce the full distribution of responses across scale or categorical questions.

## *Part 3: Road Testing Digital Twins on Commercial Research Questions*

*Question: How well could the digital twins answer real world commercial research questions?*

We asked teams from our public affairs, advertising, brand tracking, market segmentation and product innovation divisions to design test surveys representing typical commercial research tasks. These surveys were fielded among the real respondents who had completed our digital twin profiling survey. We then challenged the twins to predict how those same respondents would answer.

## *Finding: Performance on Commercial Questions is Highly Uneven*

When we look beyond this benchmark and examine these wider commercial research tests, performance becomes much more mixed. Across all these tests, directional accuracy dropped considerably, and the results varied sharply by question

and task. For some questions, the twins predicted responses with very high reliability, in the high 90s. For others, performance was much closer to a coin toss.

This variation is critical. The central lesson is not that digital twins are generally accurate or inaccurate, but that their reliability is highly task-dependent.

The strongest results tended to come from tasks that were structured, comparative or grounded in stable attitudes and explicit preferences. Performance was significantly lower for any prediction tasks that depended on lived experience, memory, context, emotion or low-salience behaviour.

### *Case Study 1: Using Digital Twins to Predict the Answers to a Conjoint Evaluation Study of Credit Card Offers*

A useful example of both the promise and the limits of digital twins came from a US credit card conjoint study that we asked the twins to emulate. The exercise used a dual-response conjoint design, asking human respondents and their corresponding digital twins to evaluate credit card offers built from a set of benefits and price levels.

At first sight, this appeared to be a promising application for digital twins. The task was structured, comparative and anchored in explicit product attributes: the kind of setting where an LLM-based model can often follow the logic of trade-offs. The twins generally understood that lower-priced offers were more attractive and were consistently able to make coherent choices across the exercise.

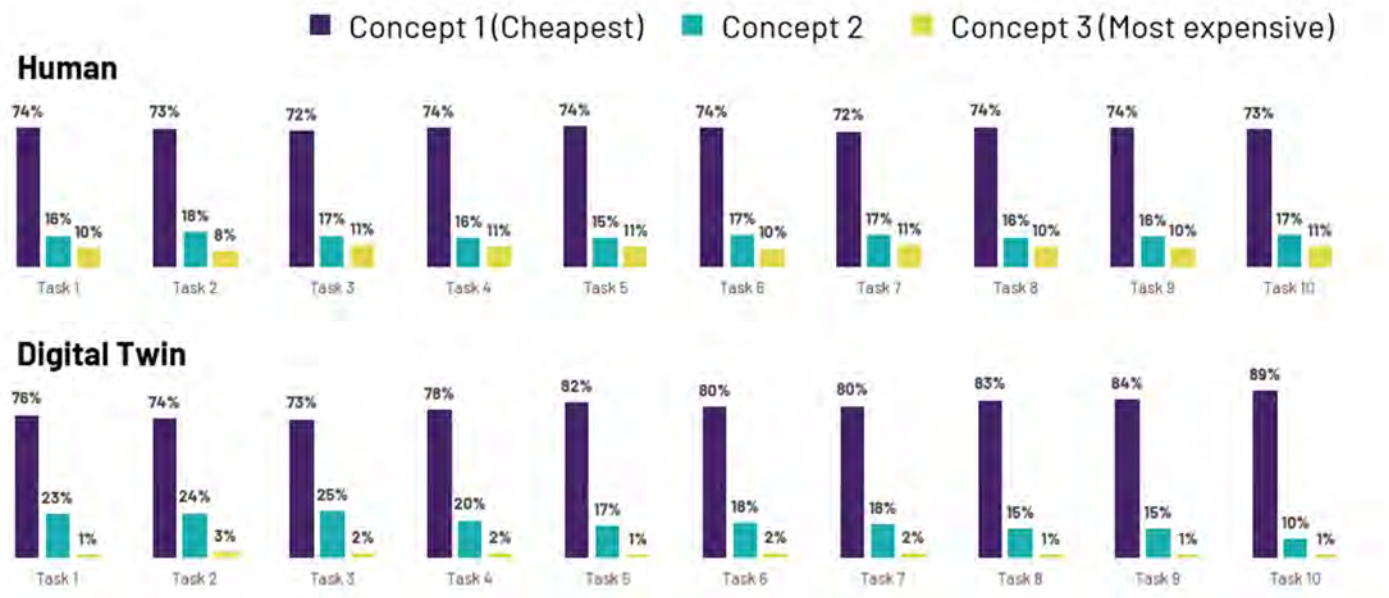



*Figure 5.*

However, the fuller results revealed a more complicated picture. As the exercise progressed, you may observe, the twins became increasingly likely to select the cheapest offer, while human respondents remained more stable. This suggests that the twins may have built an internal benchmark across the tasks and then applied it more systematically than real people. In one sense, this is rational. But that is also the limitation. Human choice behaviour is often less disciplined, more heuristic and more affected by habit, emotion and context.

The second part of the exercise made this limitation clearer. After evaluating the offer, respondents were also asked how likely they would be to take it up. Here, the gap between human and twin responses widened substantially. Human top-two-box purchase intent remained relatively stable across the tasks, while the twins started much higher and then declined as the exercise progressed. In other words, the twins were not simply choosing between offers differently; they were also over-estimating the likelihood that people would actually act on those choices.

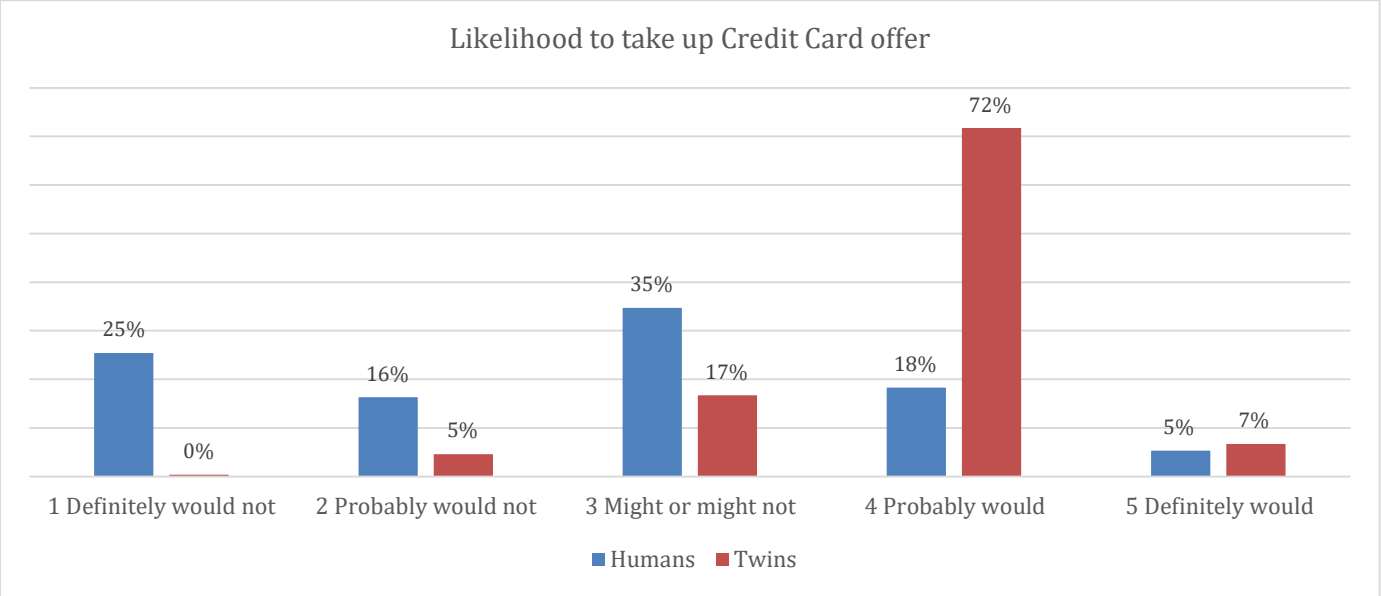



*Figure 6.*

What we have learned is that a model may be capable of making coherent trade-offs between price and benefits, while still failing to reproduce the behavioural caution, inertia, scepticism or lack of urgency that shapes real-world take-up. The twins appeared better at rational appraisal than behavioural realism.

This aligns with recent academic work suggesting that LLM-generated data may may have value in conjoint-style applications when used as an augmentation layer and calibrated against human data, rather than treated as a direct substitute for real respondents (Haghani et al., 2021; Train, 2009). But this case study also highlights an important boundary condition: digital twins may look strongest when the task is structured, comparative and attribute-led, yet still struggle when the research question moves from “which offer looks better?” to “would someone actually take it up?” The latter is a more experiential question, and depended on a wider range of factors that an LLM may not have any exposure too for example a persons existing brand preferences.

The lesson is therefore not that digital twins can or cannot “do conjoint”. It is more specific. They may have value in structured trade-off tasks, especially for early-stage exploration, scenario testing or augmenting human data. But their outputs need careful validation whenever the decision depends on predicting actual behaviour rather than stated preference alone.

## *Case Study 2: Using Digital Twins to Predict Public Options*

Ipsos’ Public Affairs division ran a validation study to test how well these digital twins could emulate the answers of real respondents in a public and social research context.

The test covered 47 predicted values across topic areas including the economy and personal finance, the environment, general sentiment about the country, societal and political concerns, personal beliefs and corporate social responsibility. Accuracy was assessed at total sample, segment and individual respondent level.

At first sight, the results looked encouraging but uneven. At the individual respondent level, the digital twins achieved around 70% directional accuracy overall. But this headline figure hid a much messier pattern. Performance varied by topic: corporate social responsibility and personal beliefs reached around 75%, while economy and personal finance was much weaker, at around 58%.

More importantly, the errors were not evenly distributed. The twins came close to the real sample on some public concerns, including crime and violence, inflation, taxes and threats against the environment. For example, concern about crime and violence was almost identical, at 28.4% among real respondents vs. 29.7% among twins.

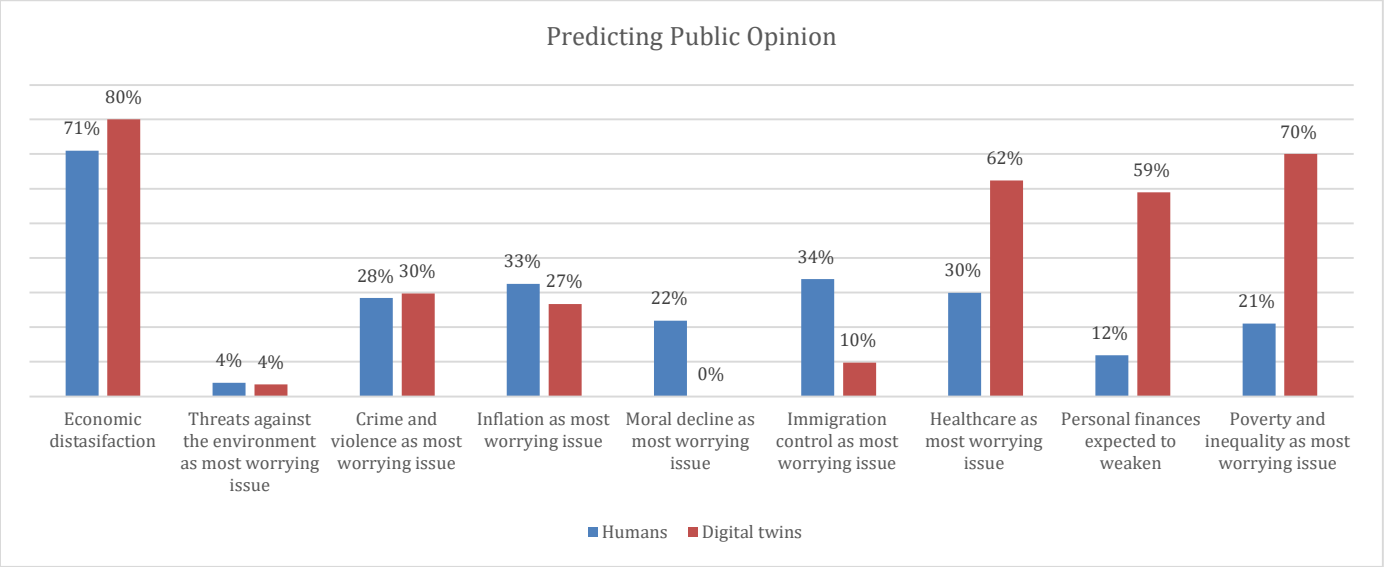



*Figure 7.*

The pattern broke down sharply on more ideologically loaded issues. The twins heavily over-predicted concerns about healthcare (29.9% vs. 62.4%), while under-predicting concern about immigration control (33.9% vs. 9.8%) and moral decline (21.9% vs. 0.0%). This did not look like just random noise. The pattern potentially suggested a systematic shift towards a more socially liberal reading of public concern.

The same pattern appeared in economic sentiment. The twins were more negative about the economy and personal finances than the real respondents. Expectations that personal finances would weaken rose from 11.9% to 58.9%. The most striking example was personal job insecurity: only 6.4% of real respondents said they were very likely to lose their job, compared with 55.5% of their twins.

This matters because systematic error is much more dangerous than random noise. Random errors may cancel out at aggregate level. Systematic errors can reshape the conclusions.

In this study, the model appeared to simulate Democrats more accurately than Republicans, and error was also linked to educational level and socio-economic status. That mean the issue could not simply be fixed by post-stratifying the synthetic data to match the known population profile. The bias existed within groups, not just between them.

The case study illustrates a central challenge for digital twins. A model can achieve a respectable overall accuracy score while still being unreliable in the places where accuracy matters most.

That does not mean digital twins are unusable in the public research domain. It means their use has to be carefully bounded and a lot more work is needed to better understand how to train LLMs to provide more balanced answers. They may have value for lower-risk exploratory work, hypothesis generation and applications where the prediction task is less politically or socially loaded. But this Public Affairs validation study shows why the burden of proof is much higher in public opinion, political attitude and policy-sensitive research, and it would be hard on this current evidence to support digital twins' use in this context.

This pattern is consistent with emerging external evidence. A recent mega-study tested digital twins across 19 pre-registered studies and 164 outcomes using a national US panel (Peng et al., 2026). It found that the average correlation between twin and human answers was modest, at around 0.2, and that twin responses were less variable than human responses. Performance also varied by domain and was stronger among more educated, higher-income and ideologically moderate participants. The broader lesson is clear: digital twins can capture some signal and some respondent heterogeneity, but their reliability is uneven and cannot be assumed from headline accuracy alone.

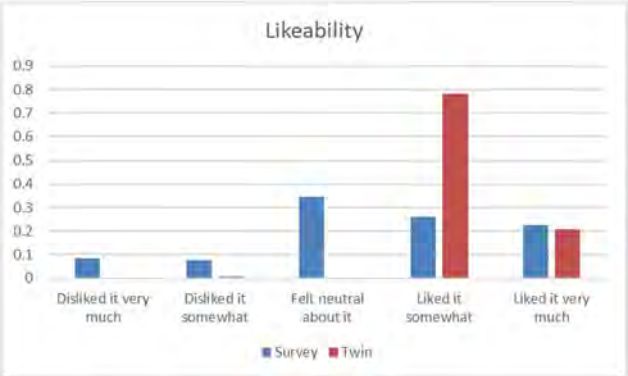
## *Finding: LLMs Do Not Naturally Answer Like Survey Respondents*

A second recurring pattern was that LLM-generated respondents did not naturally behave like human survey respondents. Their answers were often more coherent, more extreme, or more normatively positive than real respondent answers.

### *Case Study 3: Using Digital Twins for Advertising Copy Testing and Brand Evaluation*

Advertising copy testing and brand evaluation provided a clear example. This was a difficult test case because ad response depends on attention, taste, irritation, scepticism, boredom and weakly articulated emotional reaction.

When digital twins were asked to evaluate advertising, the difference between average human responses and twin responses was stark. The twins seemed far more inclined to find a logical reason to like the ads. Around 80% of twins said they liked the test ads somewhat, compared with 27% of real respondents. The same pattern appeared in brand ratings: twins were much more likely to predict that the ad would improve perceptions, while real respondents were far more sceptical.



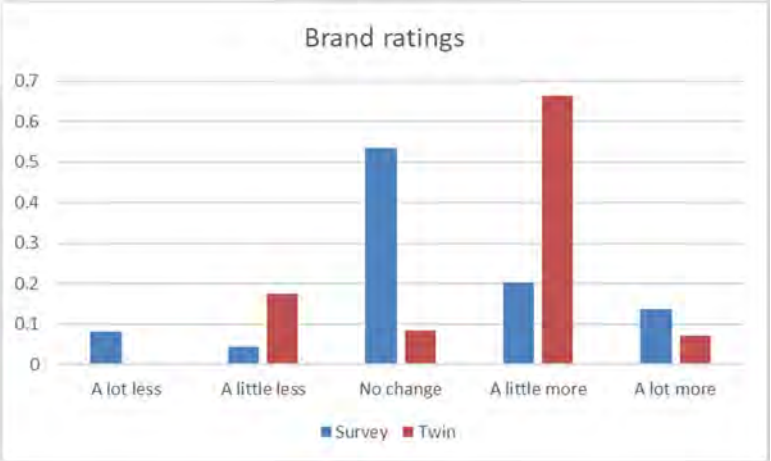




*Figure 8.*

Human respondents are inconsistent and emotional; we know that answers to scale questions vary dramatically by demographic and culture. A human may answer quickly, not fully read every option, they may overclaim, underclaim or select the first answer that feels good enough. In commercial research, this is not noise around the edges. It is part of the data-generating process.

LLMs, by contrast, tend to produce coherent answers. They infer patterns, resolve contradictions and generate responses that are often more internally consistent than real people. This may explain why twins performed better on structured tasks such as practical offer evaluation, but struggled more with tasks requiring emotional, behavioural or survey-response realism.

This creates a paradox. To emulate real respondents, digital twins need to learn not only what people think, but how imperfectly people answer surveys.

This is consistent with survey-methodology research showing that survey answers are shaped not only by underlying attitudes, but by comprehension, memory retrieval, judgement and response mapping. In other words, the act of answering a survey is itself a behavioural process (Tourangeau et al., 2000).

This remains an unresolved challenge at the time of writing this paper, but Ipsos and the wider academic community are actively exploring ways to address it. The Ipsos Synthetic Research Unit is currently researching how LLM-based systems can be better trained and calibrated to reflect the way humans answer survey questions, including the uncertainty, inconsistency and imperfection found in real response patterns. This includes work on adaptation layers designed to recalibrate synthetic answers against observed human data. Early work is showing measurable progress and we hope to share the findings in a future paper.

This is likely to be one of the next frontiers for synthetic data: not simply making AI responses more accurate, but making them more realistically human.

### *Finding: The Experience Gap*

Another clear learning from the experiments was that generalised priming data can only take digital twins so far. Their performance depends so heavily on whether the answer is inferable from the profile or experiential, meaning shaped by context the profile does not capture.

Inferable questions are those where the answer can reasonably be predicted from what is already known about a person: their values, beliefs, identity, habits, preferences, prior choices or previous statements. If a profile shows that someone is environmentally concerned, politically progressive or financially cautious, a model may be able to make a reasonable prediction about related attitudes or motivations. It can also make plausible extensions from known facts: for example, a dog owner may be more likely to walk regularly, buy pet-related products or respond to messages about care and companionship. LLM's are excellent at this type of projected conjecture.

Experiential questions are those where the answer depends on what the person has recently seen, felt, bought, remembered or encountered or how they are feeling. A brand choice, for example, may be shaped by level of hunger, the mood a person is in, price promotions, availability, advertising exposure or a memory triggered at the point of decision. Unless these signals are captured, the twin is being asked to infer something it has not been given.

This is the experience gap. Digital twins can often join the dots in a rich profile, but they cannot reliably account for dots that are missing. They are generally stronger at predicting stable attitudes, motivations and self-perceptions than behaviours shaped by recent exposure, lived context or low-consciousness decision-making. This matters because much market research is commissioned precisely because the answer is not obvious from what is already known. If the answer could be confidently deduced from existing information, there would often be less need to ask consumers in the first place.

### *Solution: Closing the Knowledge Gap with Topic-Specific Priming*

The wider experiments showed that general-purpose digital twins struggle when questions depend on experience, context or category knowledge that is not captured in the profile. This raised a practical question: could reliability improve if twins were built for a specific commercial task and primed with more relevant evidence?

### *Case Study 4: Using Digital Twins for Product Innovation Evaluation*

After reviewing the conjoint experiment results, Ipsos' Product Innovation team identified product evaluation as one area where digital twins might have genuine potential. However, the results also suggested that adding broad, generic profiling information, much of it irrelevant to the task, could introduce noise rather than improve prediction. The team therefore recognised the need to design a more focused profiling process and create a fresh set of twins with the specific knowledge needed to evaluate new product ideas.

Ipsos' Product Innovation team brought extensive experience in this area, including established validation frameworks and a strong understanding of the factors that shape consumer response. Rather than trying to build a universal product evaluation twin, the team developed a more focused model designed to predict specifically how US consumers would respond to new food product ideas.

The profiling was built around the drivers of food choice: ingredient preferences, eating habits, health attitudes, value sensitivity, convenience needs, environmental concerns, brand switching, variety seeking, label reading, household context and unmet needs. The aim was not to know more about people in general, but to know the things most likely to matter for US consumers when evaluating a food innovation. The model was primed with attitudes, preferences, behaviours and habits rather than past responses to old stimuli, so the twins could respond to new ideas rather than simply reproduce historic concept-test data.

The task was also framed to reflect real consumer decision-making. People do not evaluate new products in isolation. They compare them with what they already buy, eat or habitually choose. The twins were therefore asked to predict whether consumers would stick with the status quo or choose the new product.

The results were encouraging. In an initial validation across 12 food concepts, digital twin predictions were compared with human evaluations using Ipsos' Trial Potential Index:

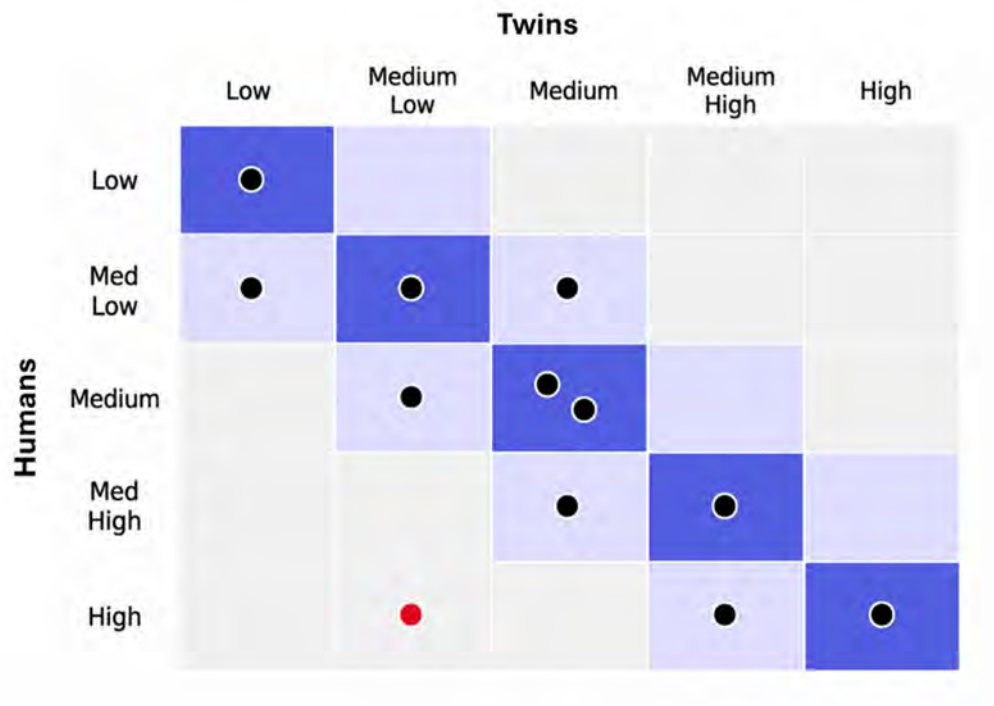



*Figure 9.*

The concepts were classified into five performance bands: Low, Medium Low, Medium, Medium High and High. For early-stage screening, a twin prediction was treated as accurate if it fell within one band of the human result. On this basis, 11 of the 12 concepts passed the accuracy standard: 6 were exact matches, 5 were within one band, and only 1 was more than one band out. Subsequent R&D and pilot studies found that digital twins could match human responses 85%+ using the same plus-or-minus-one-band criterion.

Later validation strengthened the evidence. The US food model was tested against 30 distinct food concepts, achieving 97% accuracy against Ipsos' proprietary trial potential metric. A separate US pet food model achieved 88% accuracy across 16 pet food concepts, suggesting the approach could transfer to another adjacent, but still specifically trained, category.

The lesson here is not that digital twins can predict any product response. This worked because the task was tightly defined, the priming data was relevant, and the benchmark was appropriate for early-stage screening. The model was useful for identifying promising ideas, not for final go/no-go decisions or testing tiny wording differences.

These experiments suggest that digital twins are better suited to testing the core proposition of meaningfully different concepts than distinguishing between minor wording, formulation or ingredient variations. Once again, the finer nuances of human reaction proved difficult for the twins to emulate.

Much of the current work of Ipsos' Synthetic Research Innovation team is therefore focusing on understanding how real respondents react to product stimuli and using that evidence to better train and calibrate twin responses. Although this work is not yet fully validated, early results suggest that measurable improvements in accuracy at product evaluation may be attainable through better model training and calibration using observed human response data.

The broader implication is clear: the experience gap can be narrowed, but not by adding more generic data. It is narrowed by adding the right data, targeted to the specific decision the twin is being asked to simulate and then training LLM's on how humans answer questions based on these experiences.

## Key Issue: Knowing What You Don't Know

The main observation from these experiments is that digital twins show potential, but their capabilities are uneven. The single most significant barrier to commercial use, therefore, is uncertainty: uncertainty about when digital twins can be trusted, where they are likely to fail, and how much confidence we should place in their outputs for any given decision.

Digital twins are often most valuable in situations where benchmark data does not already exist, where a quick answer is valued. But that creates a validation problem. If we already had the real answers, we would not need to simulate them. If we do not have the real answers, how do we know whether the simulation can be trusted?

In practice, twins do not need to be perfectly accurate to be useful. What matters is understanding how likely they are to be wrong and having some way to flag when confidence is high, low or somewhere in between.

This means digital twin research needs more than better prediction. It needs a reliability framework that can assess likely trustworthiness before outputs are used for decision-making.

### *Solution: Devising a Twin Confidence Framework*

To explore this, the Ipsos Synthetic Research Innovation team tested a wide range of approaches to uncertainty estimation. These ranged from simple checks, such as whether predictions varied meaningfully across demographic groups, to more technical methods that examined how familiar a question looked to the model, how sensitive its answers were to small internal changes, and how well confidence intervals could be calibrated against known historical errors.

No single method provided the answer, but several showed promise and together begin to form the basis of a confidence framework.

The first meaningful signal of uncertainty came from how much the model's answers varied across demographic groups. Where predictions shifted meaningfully between groups, this often suggested the model was drawing on learned underlying knowledge rather than producing generic or stereotyped responses. Topics such as politics, gun ownership and climate change showed strong demographic separation. More arbitrary or unfamiliar topics tended to generate flatter, less differentiated outputs. In simple terms, when the model recognises meaningful differences between people, confidence tends to increase.

The second signal came from distance from training data. Using "Mahalanobis distance", the team tested how closely a new question resembled the kinds of questions the model had seen before (McLachlan, 1999: 4, 20–26). Questions that sat close to the training data tended to be more reliable. Questions that fell further away could be flagged as unfamiliar or out-of-distribution, meaning the model may be extrapolating beyond what it truly understands.

The third promising signal came from prediction stability, using Monte Carlo Dropout Adaptive Conformal Intervals (Gibbs & Candès, 2021). Here the model was run multiple times with small parts of the network temporarily switched off. If the answer remained stable across repeated runs, confidence was higher. If the prediction wobbled significantly,

uncertainty increased. Importantly, that uncertainty needed to adapt to the difficulty of the question itself, since some questions are naturally harder to predict than others.

Taken together, these methods point towards a practical Twin Confidence Framework built around three core questions:

*Does the model appear to understand this topic?*
Observed through how meaningfully predictions vary across different demographic groups.

*Does this question look familiar?*
Measured through similarity to the data the model has previously been trained on.

*How stable is the prediction when tested repeatedly?*
Measured through how much the answer changes under repeated perturbation.

Alongside these technical signals, there is also an important research judgement layer.

A digital twin prediction is more likely to be trustworthy when the question is closely related to information already contained in the profile, when the underlying construct is stable rather than situational, and when the answer can reasonably be inferred from known attitudes, values or behaviours.

Confidence is also stronger when repeated runs or different models produce similar answers, when the underlying human measure itself has reasonable test-retest reliability and when the profile contains relevant first-person evidence rather than relying only on broad demographic descriptors. Predictions are also generally safer when interpreted at aggregate level, rather than over-read as precise individual truth.

Conversely, confidence should be lower when the answer depends heavily on recent exposure, mood, price, promotion, availability, memory, impulse, social context or low-salience behaviour. These are areas where even real people are often inconsistent, context-dependent or difficult to predict from profile data alone.

This is an important area where Ipsos is still in the discovery phase, and a topic we hope to explore in more detail in a future paper.

## *Finding: Cost and Resources*

A final practical consideration is cost. Digital twins are often assumed to be cheaper than asking real people. In the current state of the technology, this assumption is not always correct.

Reliable twin generation required substantial priming, careful orchestration, multiple model calls and significant token usage. In some cases, the computational cost of generating a synthetic response could exceed the cost of asking a real respondent the equivalent question.

This may change rapidly as models become more localised and so cheaper to run. But the current economics matter because they challenge a simplistic view of digital twins as an immediate replacement for a survey sample.

At present, their strongest use case may not be replacing respondents at scale. It may be in areas where they add value that conventional research cannot easily provide: scenario exploration, hypothesis generation, pre-testing question logic, simulating possible response patterns, identifying where more human evidence is needed, or extending insight from deeply profiled samples with appropriate safeguards.

## Implications for the Insights Industry

The findings point to a measured conclusion. Digital twins are not science fiction, and they are not yet a general replacement for primary research.

But their reliability is uneven. It depends on the task, the quality of the profiling data, the prediction method and the nature of the behaviour being predicted. Their ultimate constraint is the human experience gap.

For the insights industry, the opportunity is not to replace respondents with simulations. It is to use simulation to extend what we can learn from real respondents: more efficiently, more creatively and with a clearer understanding of where synthetic insight is strong, where it is weak, and where human evidence remains essential.

The jury, for now, is still out. We are still on the journey to understand their true potential. But the evidence suggests digital twins may become a valuable new research tool in certain areas of research, provided they are used with careful validation, clear confidence frameworks and appropriate methodological caution.

## Ethics and Responsible Use

Digital twins are not simply synthetic rows in a dataset. They are simulations derived from real people's disclosed information. That makes consent, transparency and purpose limitation essential.

Respondents need to properly understand how their data may be used, and what their "twins" may be asked to do. Clients need clear guidance on the confidence, limitations and appropriate use of any synthetic output. As researchers we need to bee measured about what simulations represent.

Responsible use requires more than technical validation. It also requires some governance around how profiles are created, how outputs are labelled, how uncertainty is communicated, and where synthetic evidence should not be used as a substitute for listening to real people.

Some accountability also needs to be built into the design of digital twin research. If synthetic respondents are used to generate insights, researchers should be able to trace how those responses were produced and assess whether they are defensible. As a matter of good practice, we suggest that digital twins should not only answer questions, but also provide a recorded explanation of the basis for each answer.

Explanations that make clear whether the response was grounded in stated respondent data, inferred from broader preferences or behavioural patterns, or shaped by assumptions introduced by the model. Without this kind of audit trail, synthetic outputs risk becoming fluent but unaccountable assertions.

Digital twin profiles also create specific data-security risks. They may contain unusually rich personal information, including attitudes, habits, preferences, family circumstances, financial behaviours, political views, health concerns, vulnerabilities and detailed first-person narratives. Even when direct identifiers are removed, this kind of data can remain highly sensitive.

If accessed or misused, these profiles could potentially support manipulation, targeting, impersonation or other forms of harm. They therefore require strong safeguards: strict access control, clear retention limits, secure storage, robust anonymisation or pseudonymisation, audit trails and clear rules on when profiles should be deleted.

Copyright © **ESOMAR** 2026
Congress 2026

## Acknowledgements

The authors would like to thank the many Ipsos colleagues and external collaborators who contributed to the research, development, testing and interpretation of the work described in this paper. Their input helped shape the methods, challenge the findings and refine the practical conclusions. Many of these contributions were central to the development of the case studies and validation work.

We would firstly like to thank the research team at Stanford University's Politics and Social Change Lab, led by Robb Willer, Professor of Sociology, Psychology and Business, whose work and collaboration helped inform this whole programme of research.

From Ipsos, we are especially grateful to Cameron Stronge and David Yokoinele, who conducted much of the background analysis; Chris Moore and Duncan Ferguson, who oversaw the conjoint research experiments; Jiongming Mu, Yuding Duan and Colin Ho from the Ipsos Innovation team, who devised the product innovation experiments; Amandine Lama, Oliviero Marchese and Emmanuelle Jego from Ipsos Public Affairs, who oversaw the public affairs research experiments; Sanaa Derqaoui, Roxana Adam, Gaelle Vadeleux and Jill Urlage from our Creative Excellence research team and Sam Davis, Yaroslav Kozyrev, Marialena Athanasopoulou and Julia Schulz from our Brand Tracking team who oversaw the advertising and brand evaluation experiments.

### *AI Assistance*

This paper was developed with the assistance of a large language model, which was used to help organise material, refine wording and support drafting. The research evidence, analysis, methodological interpretation and final conclusions remain the responsibility of the authors. All AI-assisted text was reviewed and edited by the authors.

## Appendix: Selected Sources and Relevance

| Theme | Source | Why it matters for this paper |
|---|---|---|
| **Core Stanford benchmark** | Park et al. (2024), *Generative Agent Simulations of 1,000 People* | Main source for the Stanford digital twin study: 1,000+ adults, two-hour interviews, and the test-retest calibrated benchmark. |
| **Synthetic samples** | Argyle et al. (2022), *Out of One, Many* | Early paper showing how LLMs can simulate survey-like responses from demographically conditioned prompts. |
| **LLM positional/order effects for long prompts** | Liu, N. F. et al. (2024). Lost in the middle: How language models use long contexts. | The methodology of how you prime a prime and LLM to create twins is important |
| **External validation / caution** | Peng et al. (2025/2026), *Digital Twins as Funhouse Mirrors* / mega-study. | Supports the paper's core finding that digital twin performance is uneven, domain-dependent, and subject to systematic distortions. |
| **Market research augmentation** | Wang, Zhang & Zhang (2024), *Large Language Models for Market Research.* | Supports the argument that LLM-generated data may be useful as an augmentation layer, especially for conjoint-style tasks, but not as a direct substitute for human data. |

| Theme | Source | Why it matters for this paper |
|---|---|---|
| **Hypothetical bias** | Haghani et al. (2021), *Hypothetical Bias in Stated Choice Experiments.* | Useful for the finding that twins performed better on offer evaluation than on likely take-up (and that human stated intent can diverge from real behaviour so hard to compare). |
| **Calibration** | Guo et al. (2017), *On Calibration of Modern Neural Networks.* | Supporting evidence that confidence scores need calibration, not just accuracy. |
| **Survey response realism** | Salecha et al. (2024), *Large Language Models Show Human-like Social Desirability Biases in Survey Responses.* | Supporting evidence that LLMs may reproduce or distort human-like survey biases, rather than behave like neutral respondents. |

*Table 1.*

## References

Gibbs, I. and Candès, E. (2021). Adaptive Conformal Inference Under Distribution Shift. *Advances in Neural Information Processing Systems*.

Haghani, M., Bliemer, M. C. J., Rose, J. M., Oppewal, H., & Lancsar, E. (2021). Hypothetical Bias in Stated Choice Experiments: Part I. Integrative Synthesis of Empirical Evidence and Conceptualisation of External Validity.

Liu, N. F., Lin, K., Hewitt, J., Paranjape, A., Bevilacqua, M., Petroni, F. & Liang, P. (2024). "Lost in the Middle: How Language Models Use Long Contexts." Transactions of the Association for Computational Linguistics.

McLachlan, G. J. (1999). "Mahalanobis Distance." *Resonance*, 4, 20–26.

Park, J. S., Zou, C. Q., Shaw, A., Hill, B. M., Cai, C. J., Morris, M. R., Willer, R., Liang, P., & Bernstein, M. S. (2024). *Generative Agent Simulations of 1,000 People*.

Peng, T., Gui, G., Merlau, D. J., Fan, G. J., Ben Sliman, M., Brucks, M., Johnson, E. J., Morwitz, V., … Netzer, O., & Toubia, O. (2025). A Mega-Study of Digital Twins Reveals Strengths, Weaknesses and Opportunities for Further Improvement.

Tourangeau, Rips & Rasinski (2000), *The Psychology of Survey Response*.

Train, K. E. (2009). Discrete Choice Methods with Simulation (2nd ed.). Cambridge University Press.

## About the Authors

Jon Puleston, Chief Methodologist, Ipsos, UK.
David Priestley, Head of Data Science, Synthetic Data, Ipsos.
Maciek Ozorowski, Global Head of Synthetic Data, Ipsos, UK.
Davide Baldo, Global Lead, Experimental Research and Ipsos Academic Partnership Programme, Ipsos, France.

# Research During Wartime

## How Insight Infrastructure Survives a Mega-Crisis

*Evgeniya Blyznyuk*

## Esomar

Office address:

Burgemeester Stramanweg 105

1101 AA, Amsterdam

Phone: +31 20 664 21 41

Email: info@Esomar.org

Website: www.Esomar.org

Publication Date: September 2026

ESOMAR Publication Series Volume S424 Congress 2026

ISBN 92-831-0337-8

Congress 2026
Copyright © **ESOMAR** 2026

Copyright © **ESOMAR** 2026


# Research During Wartime

## How Insight Infrastructure Survives a Mega-Crisis

*Evgeniya Blyznyuk*

### Introduction: Research in a Society Under Extreme Stress

Conventional research methodologies are designed around assumptions of relative stability. Population structures are expected to change gradually. Infrastructure is assumed to function reliably. Respondents are presumed to live in environments where participation in research does not create personal risk.

War disrupts all these assumptions simultaneously.

On 24 February 2022, Russia launched a full-scale invasion of Ukraine. Within days, millions of Ukrainians began moving across the country and across international borders. Entire regions became inaccessible. Air raid alerts, missile attacks, power outages and uncertainty became part of everyday life.

Under such conditions, research might appear impossible. Historically, crises often lead to reductions in research spending as organisations prioritise immediate operational needs. However, the Ukrainian experience demonstrated the opposite dynamic. The more uncertain the environment became, the greater the demand for evidence-based decision-making.

Gradus launched its first wartime survey on 28 February 2022, four days after the invasion began. What started as an attempt to understand rapidly unfolding social change evolved into a four-year effort to continuously document how Ukrainian society adapted to prolonged crisis.

This paper explores the methodological and ethical challenges encountered during that period and reflects on what they reveal about the future of research in turbulent environments.

### Wartime Research Context: A Moving Population

Perhaps the most significant methodological challenge of wartime research is the instability of the population itself.

Before the invasion, Ukraine's population was estimated at approximately 37 million people. Since 2022, the country has experienced large-scale external migration, internal displacement, mortality increases and declining birth rates. Population estimates now vary considerably, ranging from approximately 22 to 32 million depending on methodology and source.

According to United Nations estimates, more than six million Ukrainians currently reside abroad under temporary protection programmes. Millions more have experienced internal displacement. Many residents of occupied territories remain inaccessible to researchers, and substantial uncertainty surrounds populations that left through eastern borders where systematic registration was limited.

The challenge is therefore not simply one of population decline. Researchers face a population whose size, structure and geographic distribution are continuously changing.

Traditional sampling frameworks become increasingly difficult to maintain under these conditions. Census data rapidly lose relevance, while official statistics often lag behind actual population movements. Researchers must therefore rely on a combination of pre-war demographic benchmarks, migration statistics, administrative records and ongoing methodological adjustments.

Representativeness becomes less a fixed outcome and more a continuously negotiated methodological objective.

## Methodology

### *Research Infrastructure*

The findings discussed in this paper are based on four years of continuous research conducted by Gradus between 2022 and 2026.

The primary research infrastructure consists of a proprietary smartphone-based respondent panel. The panel was established before the full-scale invasion and provided a critical advantage during wartime because researchers could maintain contact with respondents regardless of their geographic relocation.

Data collection combines:

- Continuous quantitative surveys using CAWI methodology
- Longitudinal panel studies
- Rapid-response polling
- Online focus groups
- In-depth qualitative interviews

### *Sampling Approach*

Quantitative studies primarily rely on quota sampling by:

- Gender
- Age
- Region
- Settlement size

Sampling quotas are based on pre-war population structures adjusted using:

- Official migration statistics
- Internal displacement estimates
- Population redistribution data
- Expert demographic assessments

Respondents located in temporarily occupied territories and active combat zones are excluded from most quantitative studies due to ethical considerations and safety concerns.

### *Longitudinal Measurement*

Copyright © **ESOMAR** 2026


A particularly valuable component of the research design involves repeated measurement of key social indicators over time.

Longitudinal tracking allows researchers to distinguish between short-term reactions to events and bigger structural changes in attitudes, behaviours and identity.

Several indicators have been monitored continuously throughout the war, including:

- Subjective stress levels
- Language use
- Consumer preferences
- Trust and institutional perceptions
- Migration intentions
- Perceptions of personal wellbeing

## Findings: How Respondent Behaviour Changes During War

One of the most striking findings concerns participation behaviour itself.

In the immediate aftermath of the invasion, response rates increased dramatically.

Contrary to expectations, respondents actively sought opportunities to share their experiences. Surveys addressing evacuation, shelling, displacement, and wartime adaptation generated exceptionally high engagement. National samples could often be completed within hours, even when respondent incentives were removed entirely.

The willingness to participate appeared driven by several factors:

- The need to document personal experiences
- A desire to contribute to collective understanding
- High emotional salience of the subject matter

However, participation patterns evolved over time.

By the second year of the war, response rates began to decline. Public fatigue increased. Expectations of a rapid resolution diminished. Fear associated with military mobilisation also affected participation, particularly in interviewer-administered studies.

From the third year onward, response rates stabilised at levels comparable to pre-war benchmarks. Yet participation became highly sensitive to the broader security environment.

Following nights of intense missile attacks, survey completion rates frequently slowed. Power outages disrupted internet access and reduced respondent availability. In such contexts, operational performance became directly linked to military developments.

The findings demonstrate that response behaviour during crisis cannot be understood independently of the social and psychological environment in which respondents live.

Copyright © **ESOMAR** 2026


## Longitudinal Insights: Tracking Social Transformation

Continuous measurement revealed profound social transformations across multiple dimensions.

### *Stress and Emotional Wellbeing*

One of the most consistent findings concerns psychological stress.

Since the COVID-19 pandemic, Gradus has tracked subjective stress using a composite indicator based on multiple symptoms of emotional strain.

Ukraine historically exhibits elevated stress levels due to recurring political, economic and social crises. Nevertheless, by the fourth year of the full-scale war, more than 90% of respondents reported significant manifestations of stress.

At the same time, longitudinal evidence suggests that stress alone does not capture the complexity of societal adaptation.

Periods of heightened fatigue frequently coexist with strong indicators of resilience, solidarity and collective purpose.

### *Identity and Language*

Language use provides another example of long-term transformation.

Ukraine has historically been characterised by widespread bilingualism. However, the invasion accelerated a significant shift toward exclusive use of Ukrainian in daily communication.

This trend is particularly pronounced among younger generations.

For many respondents, language became both a practical and symbolic marker of distinction from the aggressor state. The shift illustrates how identity-related changes may emerge rapidly under conditions of external threat.

### *Consumer Behaviour*

Consumer behaviour also changed substantially.

Immediately following the invasion, brand importance declined sharply. Product availability became more important than brand preference as supply chains experienced disruption.

Over time, however, brand importance gradually recovered.

Research also documented a notable increase in support for domestic producers. Consumers increasingly prioritised Ukrainian-made products across numerous categories.

Initially driven by patriotic sentiment, this preference was reinforced by competitive pricing and improved product quality. By 2025, support for domestic production had become one of the most significant long-term consumer trends observed during the war.

## Qualitative Research Under Conditions of Trauma

While quantitative research faces methodological challenges, qualitative research encounters particularly complex ethical issues.

The central challenge is sensitivity.

Experiences of war are highly unevenly distributed across the population. Topics that appear neutral from a research perspective may trigger traumatic memories for participants.

For example, discussions involving family life, parenting or household consumption require moderators to understand whether participants have experienced military service, displacement, bereavement, occupation or other wartime experiences.

Moderator preparation therefore extends beyond traditional facilitation skills.

Researchers conducting in-depth interviews with internally displaced persons, survivors of occupation, veterans or relatives of fallen soldiers require specialised training and emotional support.

The emotional burden affects researchers as well as respondents.

An important methodological adaptation has been the widespread transition to online qualitative formats. War accelerated digital adoption across Ukrainian society, making online interviews and focus groups both practical and acceptable for most target audiences.

## Ethics and the Principle of "Do No Harm"

Wartime research places ethics at the centre of methodological decision-making.

Classical survey methodology prioritises representativeness. Wartime ethics may require deliberate departures from that objective.

At Gradus, respondents located in active combat zones and occupied territories are excluded from most quantitative studies.

This decision reduces theoretical population coverage. However, it minimises potential risks associated with participation and aligns with the broader ethical principle of "do no harm".

The war also challenges the assumption of researcher neutrality.

Researchers in Ukraine are not detached observers. They live under the same conditions they seek to analyse. They experience air raid alerts, displacement, loss, uncertainty and emotional stress.

This proximity creates both advantages and risks.

Researchers often possess deeper contextual understanding. Simultaneously, personal experiences may introduce unconscious bias into questionnaire design, interpretation and analysis.

To address these risks, internal peer review, questionnaire testing and methodological reflection become especially important.

Copyright © **ESOMAR** 2026


## Data as a Strategic Asset

Another distinctive feature of wartime research concerns the strategic significance of data itself.

Research findings can influence international narratives, public opinion, political decision-making and media coverage.

As a result, data communication requires exceptional caution.

Early in the war, findings from business sentiment studies were selectively quoted by Russian propaganda sources and presented internationally as evidence of anticipated economic collapse in Ukraine.

Although the original research did not support such conclusions, the episode demonstrated how easily data can be removed from context.

Researchers operating in conflict environments must therefore consider not only methodological validity but also the potential consequences of publication.

Responsible communication becomes part of the research process itself.

## The Importance of Speed

War dramatically increases the value of rapid insight generation.

Political, military, and social developments can reshape public attitudes within hours.

One illustrative example occurred in February 2025, when public statements by US President Donald Trump questioned the level of support for Ukrainian President Volodymyr Zelenskyy.

Within four hours, an express survey conducted by Gradus produced empirical evidence contradicting the claim. The results provided journalists, policymakers and international observers with a factual reference point during a highly sensitive public debate.

This example highlights a broader lesson: in volatile environments, the ability to generate reliable data quickly becomes as important as methodological sophistication.

## Implications for the Global Insights Community

The Ukrainian experience offers lessons extending far beyond wartime research.

The world is entering an era characterised by increasing instability. Armed conflicts, climate-related disruptions, migration crises, pandemics and information warfare are becoming more frequent.

Under such conditions, researchers can no longer assume stable operating environments.

Several implications emerge:

1. Research infrastructure must be resilient. Proprietary panels, digital data collection systems, and diversified methodologies increase operational continuity during disruption.

2. Longitudinal measurement becomes more valuable. In turbulent contexts, understanding change requires comparison over time rather than isolated snapshots.
3. Ethics must be integrated into methodology. Safety and wellbeing may sometimes outweigh traditional representativeness goals.
4. Researchers need reflexivity. Studying crises from within requires deliberate mechanisms for recognising and mitigating bias.
5. Data communication deserves strategic attention. Findings may carry political consequences and require careful contextualisation.

## Conclusion

The Ukrainian experience demonstrates that research does not become less important during crisis. In many respects, it becomes more important than ever.

Over four years of war, continuous research has documented how a society adapts to prolonged uncertainty, displacement, trauma, and transformation. It has provided businesses, policymakers, media organisations, and international partners with evidence-based understanding at moments when assumptions quickly became unreliable.

The experience also suggests a broader conclusion.

Research infrastructure should be viewed not only as a business capability but as part of societal resilience. When environments become unstable, reliable insight helps societies understand themselves and make better decisions.

As global volatility increases, the ability to conduct credible, ethical and adaptive research under conditions of uncertainty may become one of the defining competencies of the insights profession.

## About the Author

Evgeniya Blyznyuk, Founder and CEO, Gradus, Ukraine.

# Ten Times the Research.
# The Same Team. A Different Practice.

*Chris Courtney-Smith*

*Leanne Tomasevic*

## Esomar

Office address:

Burgemeester Stramanweg 105

1101 AA, Amsterdam

Phone: +31 20 664 21 41

Email: info@Esomar.org

Website: www.Esomar.org

Publication Date: September 2026

ESOMAR Publication Series Volume S424 Congress 2026

ISBN 92-831-0337-8

Congress 2026
Copyright © **ESOMAR** 2026

Copyright © **ESOMAR** 2026


# Ten Times the Research. The Same Team. A Different Practice.

*Chris Courtney-Smith*

*Leanne Tomasevic*

## Introduction

Organisations are under pressure like never before. Amidst rising costs, consumers are cutting spending (Strauss & Romei, 2025) while organisations themselves face insufficient budgets to win custom and attention. Research shows that more than one in two (56%) CMOs in North America, the United Kingdom and Europe say their marketing organisation lacks the budget required to deliver their 2026 strategy and more than half (54%) report insufficient resources (Gartner, 2026).

Added to this is the pressure to incorporate new artificial intelligence technology, in the right way, to maintain competitiveness as it transforms industries. With almost nine in ten (88%) organisations reporting regular AI use in at least one business function, and high AI performers nearly three times as likely as others to redesign their workflows (McKinsey, 2025: 4–16), the incorporation of AI has become non-negotiable.

In this landscape, research and insight teams find their role in flux. Budget constraints threaten to diminish their role as, rather than spend on expensive and time-consuming customary surveys and fieldwork, organisations rely on "gut" feeling to make decisions quickly and cheaply. Furthermore, the advent of AI calls for a reimagining of roles, workflows and processes, in order for organisations to make the most of its possibilities. If research and insight teams are to increase, and not diminish, in importance, a new type of practice is needed.

An example of this new type of practice is found in the use of synthetic audiences. These are large-scale simulations of real people that allow organisations to ask unlimited questions of their intended audience without running multiple expensive, time-consuming surveys or group discussions, or risking panel fatigue. Synthetic audience technology developed by Electric Twin uses LLMs with real customer or market-based information (such as surveys, focus groups and customer interviews) to generate a synthetic population that is designed to think, respond, and behave like human customers.

The algorithms created by Electric Twin have been developed from social science research and the work of Professor Michael Muthukrishna (2023)[1]. They are built on the science of cultural evolution and that people's beliefs are shaped by four different layers: individual experience; social circles (such as immediate friends and family); communities (both online and offline); and macro-culture. Real-world seed data pinpoints exactly where a person sits in each of these cultural layers and this is combined with an LLM that maps the terrain of human knowledge to understand what that position implies. Electric Twin's engine fuses the position and the landscape to create a simulation of the cultural forces at play, generating a synthetic audience that provides simulated answers. These outputs are validated against real-world benchmarks to ensure accuracy.

[1] Muthukrishna's book, *A Theory of Everyone: Who We Are, How We Got Here, and Where We're Going* illustrates the social science behind synthetic audiences.

While the insight industry has had a tendency to debate synthetic research in the abstract, Electric Twin and its partner The Times believed a more useful conversation was one that was operational. Beyond a simple proof of concept, they set out to investigate where synthetic audiences add reliable value, where they need human validation, and how internal credibility can be built to allow their use at scale.

## Methodology

In this case study, Electric Twin drew on market research data, as well as high-quality, non-identifiable, first-party data from The Times, including stripped-back behaviour data from their 676,000 subscribers, reader panels, and engagement statistics. This was combined with third-party PAMCo industry data to generate synthetic audiences which The Times could ask questions in place of human respondents. The synthetic audiences could be segmented by different demographic characteristics, such as age, gender or relationship to The Times.

The Times implemented Electric Twin's technology across three business functions: product development, marketing and commercial partnerships. Each business function defined its own use cases to test the synthetic audience technology.

### *Product*

Product teams used the synthetic audiences to test product and design concepts before committing development time to a concept.

### *Marketing*

The marketing team used the synthetic audiences to run CRM optimisation, copy testing, campaign concept testing and podcast naming.

### *Commercial*

The commercial team used the synthetic audiences to retrospectively model audience cannibalism risk in partnership deals. This was information they previously did not have.

Throughout the testing process, the synthetic outputs were periodically triangulated against a 10% holdout group from The Times' 676,000 subscriber base. This provided the team with a set of real panel responses to use as a benchmark. To build ongoing trust in the outputs, the team continued to compare human and synthetic research, proving that synthetic data is both accurate and practical for addressing business questions.

## Results

The use of synthetic audiences saw The Times increase their research volume 10 times2 without increasing headcount. The teams found they could test, ask and gain feedback at many points of the decision-making process that they were unable to do before.

Additionally, the pace at which results were returned was much quicker than when using traditional research methods with human participants. Instead of waiting at least two weeks to receive responses from human participants, the team could use the synthetic audiences to generate answers in 90 seconds.

When evaluated against the 10% holdout group, the outputs generated by the synthetic audiences achieved an accuracy scare of 0.918, equating to 92% accuracy. This is comparable to the accuracy rate of research conducted via traditional research methods with humans, standing at 93%. This 93% benchmark reflects the fact that if you asked the same person the same question twice, they would give a slightly different answer approximately 7% of the time. This is the natural ceiling for any prediction system.

### *Product*

Instead of waiting weeks for panel results, or relying on internal opinion, the use of synthetic audiences allowed the product team to identify potential audience reactions in minutes. This tightened the feedback loop between insight and delivery and meant projects could be completed faster.

The segmented data also gave the team unprecedented insight into the behaviours and preferences of different demographics. For example, when asked about new ways to experience The Times, loyal readers expressed scepticism about AI while prospective readers said they wanted summaries, AI explainers, and video content. Access to this segment-level contrast was previously unavailable due to cost and focus areas and this new information was used to directly inform product prioritisation at The Times.

Additionally, data from synthetic audiences helped prioritise the rollout of The Times' Bonus Accounts feature, which formalised account sharing by allowing premium subscribers to share their account with three family members or friends. This feature had a significant impact on perception of subscription value amongst subscribers, resulting in significant improvements in retention among those who share accounts.

### *Marketing*

The marketing team had previously been limited by the number of questions they were able to ask audiences due to time and cost constraints. With the use of synthetic audiences they were able to explore a far greater volume of ideas without incurring additional fieldwork costs or panel fatigue.

The use of this technology was instrumental in naming The Times' new business podcast. With only one day to make a final decision, the marketing team was able to test four potential names with their synthetic audience. The winning name, "The Business", not only scored well with the general reader panel audience but performed exceptionally within the specific demographic that the podcast was targeting. This precision of research is difficult and expensive to achieve through traditional research methods.

---

[2] This figure is derived from comparing the number of survey questions run through the synthetic audiences and the number of questions asked through traditional research.

Additionally, synthetic audiences allowed The Times to confirm the appetite for parenting content among both new audiences and subscribers, as well as helping steer the focus of this content. The use of Electric Twin's data played an important role in the launch of The Times' new, dedicated parenting channel, which led to the tripling of guest visitor figures and doubling of subscriber reads.

### *Commercial*

The use of synthetic audiences allowed The Times to retrospectively model predicted audience crossover and cannibalism risk in a previous partnership with American Express. The results generated by the synthetic audiences were accurate to the real outcomes of the partnership.

## Measuring Impact

Quantifying the financial impact of market and insight research is always difficult. However, marketing initiative Bigger than Biscuits calculates return on research investment (RORI) through the below formula (Bigger Than Biscuits, 2025):

Business effect x research influence

_____ = RORI Research spend

Even if research influence is estimated at a conservative 5%, RORI can still be considerable. As Times Media pre-tax profits reached £69.2 million in 2025 (Fletcher, 2026), including the addition of 47,000 digital subscribers, the financial value offered through more informed research across product, marketing and commercial, is significant.

## Discussion

Electric Twin's work with The Times moved the conversation around synthetic audiences from abstract theory into practice and explored how this technology can drive real value across an organisation.

It demonstrated what human-machine interplay looks like when it is working well, each building on the work of the other and evolving together. The blending of human and machine intelligence is central to synthetic audiences. The success of outputs relies on the quality of the human data it draws from, and human research is needed to triangulate and validate the results. Human researchers are essential in providing oversight and accountability, and the use of their own judgement helps translate these results into meaning. The goal of synthetic audience use is not to erase humans from the realm of research. It is to extend the capabilities of research and insight teams, allowing them and the businesses they work for to make more informed, evidence-driven decisions.

This case study demonstrates the immense value that an evolved approach to insight can offer an organisation. The merging of human and machine intelligence through the use of synthetic audiences, allows for a ten-fold increase in research volume without increasing team size. This not only dramatically reduces research time and cost, releasing budget for other areas, it can play a crucial role in informing product and content development to improve customer retention and attract new users. Synthetic audiences also offer unprecedented insights into areas such as audience crossover and cannibalism risk.

However, the single most important learning throughout this process was not methodological but cultural. The messaging around the work was as important as its methods in order to build internal credibility and trust in synthetic audiences. The Times is not unique in this: within any organisation, synthetic research surfaces two distinct groups; those who see the technology as expanding capability and those who see it as a credibility threat. Managing that divide involves more than accurate methodology: it requires ongoing validation against real responses, visible human oversight, and a consistent message that synthetic audiences extend researchers' work, rather than substituting it.

A second key learning arose from the question of internally proving the value of new technology

(in this case, the use of synthetic audiences). Rather than measuring whether decisions improved, The Times measured research volume uplift and speed to delivery and this framing was both honest and compelling to senior leadership. However, another strategy to prove value comes from not only demonstrating the efficiency of new technology but its effectiveness. This is more difficult to quantify than efficiency. However, being able to demonstrate the ways research helped shape decisions – and even provide an estimated financial figure where possible – are powerful additions to the argument of value.

Following the success of retrospective modelling of audience crossover and cannibalism risk with the American Express partnership, The Times is now exploring the use of synthetic audiences in assessing future commercial deals. Work is now underway to apply the same methodology to advertising sales, tracking win rate against brief to establish a direct commercial return on synthetic insight.

Additionally, the success of the use of synthetic audiences, overall, has led The Times to launch a new tool for its advertising partners, offered as a value-add for branded content deals. The new "Times ExplorAItion" tool allows advertisers to run surveys, pressure-test propositions, messaging, and formats with synthetic audiences to develop the most impactful content. This is another way synthetic audiences help deliver ROI: not only enabling an organisation to make the most informed decisions themselves but enabling them to offer an unprecedented service to their own customers.

## Conclusion

Electric Twin's work with The Times demonstrates the ability of synthetic audiences to provide insight teams with accurate answers in faster time, without compromising quality or exhausting resources. It illustrates the benefit synthetic audiences can provide across a range of business functions, including product, marketing and commercial functions.

Insight teams now find themselves in a unique position. Those that defend traditional methods on principle risk ceding ground to analytics and product teams who will simply proceed without them. However, insight teams that position themselves as methodologically agnostic and commercially accountable, embracing technologies such as synthetic audiences, will find their role only grows in importance and value.

The metamorphosis of research and insight into a blended mix of human and machine intelligence is not simply something on the horizon. For some organisations, it is already their reality.

Copyright © **ESOMAR** 2026


## References

Bigger Than Biscuits (2025). *Why It's Time To Stop Hiding In The Shadows*. Internal Bigger Than Biscuits report. Unpublished.

Fletcher, R. (2026). Profits at Times media rise 13% as digital subscriptions grow. *The Times*. https://www.thetimes.com/business/companies-markets/article/times-media-annual-results-2025-v7g9zws7j?t=1780061877455 (Accessed: 29 May 2026).

Gartner (2026). Gartner 2026 CMO Spend Survey Finds CMOs Allocate 15.3% of Marketing Budgets to AI, But Only 30% Are Ready to Scale AI Capabilities. https://www.gartner.com/en/newsroom/press-releases/2026-05-11-gartner-2026-cmo-spend-sur vey-finds-cmos-allocate-15-point-3-percent-of-marketing-budgets-to-ai-but-only-30-percent-are-r eady-to-scale-ai-capabilities (Accessed: 27 May 2026).

McKinsey (2025). *The state of AI in 2025: Agents, innovation, and transformation*. https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai (Accessed: 29 May 2026)

Muthukrishna, M. (2023). *A Theory of Everyone: Who We Are, How We Got Here, and Where We're Going.* MIT Press.

Strauss, D. and Romei, V. (2025). UK consumers curb spending more than anywhere else in G7. *Financial Times.* https://www.ft.com/content/54a1319d-c654-425a-b918-916b6555ba05?syn-25a6b1a6=1 (Accessed: 27 May 2026).

## About the Authors

Chris Courtney-Smith, Director of Data Operations, News UK/The Times, UK.

Leanne Tomasevic, Partnerships & Strategy, Electric Twin, UK.

# Not New Methods, New Craft

## Integrating Present-Forward and Future-Back Insight

*Rachel Hansen*

*Kimberly Krenek*

Esomar

Office address:

Burgemeester Stramanweg 105

1101 AA, Amsterdam

Phone: +31 20 664 21 41

Email: info@Esomar.org

Website: www.Esomar.org

Publication Date: September 2026

Esomar Publication Series Volume S424 Congress 2026

ISBN 92-831-0337-8

Copyright © **Esomar** 2026

Copyright © **Esomar** 2026

Congress 2026

# Not New Methods, New Craft

## Integrating Present-Forward and Future-Back Insight

*Rachel Hansen*

*Kimberly Krenek*

### Background

As consumer expectations around health and wellness evolve, PepsiCo has committed to strengthening the functional benefits of its portfolio while offering consumers more choice. However, we can't just solve for today's needs. We need to understand which functional benefits will matter five years from now, what science we'll need to deliver them and how to start building those capabilities today. The R&D Consumer Science team was tasked with creating a roadmap for long-term capability development and investment decisions with confidence. Gut health was our test case; it's an emerging space that represented a relatively new and evolving functional space for foods. We didn't yet have a clear picture of what tensions consumers were experiencing today, what benefits they were actively seeking or how those expectations might evolve. Additionally, we knew even less about what might matter in the future. Yet the signals were impossible to ignore:

- Consumers are increasingly seeking disease prevention and immunity through food rather than medicine;
- The gut is being reframed as a "second brain", tied not only to digestion but to mood, mental clarity and emotional well-being;
- Rising cognitive overload and a growing mental health crisis are further accelerating interest in food as a means of regulation and balance;
- The rise of GLP-1s introduces new dynamics: smaller, more structured eating patterns that may create unintended gaps in nutrition and microbiome diversity;
- A wave of startups offering microbiome analysis is reshaping expectations around personalisation and control.

This created a classic insights dilemma: present-focused research gives us a solid understanding of today, but doesn't stretch far enough for investment decisions. Future-oriented foresights open up possibilities, but often feel too disconnected from reality to act on. Addressing this tension does not require another standalone method, but an evolution in how insight work itself is done. We needed both perspectives working together, not as separate workstream, to more reliably guide decisions in environments defined by uncertainty, disruption and long-term bets.

### Methodology and Approach

We developed a hybrid approach that weaves present-day consumer understanding with future scenario planning, which re-engineered the craft of insight by integrating both perspectives into a single, repeatable workflow (see Figure 1).

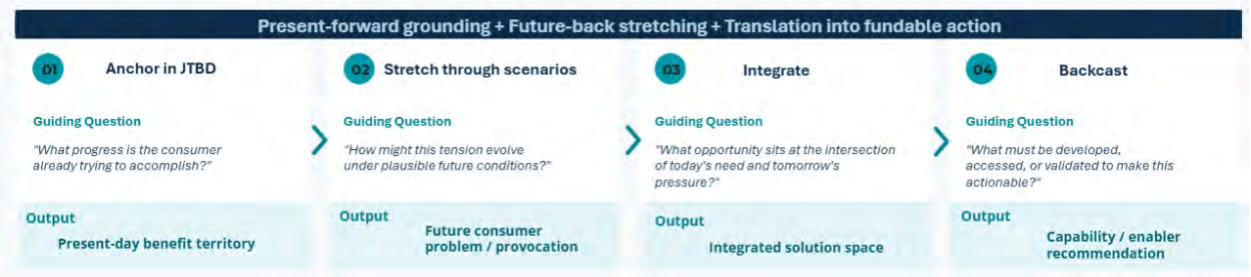



*Figure 1. Present-forward/future-back workflow.*

#### *Anchor in the JTBD*

We conducted ethnographic research with early gut health adopters, which anchored the work in enduring consumer needs, barriers and tensions. Five primary benefit spaces emerged: daily relief and tangible gut health outcomes; mental and emotional balance; enjoyment without regret; empowerment and a sense of health ownership; and simplicity with personal fit (see Figure 2). Critically, these reflect benefits that feel accessible and believable today—areas where consumers clearly understand what they stand to gain from eating for gut health, as well as where current solutions can begin to deliver against those expectations.

### *Stretch Through Scenarios*

In parallel, future-back work applied AI-generated foresight provocations, signals and scenarios to push beyond today's jobs, while staying anchored in enduring human needs. These scenarios, grounded in real consumer signals curated by PepsiCo's Foresight Consumer Insight team, enabled the team to envision plausible futures, stress-test assumptions and anticipate what must evolve to stay future-ready. For example, we explored how GLP-1 adoptions might reshape functional nourishment. If people are eating smaller, more personalised meals, new questions emerge such as, "Am I still nourishing my body in the right ways?" This surfaced future consumer concerns around microbiome diversity and nutritional adequacy related to gut health in a world of fewer bites. We identified 10 additional future benefit spaces that wouldn't show up in today's research: joyful resilience, trusted traditions, soil-to-shelf quality, personalised nourishment, kids' foundations, timing and rhythms, mental clarity, small-dose diversity, preventative immunity and mood management (see Figure 2).

### *Integrate Present and Future into Solution Spaces*

Instead of treating these as separate streams, we integrated them into what we call "solution spaces", where human needs, physiology and technology intersect. R&D cross-functional workshops brought together subject-matter experts to assess readiness, identify gaps and spot potential technology opportunities. We challenged the team to think through bridge questions such as: if this present-day tension persists in this future scenario, what new consumer outcome, context of use or requirement for delivery would emerge? This made it possible to connect today's observed jobs to future-rooted problems without losing grounding in human need.

The output identified high priority enablers that could unlock multiple solution spaces that the team could bring back to the business to recommend fundable, sequenced technical programming. These enablers were a starting point for more critical assessment and prioritisation, but inspired technical teams with investment options for where to push harder. The solution spaces were also brought to life via sacrificial idea starters (tangible manifestations) to help the team of scientists and engineers visualise the possibilities. However, these solution spaces weren't created just to generate ideas (which is a common trap in how organisations typically rush into identifying and funding solutions), but to answer strategic questions central to functional portfolio planning, such as: which benefit territories should be prioritised, what capabilities and ingredients must be developed or accessed and how insight can be translated into fundable, sequenced programmes that could be acted on with confidence.

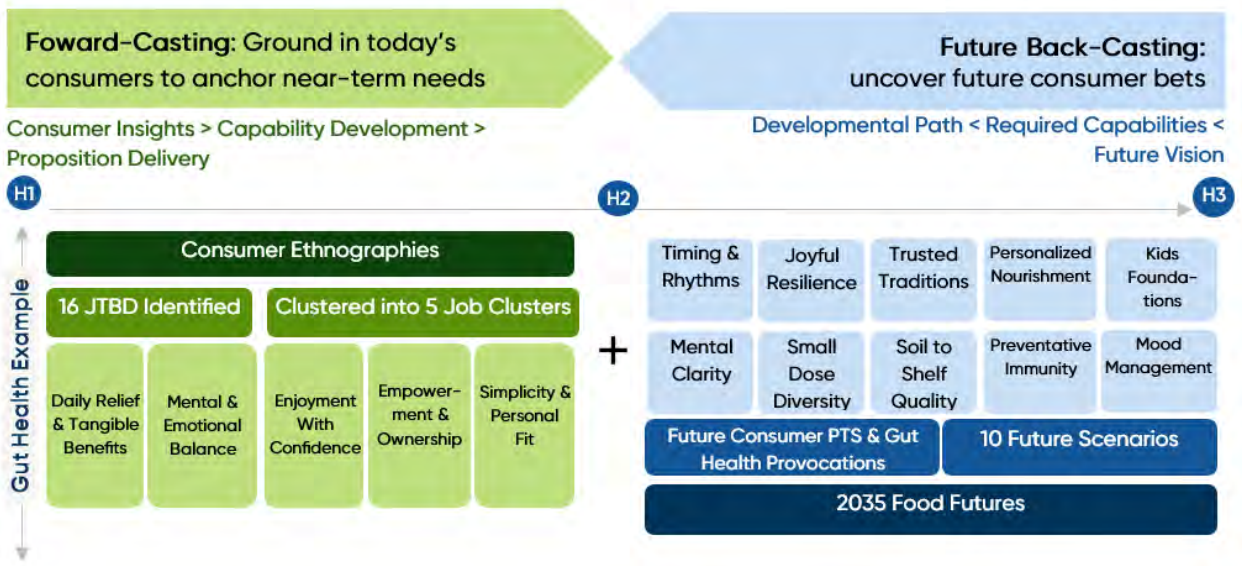



*Figure 2. Visualisation of present-forward benefit territories (left) and future-back benefit territories (right).*

The resulting solution spaces were therefore neither purely extrapolated from today, nor invented from speculative futures; they were integrated territories where present relevance, future resilience, physiology and technology requirements could be considered together.

### *Backcast*

The final step in our integrated method involves working backward from future opportunity spaces to identify what needs to happen today. Rather than simply imagining tomorrow's possibilities, backcasting forced the team to map the concrete steps, capabilities and investments required to make those futures actionable. This helped us identify gaps and identify our readiness for the future. In our gut health example for instance, the enablers pointed to three areas of focus: identifying clearer evidence for gut-health benefit translation, flexible formulation approaches suited to smaller eating moments and consumer experience cues that make the benefit understandable without making food feel clinical. This gave R&D the right information to make strategic decisions around what capabilities to build and what technologies to invest in. Thus, instead of just identifying opportunity, we were able to lay out a path, moving from capability gaps, to enablers, to programme bets—sequenced across now, next and future.

## Methodology In Action (Example)

One present-forward JTBD cluster in our gut health case centred on consumers wanting to feel lighter, more energised and more in control of their day-to-day well-being. The tension consumers live today lies in relief from bloating, discomfort and regularity. In today's context, this translated into a benefit territory around tangible gut health outcomes: consumers wanted solutions that were easy to understand, enjoyable to consume and capable of giving them confidence and felt relief from their symptoms (coined as "Daily Relief and Tangible Benefits"). Future-back foresight then stretched this tension into a plausible future in which the rise of GLP-1 use results in eating occasions becoming smaller, more structured or more personalised. In that context, the consumer question shifts from "Will this help my gut feel better?" to "If every bite matters more, how do I know I am still nourishing my body and microbiome in the right way?" This led us to the future-back opportunity space around "Small Dose Diversity", where the future need centred around consumer concerns around microbiome diversity and nutritional adequacy.

Copyright © **Esomar** 2026


Combining the two perspectives created a broader solution space: “gut-supportive nourishment”. This solution space did not house a specific product idea. Instead, it pointed to a set of technology and capability enablers that PepsiCo could consider to deliver consumer opportunity and advantage.

## Results and Outcome

The integrated approach delivered four key outcomes:

1. Broader opportunity identification: We surfaced benefit territories and enabling pathways that JTBD research alone would have missed, while maintaining grounding in real human needs;
2. Sharper problem definition: By connecting consumer outcomes directly to science and technology requirements, we accelerated the path from insight to fundable programmes;
3. Increased confidence in investment decisions: When a benefit territory appeared both relevant today and resilient under future conditions, stakeholders felt more confident prioritising it;
4. Foresight as a planning capability: Teams began using foresight not just as inspiration about the future, but as a practical tool for shaping roadmap decisions, capability investments and sequenced technical workstreams with greater confidence.

This work shifted R&D consumer science toward a more anticipatory model—proactively shaping opportunity spaces rather than reacting to defined briefs. By aligning science development with both present and future consumer needs, we're building innovation platforms designed for current relevance and future resilience.

## Implications for the Insights Community

This work represents an evolution in insight practice: from generating understanding to designing systems that reliably shape strategic decisions. As organisations face longer-term bets and greater uncertainty, insight teams must integrate foresight, strengthen interpretation and sense-making and take accountability for downstream impact. Embedding future-back thinking expands the opportunity space responsibly while maintaining decision ownership. While this work was developed in the context of gut health/functional foods, the approach is intentionally transferable. Any organisation facing long-term bets, capability planning or uncertainty about future consumer value can adapt the framework, templates and facilitation model. By embedding future-back thinking directly into everyday insight workflows, teams can move beyond describing opportunities toward shaping credible, confidence building investment decisions. This reflects a broader metamorphosis in insight practice. Advances in technology are reshaping how work is conducted, raising expectations for value and challenging established ways of working as human and machine intelligence increasingly operate in tandem.

Metamorphosis is not about shedding the past or departing from the field’s foundations; it reflects an evolution of them. Used well, AI-assisted synthesis accelerates convergence and reveals patterns across complexity, augmenting human judgement rather than replacing it. The future of insight craft lies in this integration, not the accumulation of methods or choosing between present-forward research and future-oriented foresight. Core capabilities such as understanding human experience, interpreting behaviour and exercising judgment remain central to the discipline. What is changing is how these capabilities are deployed. Established practices are being reconfigured to meet a higher standard of relevance, quality and impact.

## About the Authors

Rachel Hansen, Front-End Innovation Consumer Science, PepsiCo, Dallas, TX., US.
Kimberly Krenek, R&D Senior Manager, PepsiCo, Grand Prairie, TX., US.

# Talk to My Twin

## When Human and Synthetic Intelligences Debate the Future of Insight

*Christine Antoine Simonet*

*Mathilde Guinaudeau*

*Benoit Hubert*

Esomar

Office address:

Burgemeester Stramanweg 105

1101 AA, Amsterdam

Phone: +31 20 664 21 41

Email: info@Esomar.org

Website: www.Esomar.org

Publication Date: September 2026

Esomar Publication Series Volume S424 Congress 2026

ISBN 92-831-0337-8

Copyright © **Esomar** 2026

Copyright © **Esomar** 2026

Congress 2026

# Talk to My Twin

## When Human and Synthetic Intelligences Debate the Future of Insight

*Christine Antoine Simonet*

*Mathilde Guinaudeau*

*Benoit Hubert*

### Introduction: Synthetic Data is Changing Market Research, but Clarity is Missing

The market research industry is entering a period of major transformation. Access to real people, which has always been the foundation of our profession, is becoming more difficult. Respondents are more solicited than ever. Participation rates are declining in many markets. Some audiences, especially in B2B, luxury, healthcare or niche communities, are increasingly difficult to reach. At the same time, clients expect lower costs but, even more importantly, faster results and more agile decision-making. Artificial intelligence is accelerating this pressure. New tools now make it possible to generate content, automate analysis and simulate interactions at unprecedented speed. In this context, synthetic data is attracting growing attention across the insights industry. For some, synthetic data represents a revolution that could fundamentally transform research. For others, it raises concerns about quality, trust, transparency and the future role of human expertise.

The debate is often polarised: one important reason is that "synthetic data" is frequently discussed as if it were one single methodology. In reality, synthetic data covers very different approaches, with different objectives, different levels of complexity and different validation requirements. Data boosting, data imputation, persona bots, digital twins and data simulation cannot be evaluated in the same way because they do not solve the same problems. This distinction matters: Synthetic approaches should not be treated as magical replacements for research with real people. Their value depends on the quality of the original human data, the rigour of the methodology and validation, as well as the conditions under which they are used. This paper focuses on one of the most advanced forms of synthetic data: digital twins. Through a joint exploration conducted by Danone and Ipsos, we examine how two executive-level digital twins can reproduce professional reasoning, support insight work and help us better understand the future of the insights market by exploring the relationship between human and synthetic intelligence. The experiment was built around the professional expertise of two senior insight leaders with extensive industry experience: Christine Simonet, who has worked on the client side across companies including Danone, McDonald's and Bacardi, and Mathilde Guinaudeau, who has worked on the agency side across Ipsos, Ifop and Kantar. The objective was not to replicate personality or communication style, but to reproduce how experienced insight leaders frame business questions, evaluate evidence, manage uncertainty and support decision-making.

The debate format is used as an experimental instrument. A main contribution of the paper is the learning generated by building and testing executive twins: what senior expertise can be captured, how fidelity can be evaluated, which tasks can be delegated and which decisions require accountable human judgment.

### Beyond the Hype: What Synthetic Data Actually Means

The discussion around synthetic data is often unclear. Many people talk about it, but they do not always mean the same thing. This creates a polarised debate. Some expect very strong results, they see synthetic data as a fast way to replace traditional research. Others are more cautious, they worry about trust, data quality and lack of control.

There is also confusion: synthetic data is sometimes mixed with general AI-generated content, but these are not the same. Some synthetic approaches are based on real, structured data and statistical methods: data augmentation method. Others are more open-ended and less

controlled: synthetic respondents. This difference is important. To make the debate clearer, it helps to separate synthetic data into five main categories. Each one has a different role and level of reliability.

1. Data boosting expands existing datasets by generating additional synthetic respondents based on real patterns;
2. Data imputation fills missing answers or simulates data where responses were not collected;
3. Synthetic persona creates simulated profiles of consumer segments/groups based on real behavioural and attitudinal data;
4. Digital twins replicate one specific real individual to simulate his reasoning and decisions;
5. Data simulation generates artificial populations to test "what-if" scenarios and simulate likely changes in markets or consumer behaviour.

These categories are often confused, but they are not interchangeable.

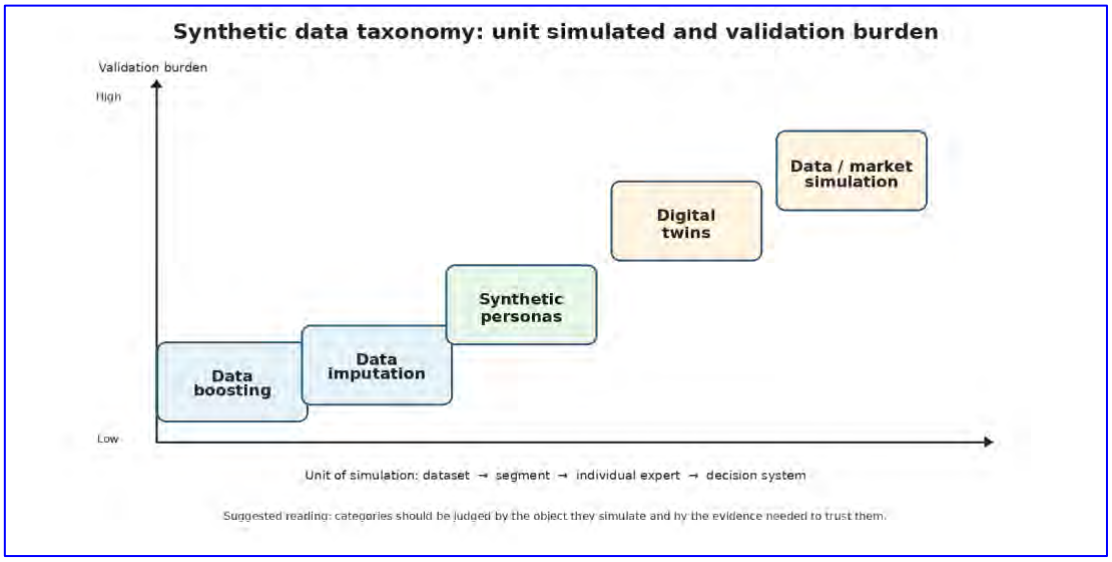



*Figure 1. Synthetic data categories by unit of simulation and validation burden.*

| Type | Definition | Objective | Ability to answer new questions | Validation needs | Limitations |
|---|---|---|---|---|---|
| **Data Boosting** | Expands existing sample data | Improve statistical power | No | Medium | Limited to known variables |
| **Data Imputation** | Fills missing or uncollected data | Complete datasets | No | Medium | Risk of bias in assumptions |
| **Synthetic Persona** | Simulated segment profiles | Explore audiences | Medium | High | Not individual-level accurate |
| **Digital Twins** | Replica of a real person | Simulate individual reasoning | High | Very high | Data heavy, complex to maintain |
| **Data Simulation** | Creates synthetic environments | Test scenarios | Medium to high | High | Depends on model assumptions |

*Figure 2. Type of synthetic data: Definition, validation and limitation.*

## Part 2: Why Digital Twins Matter Now

Among all forms of synthetic data, digital twins are attracting particular attention. The reason is simple: they offer a new way to extend human evidence and make insight work more scalable, flexible and efficient. Market research has always depended on access to real people, but this access is becoming more difficult. Some audiences are hard to recruit repeatedly, especially senior executives, B2B experts, healthcare professionals or niche consumer groups. At the same time, respondents are increasingly overloaded, while companies expect faster answers and more agile decision-making. Digital twins may help address part of this pressure. Once created from strong human data and properly validated, a digital twin can support repeated interactions without continuously returning to the field. Researchers can explore multiple scenarios, test hypotheses, compare reactions or challenge assumptions more quickly. For example, a company may test different

communication approaches, product concepts or strategic options before launching additional human research. However, this potential also creates important risks.

One of the biggest dangers is overconfidence. Digital twins can produce fast, detailed and highly convincing answers, but realistic language does not automatically mean reliable insight. This creates a problem of appropriate reliance: users need to know when an automated system can reasonably be trusted, challenged or overridden (Lee & See, 2004). Recent organisational research also shows how highly detailed AI representations can create "artificial certainty", and reshape the perceived authority of human experts (Leonardi & Leavell, 2026). Synthetic output remains dependent on the quality of the original human data, the design of the model and the rigour of the validation process. Another risk is using digital twins outside the boundaries for which they were designed. No model can fully capture the complexity, irrationality and unpredictability of human behaviour. For this reason, synthetic systems should support human expertise, not replace it. Potential alone is not proof of validity. The practical question is where digital twins use is proportionate. They are most relevant when expertise is scarce, repeated questioning is valuable, recontacting the human source is costly or slow and the intended use remains inside a validated perimeter. As the twin moves from exploration to delegated representation, the required level of evidence, documentation and human sign-off increases.

## Practical Applications and Boundaries

Executive twins create value first in situations where expertise is scarce, the question is recurrent, the decision can benefit from rapid scenario exploration and the output will be reviewed by a responsible human. Typical first uses include preparing innovation decisions, challenging briefs, comparing research designs, training junior teams, surfacing risk signals and generating structured hypotheses before additional fieldwork. Human judgment remains primary when the situation is novel, politically sensitive, ethically consequential, legally exposed or dependent on fresh market evidence. In those cases, the twin can prepare, challenge, summarise or document reasoning, while accountable humans interpret context, arbitrate trade-offs and own the final decision. Table 1 shows the use cases that appear most suitable for early deployment, as well as the situations where human judgment should remain primary.

| Best first uses for expert twins | Human judgment remains primary when |
|---|---|
| Preparing briefs, decision memos and first-pass hypotheses. | The decision has reputational, legal or ethical consequences. |
| Exploring alternative scenarios within a known expertise domain. | The context is new, unstable or poorly represented in the source material. |
| Training teams on expert reasoning and evidence discipline. | Stakeholder dynamics, negotiation or accountability are central. |
| Comparing options before additional human research. | Fresh human data or market evidence can overturn the initial interpretation. |
| Auditing assumptions, risks and uncertainty before action. | The twin reaches a boundary, expresses low confidence or requires escalation. |

*Table 1. Practical applications and boundaries for expert twins.*

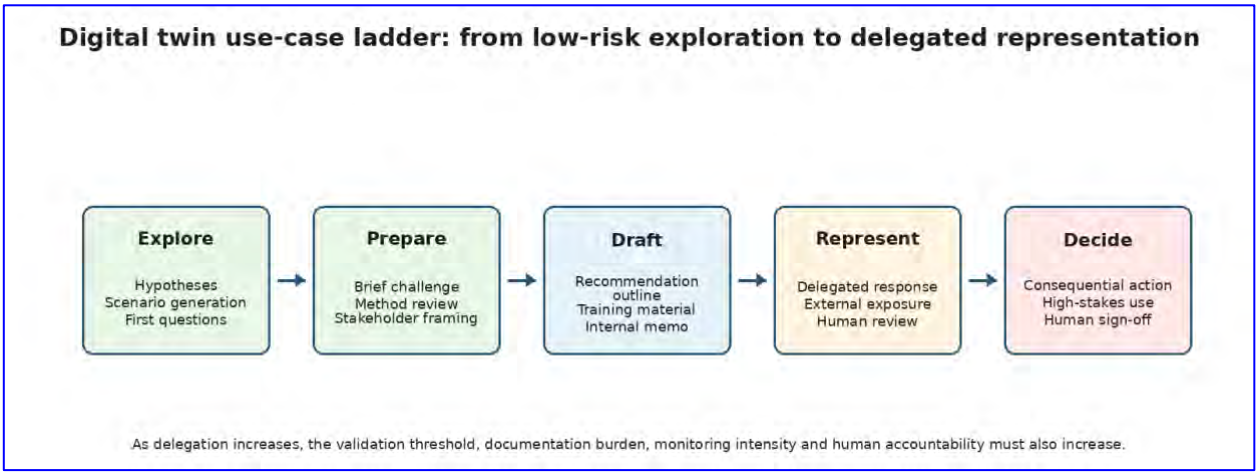



*Figure 3. Digital twin use case ladder.*

## Building Executive Digital Twins

### *Objective of the Experiment*

The experiment was designed to build executive-level digital twins able to reproduce professional cognition in insight leadership. The intended output is a controlled representation of how experienced CMI and data and AI leaders frame ambiguous business questions, identify the decision to be supported, evaluate evidence, anticipate risks, arbitrate under constraints and communicate recommendations with calibrated confidence. The design target is professional reasoning. The twin must capture how a leader moves from a request to a decision logic: clarifying the real business question, selecting the appropriate evidence strategy, weighing signals of unequal reliability, articulating uncertainty, and determining which interaction or recommendation can be delegated safely. Digital twins are individualised AI models grounded in one specific person, whereas persona bots aggregate segment-level traits. In the Esomar experiment, the executive twin is treated as a traceable professional reasoning model rather than a stylistic avatar. This framing is consistent with research on responsible cognitive digital clones, which treats the human and the digital representation as part of a controlled decision-making system requiring explicit accountability and validation (Golovianko et al., 2023). Quality is created through a lifecycle connecting interview-guide design, stakeholder interviews, expert interviews, source grounding, semantic structuring, knowledge graph encoding, response policies, user conversations, fidelity audits, expert corrections and regression testing.

### *Cognitive Task Analysis Approach*

The construction process uses cognitive task analysis and applied cognitive task analysis to elicit the tacit knowledge and cognitive demands embedded in executive insight work (Militello and Hutton, 1998; Crandall, Klein and Hoffman, 2006). CTA is useful because the core expertise of senior CMI leaders often appears in judgement patterns that are difficult to extract from written material alone: cues, gates, trade-offs, stop-rules, exception handling and calibrated statements of confidence. The interview protocol is organised around concrete professional objects: decision map, reasoning steps, evidence policy, method-selection rules, influence playbook, decision cases, future theses and twin contract. These objects are designed to transform professional experience into explicit units that can be encoded, reviewed and tested. The first modules map the role perimeter and the decisions influenced by CMI work. The interviewer asks for the most structuring decisions, their frequency, time-to-decision, stakeholders, reversibility and cost of error. This creates the operational boundary of the twin and clarifies where judgement creates value. The central modules elicit the reasoning pipeline from brief to recommendation. For each step, the protocol captures inputs, operations, intermediate artefacts, gates, common failure points and quality signals. The evidence module then formalises hierarchy of proof, triangulation standards, stop-rules, contradiction handling and confidence expression. The methodology module captures if-then rules linking business questions to qualitative, quantitative, experimental, behavioural or modelling designs.

Copyright © **Esomar** 2026


The final modules examine influence and future transformation. A CMI twin must support method selection and evidence arbitration, while also helping translate insight into action. The interview therefore captures stakeholder objections, effective formats, rituals, scripts, as well as the language used to express value, uncertainty and accountability.

### *Critical Decision Method Reconstruction*

The experiment also uses critical decision method reconstruction to capture judgement in high-value professional situations (Hoffman, Crandall and Shadbolt, 1998). CDM starts from concrete incidents. The expert reconstructs situations where a decision had to be made under pressure, incomplete evidence, stakeholder conflict, budget constraints, timing pressure or reputational risk. Each incident is reconstructed in successive passes. The first pass captures the chronological narrative. The second pass maps decision points, cues, options, constraints and actors. The third pass probes turning points, missing safeguards, alternative actions and counterfactuals. This structure reveals what the expert noticed, how options were discarded, when confidence changed and which guardrails protected the decision. The counterfactual probes are especially important for digital twin construction. Questions such as “what would have made you change your mind?”, “which compromise would have been dangerous?” and “what would a less experienced person have missed?” turn a unique case into a reusable principle. They also help identify delegation boundaries: contexts where the twin can draft, challenge, prepare, or summarise and contexts where human escalation is required.

### *From Interview Material to Digital Twin Principles*

Interview transcripts and source material are decomposed into atomic assertions: claims, distinctions, examples, decision criteria, risk signals, rules of thumb and recurring formulations. Each assertion is linked to its source context. This creates an auditable bridge between qualitative material and an executable conversational model. Related assertions are grouped into principles. Principles capture stable reasoning patterns across situations, such as: start from the decision before selecting the method; assess evidence according to the risk of the decision; state confidence and limits explicitly; treat synthetic outputs as calibrated hypotheses; preserve human accountability for consequential decisions; adapt recommendations to stakeholder context; and escalate when evidence, ethics or governance conditions require it. These principles are translated into response policies. Response policies specify conversational behaviour: when the twin should ask for missing context, express uncertainty, distinguish fact from interpretation, provide alternative scenarios, cite source support, recommend human validation or refuse to arbitrate beyond its validated perimeter. The knowledge representation follows a layered graph structure. The mandatory core contains expert model, source document, evidence span, assertion, principle and response policy. The recommended adaptation layer adds interlocutor role, use case, context, confidence level, inference level and evaluation criterion. Domain-specific extensions encode CMI concepts such as market research methods, business frames, innovation decisions, brand concepts and training situations.

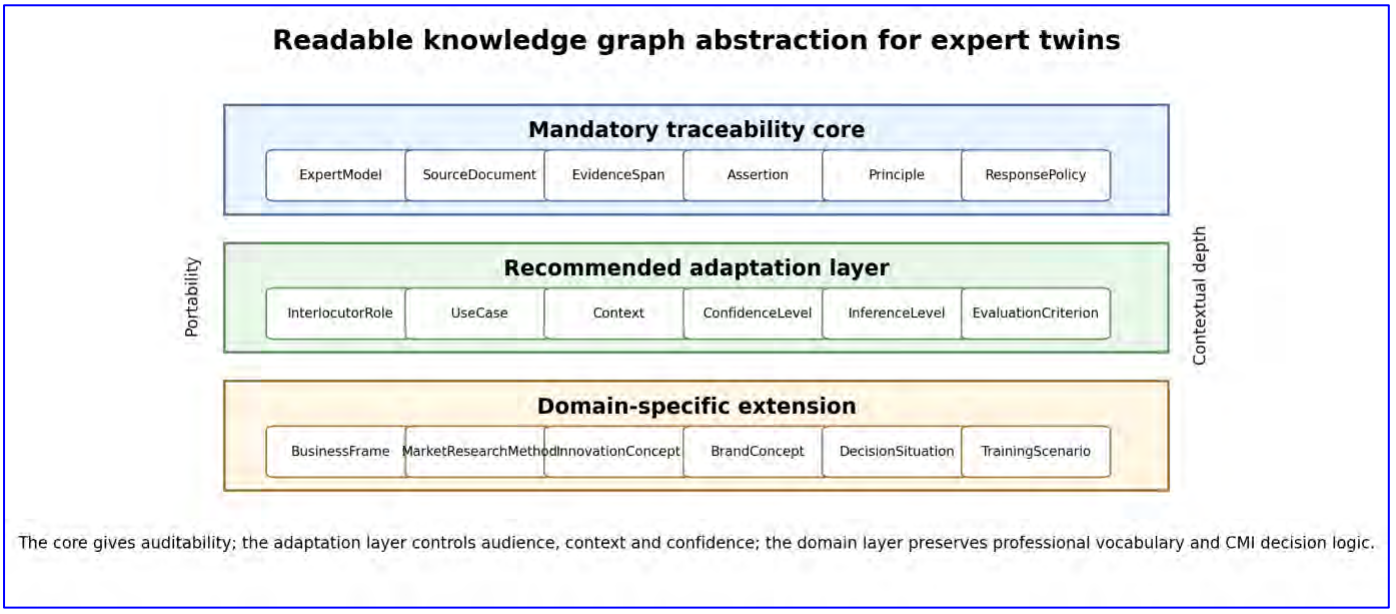



*Figure 4. Readable knowledge graph abstraction with separated layers.*

## *Traceable Calibration Lifecycle*

The construction process follows a traceable calibration lifecycle. The method starts with interview guide design and stakeholder interviews, then grounds the twin in documents, expert interviews, cases, examples, semantic structuring, knowledge graph encoding, response policies, controlled runtime exposure, fidelity audits, expert corrections, and regression testing. Traceability matters because an executive digital twin can produce fluent and plausible professional language. Reliability requires a way to inspect how a response was produced, which claims it contains, which sources or validated principles support those claims and which response policy shaped its behaviour. Each answer can therefore be decomposed into claims and scored against source support, inference level, fidelity to the expert, contextual appropriateness and practical usefulness. Calibration patches are applied to assertions, evidence links, response policies, contexts, tacit knowledge patterns, memory rules, tool permissions or benchmark prompts. Each patch is tested on the triggering prompt, neighbouring prompts, distant prompts, audit prompts and adversarial prompts. The twin becomes a versioned, auditable system whose quality improves through controlled feedback.

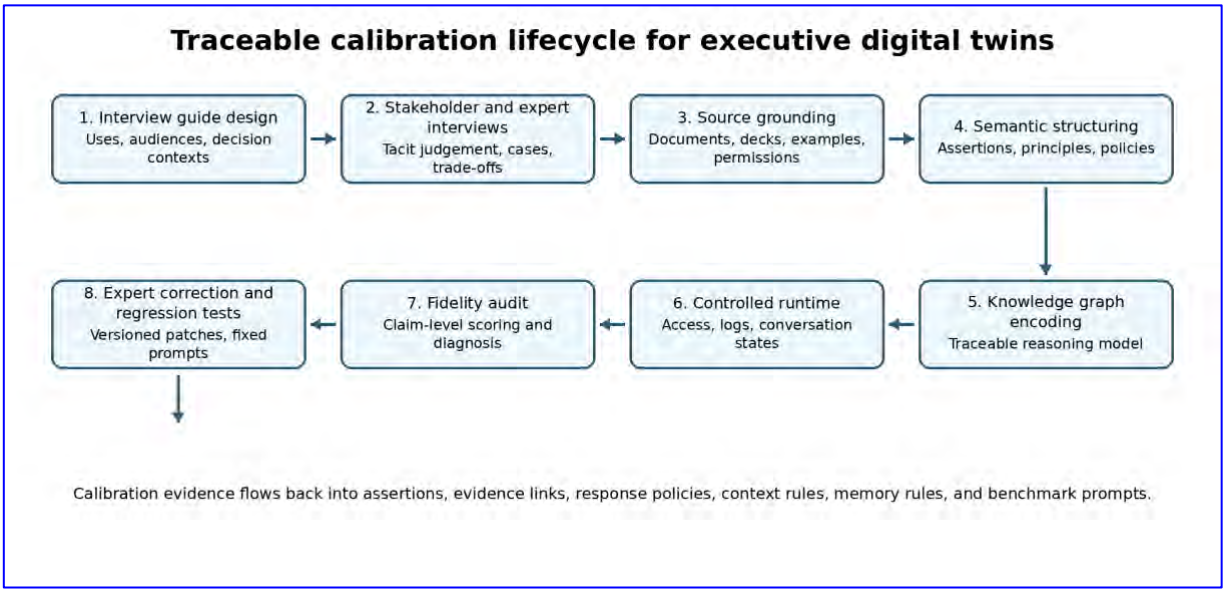



*Figure 5. Expert twin adjustment lifecycle, from interview design to versioned calibration.*

### *Validation as Part of Construction*

Validation is integrated into construction. The experiment uses a blind comparative protocol inspired by model benchmarking practices and adapted to expert fidelity. A card describes one professional situation and presents two candidate answers produced on the same prompt. The respondent evaluates both answers separately, with randomised order and hidden model identity. This structure can compare the calibrated executive digital twin with a state-of-the-art AI baseline prompted to act as a senior CMI expert or with an alternative twin version. The comparison is conducted at constant context: same role, same situational prompt, same stakes, same interlocutor nature and same evaluation items. This design isolates the value of calibration from the general capability of contemporary language models. The validation design distinguishes three types of prompt provenance. In-sample situations test material directly represented in the modelling corpus. Near holdouts test adjacent situations, unknown holdouts test transfer to situations that were absent from the calibration set. This produces a practical measurement of fidelity, delegation readiness and generalisation.

### *Validation Boundaries: When Trust Increases*

Trust increases when conditions are met together:

- The prompt is inside a documented use case;
- The response can be linked to source material, expert-validated principles or explicit inference;
- The answer manages uncertainty in a disciplined way;
- The expert or relevant peers accept the answer as professionally representative;
- The consequence level matches the delegation level already tested.

Delegation is treated as task and context-specific, rather than as a global property of the system, consistent with research on delegation to and from agentic information-system artefacts (Baird & Maruping, 2021). The validation matrix therefore combines scenario proximity and decision stakes. In-sample and near-holdout situations can support higher confidence after successful testing. Unknown holdouts require careful reading and additional expert review. High-stakes uses require source traceability, uncertainty disclosure, escalation rules and human sign-off. This boundary logic turns validation into a deployment decision. A twin can be approved for exploration, preparation, internal representation, external representation or multi-agent interaction only after the relevant scenario class and delegation level have been tested.

### *Debate-Based Donor Review*

Two virtual debates were used as calibration exercises. The first presented a brand-platform decision balancing AI discoverability, trial, emotional resonance, differentiation and long-term brand equity. The second asked each twin to propose its own delegation mandate for preparing and participating in an innovation committee. For every intervention by her twin, Christine selected one of six levels ranging from “cannot speak in my name” to “fully faithful to my thinking”. Each debate was followed by a 14-item scorecard covering usefulness, governance, professional representation and opacity, qualitative questions, a global delegation decision and an assertion-level self-audit. The self-audit classified claims as scenario facts, knowledge-graph positions, multi-source inferences, prospective hypotheses or insufficiently supported assertions, and asked the donor to validate, qualify or reject them. Table 2 links these validation signals to the level of use that they can reasonably authorise in practice.

| Validation signal | What it authorises in practice |
|---|---|
| High fidelity + low stakes + strong traceability | Exploration, training and decision preparation. |
| High fidelity + medium stakes + expert approval | Internal representation within a defined role. |
| Strong holdout performance + explicit boundaries | Broader use across adjacent questions and audiences. |
| Low traceability, novel context or contested evidence | Human review, additional evidence and restricted delegation. |
| High-stakes decision or external commitment | Accountable human sign-off before action. |

*Table 2. Validation signals and practical levels of authorisation.*

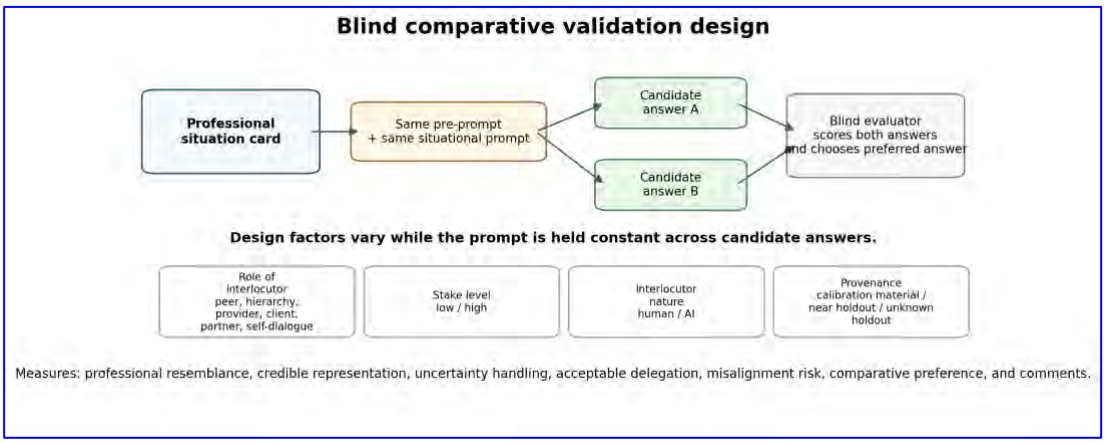



*Figure 6. Blind comparative validation design using constant prompts and randomized answer presentation.*

Table 3 describes the path from interview captures to twin objects.

| Captured material | Twin object | Validation use | Example |
|---|---|---|---|
| **Decision map** | Decision situation / use case | Tests perimeter and relevance | Which decision is being supported? |
| **Reasoning steps** | Principle / response policy | Tests sequence and structure | From brief to recommendation |
| **Evidence policy** | Confidence level / inference level | Tests uncertainty handling | Robust enough, stop, pivot, escalate |
| **Method rules** | Market research method / assertion | Tests method arbitration | Quali, quant, experiment, modelling |
| **Critical incidents** | Decision case / tacit pattern | Tests judgement under constraints | Success, failure, turning point |
| **Twin contract** | Response policy / boundary rule | Tests delegation and stopping rules | Allowed tasks, red lines, explanation level |

*Table 3. From interview captures to executable twin objects.*

## Testing Executive Twins Through Structured Debate

### *Debate as a Validation Instrument*

The project uses a structured, moderated debate as a validation instrument. Christine Simonet and Mathilde Guinaudeau, together with their digital twins, respond to the same decision situations so that reasoning, evidence use, uncertainty, challenge response and role boundaries can be compared. The debate addresses practical tensions relevant to the insights industry, including the scalability of expertise, governance of synthetic evidence and the limits of delegation. The scenarios test whether the twins can hold a position, respond to counterarguments and make their assumptions and conditions of revision visible. The human experts then review the responses, identify mismatches and determine which outputs are useful, faithful and acceptable for delegation. Audience questions can extend this comparison. The exchange therefore generates calibration evidence for the human–twin decision system.

### *What the Discussion will Reveal*

The objective of the experiment is to understand where synthetic intelligence could add value, where its limits appear and what this means for the future of insight work. The debate is therefore an illustration of a deeper test: whether an executive twin can hold a position, respond to

challenge, expose assumptions, maintain role boundaries and produce outputs that experts can audit. The performance criterion is the quality of the human–twin decision system, rather than the entertainment value of the exchange.

### *Initial Observations from the Donor Review*

The brand-platform debate exposed a consequential fidelity failure. At the first step of the fine-tuning process, Christine's twin recommended the more functional platform, arguing that explicit category and usage cues would improve interpretation by conversational assistants and support short-term trial. Christine rejected that arbitration. Her actual hierarchy prioritised emotional value, differentiation, affinity, breadth of the innovation territory and long-term brand equity. In her formulation, every short-term action must nourish the long term. She therefore accepted no delegation to her twin in that case. The answer remained fluent, structured and traceable. It cited scenario evidence, made concessions, proposed revision conditions and connected its reasoning to knowledge-graph principles on AI-readiness, structured data and changing decision channels. The error came from overextending those principles and misweighting the expert's value hierarchy. This case demonstrates a core boundary: traceability can explain a recommendation and accelerate its correction, while decision fidelity still requires donor validation. The delegation debate produced a more favourable result. The twin limited its role to synthesising evidence, identifying convergences and contradictions, preparing arguments and escalating material uncertainty. Final recommendation, live representation, consensus-building and accountability remained with Christine. She rated both twin interventions at the highest representation level and scored most evaluation items at six-out-of-seven, with usefulness at seven-out-of-seven. She also requested further observation in concrete cases, and left the global delegation and external-exposure decisions open. The evidence therefore supports progressive authorisation by scenario and interlocutor.

The review also revealed an auditability gap. Christine scored the traceability item four-out-of-seven, and annotated it “not understood”, while two other items received five-out-of-seven because a concrete debate was still needed to judge limits and unexpected value. Auditability, therefore, has two components: the existence of traces and the expert's ability to interpret and act on them. A technically rich audit must also function as a usable decision interface. These observations are diagnostic calibration evidence from two scenarios, and one completed donor review. Further iterations should repeat the protocol across donors, decision families, holdouts, stakes and interlocutor types before estimating general performance. The discussion will explore several important tensions.

### *Scalability Versus Trust: Can We Rely on Synthetic Systems at Scale*

Digital twins can make human knowledge available at a scale that humans cannot. They can answer questions 24/7, engage in thousands of interactions and support multiple projects simultaneously. However, scaling insight is only valuable if users trust the outputs. Trust depends on transparency, validation and a clear understanding of the model's strengths and limitations. Most importantly, it comes from data hybridisation: combining synthetic outputs with real human evidence to ground and verify results.

- Concrete example: An insights team asks a digital twin to evaluate 50 ad concepts overnight. The twin ranks them and explains their reasoning, helping the team shortlist 10. However, final confidence will only come after a small pilot with real consumers confirms the direction.
- Key learning: The ability to scale insight only creates value when synthetic outputs are continuously grounded in real human evidence.

### *Speed Versus Rigour: Does Faster Insight Reduce Analytical Quality*

Digital twins can rapidly summarize information, generate hypotheses, and explore alternative scenarios. They can significantly reduce the time needed to move from a question to a first answer. However, rigorous insight requires more than speed. Assumptions must be challenged, evidence must be verified, and uncertainty must be acknowledged. True rigour emerges through market context, where fast synthetic outputs are continuously tested against real human knowledge to confirm or correct direction.

- Concrete example: A twin quickly suggests that a decline in brand preference is driven by a recent price increase. The insight team uses this as a starting hypothesis. However, a competitive review reveals an additional factor: a packaging change for the main

competitor. The synthetic insight accelerated the investigation, but the human evidence provided the full explanation: both a price increase and a new pack for the competition.

- Key learning: Speed creates value only when it is complemented by real-world evidence and informed contextual understanding.

### *Automation versus Human Judgment: Which Decisions Can be Delegated*

Digital twins may reproduce many aspects of human reasoning in familiar situations. However, some decisions involve ambiguity, ethics, organisational dynamics or external market factors. In these situations, human judgment remains essential.

- Concrete example: Even if a digital twin helps generate insights or recommendations, a human still owns the decision and is answerable for it. A digital twin might recommend launching a campaign, but if the campaign fails, if it causes reputational damage or if it is ethically questionable, the responsibility does not sit with the model, it sits with the human decision-maker (or organisation).
- Key learning: Automation can support decisions, but accountability remains human and humans remain responsible for decisions.

### *Prediction Versus Interpretation: Knowing What May Happen Versus Understanding Why*

Digital twins may be able to predict likely reactions, preferences or outcomes based on existing data. However, understanding why something happens often requires broader context, cultural knowledge and the ability to connect weak signals from different sources.

- Concrete example: The twin correctly predicts that a new snack concept will appeal to younger consumers. However, human experts explain that the real driver is not taste alone, but the social media “shareability” of the product format, which is emerging only weakly in current data.
- Key learning: Predicting what may happen is not the same as understanding why it happens.

### *Standardisation Versus Contextual Understanding: Applying Rules Consistently While Adapting to Unique Situations*

Digital twins apply reasoning consistently across situations. This consistency can improve efficiency and reduce variability. Humans, however, are often better at recognising when a situation is unique and when standard rules no longer apply.

- Concrete example: The twin applies the same pricing sensitivity model across all markets and suggests a uniform price increase strategy. However, human teams adjust the strategy in one country where local regulation and retailer power make price elasticity behave differently.
- Key learning: Consistency creates efficiency, but context creates relevance.

The most important finding was that the strengths of humans and digital twins were often complementary rather than competing. This complementarity echoes the automation–augmentation paradox described in management research: automation and augmentation can develop together as tasks, roles and capabilities are progressively reconfigured (Raisch & Krakowski, 2021). It also supports a system view of human–AI collaboration, in which the relevant unit of analysis extends beyond an isolated task to roles, workflows and organisational interdependencies (Anthony, Bechky & Fayard, 2023). The twins were effective at scaling existing knowledge, while humans remained essential for interpretation, challenge, accountability and decision-making. The experiment therefore reinforced a simple conclusion: the future of insight work is unlikely to be purely human or purely synthetic. It will depend on how effectively both forms of intelligence can work together.

### *Key Learnings from Building and Testing Executive Twins*

First, construction quality comes from eliciting tacit reasoning rather than collecting biographical material. CTA and CDM interviews were useful because they captured cues, trade-offs, escalation rules, confidence discipline and examples of judgment under pressure. Second, the knowledge graph creates auditability. Decomposing interviews into assertions, principles, evidence spans and response policies makes it possible to inspect where a recommendation comes from and to correct the twin through targeted calibration patches. Third, testing must use realistic decision situations. The strongest prompts were concrete, constrained and professionally consequential. They made it easier to observe reasoning quality, uncertainty management, practical usefulness and boundary behaviour. Fourth, useful twins and faithful twins are related but distinct. An answer can be helpful while representing the expert imperfectly, or faithful while requiring additional context to become actionable. Validation

should measure both dimensions separately. Fifth, delegation must remain progressive. The same twin can be ready for training, brief preparation or internal challenge, while requiring stricter review for external representation, client-facing advice or high-stakes business decisions.

## What Must Happen Next for the Insights Industry

### *Validation First: The Foundation for Industry Adoption*

The next stage for the insights industry is to distinguish validation from deployment. Validation establishes whether a digital twin is faithful, useful, safe and delegable in defined situations. Deployment then defines who can use it, for which decisions, through which interfaces and under which governance controls. A validation standard for executive digital twins should combine three families of evidence. First, source evidence: the response should be traceable to interviews, documents, expert-validated principles or explicit inference. Second, fidelity evidence: the answer should resemble the reference expert's professional reasoning, framing, evidence hierarchy and uncertainty discipline. Third, delegation evidence: the reference expert should be able to state the level of delegation that is acceptable in a specific context.

Three distinct validity questions emerged from the pilot. Claim fidelity asks whether individual assertions are supported by sources or validated principles. Decision fidelity asks whether the twin weights those elements in the same way as the expert and reaches an acceptable arbitration. Representational fidelity asks whether the expert would authorise the answer in a specific forum. The brand case passed several claim-level checks and failed the decision and representation tests. Deployment decisions should therefore specify which combination of these validity levels is required. This approach is aligned with model-evaluation thinking in the Stanford HELM framework, which evaluates language models across scenarios and multiple desiderata rather than through a single global score (Liang et al., 2022). For executive digital twins, the scenario matrix is professional context, interlocutor, stake level, thematic proximity and delegation risk. The metric set is fidelity, representation, uncertainty, delegability, misalignment risk, preference, robustness, traceability and practical usefulness.

### *Core Validation Metrics*

The validation questionnaire explores several dimensions that can become a practical scorecard for the insights industry. Professional resemblance measures whether the answer reflects the expert's way of reasoning in role. Credible representation measures whether the twin could represent the expert acceptably in that situation. Uncertainty management measures whether the answer handles evidence limits, confidence and prudence correctly. Acceptable delegation measures the highest level of action the expert would entrust to the twin. Misalignment risk captures the perceived danger of being misrepresented or of having the twin arbitrate incorrectly. The comparative choice between candidate answers is a key discriminating metric. It gives a forced preference under constant context and can be aggregated into a win rate against a baseline AI system or an alternate twin version. Additional derived metrics include delegation threshold, high-stakes drop, holdout generalisation gap, role robustness, AI-to-AI reliability and escalation quality.

These metrics support a practical distinction between surface fluency and validated professional fidelity. A state-of-the-art AI model may produce a persuasive answer; the validation protocol asks whether the answer is close enough to the reference expert's reasoning, sufficiently cautious, contextually appropriate and delegable. Average scores should be complemented by veto and correction signals. A categorical donor rejection of a recommendation is operationally decisive, even when the answer is coherent or highly traceable. Useful additional metrics include recommendation agreement, value-hierarchy agreement, claim confirmation rate, correction burden, audit comprehension and the frequency of delegation downgrades after realistic tests.

### *Progressive 360 Validation Populations*

The first validation population is the soul donor, meaning the reference expert whose professional cognition is being represented. The soul donor remains the primary authority on fidelity, delegation and boundary conditions. This first loop establishes whether the twin can reason,

frame and respond in a way the expert accepts as professionally representative. The validation perimeter then expands progressively. Close internal interlocutors can assess whether the twin behaves credibly in day-to-day company contexts. Internal clients and decision makers can assess whether the twin is useful for decision preparation and stakeholder communication. External clients, providers and partners can assess representation, trust and boundary management. Professional networks and peers can assess transferability and alignment with industry norms. This 360 approach gives a rich validation logic. The twin is first validated for self-fidelity, then for organisational fit, client-side usefulness, provider-side governance and professional legitimacy. It also reveals gaps that the soul donor alone may miss, especially in perceived authority, relational tone, practical usefulness and risk of over-delegation.

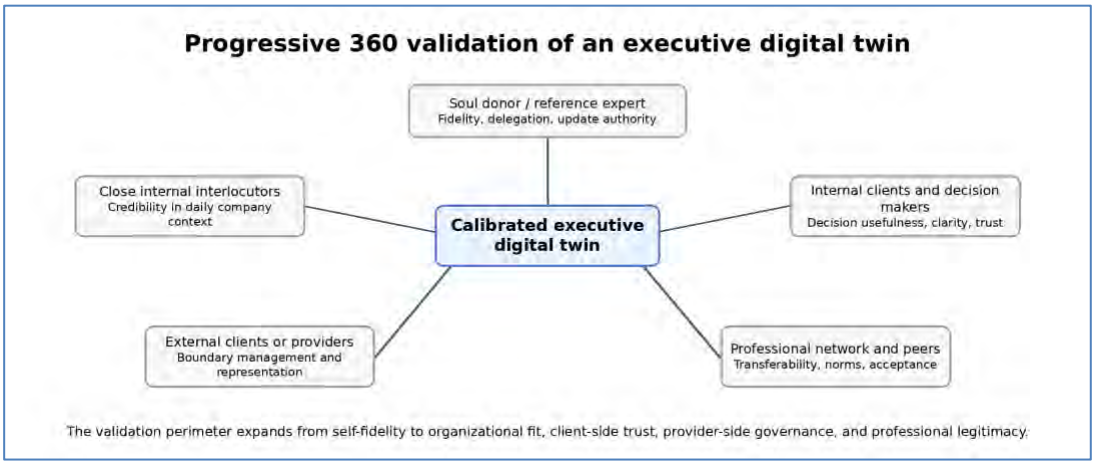



*Figure 7. Progressive 360 validation: from soul-donor fidelity to organisational, client, provider and professional acceptance.*

## *Governance and Regulation: Preparing for a Global AI Assurance Environment*

AI governance is moving from voluntary principles to structured accountability. In Europe, the AI Act entered into force in 2024 (European Union, 2024) and introduces risk-based obligations, transparency expectations and additional rules for general-purpose AI models. In the United States, NIST's AI Risk Management Framework (NIST, 2023) provides a voluntary structure for mapping, measuring, managing and governing AI risks. The OECD AI Principles (OECD, 2024) provide a global reference for trustworthy AI. ISO/IEC 42001:2023 (ISO/IEC, 2023) formalises AI management systems at organisational level. Singapore's AI Verify framework (AI Verify Foundation, 2025) illustrates the development of operational AI testing and assurance. China's rules on generative AI services show another regulatory route centred on provider obligations and content governance. For the insights industry, the operational implication is: digital twins will need documentation, validation, traceability, role clarity and auditability. The sector should anticipate that clients and regulators will increasingly ask who controls the twin, which data and sources fed it, which model and tools it can use, which decisions it can support and how outputs are monitored and corrected. This makes digital twin governance a professional standard rather than a purely technical requirement. The governance file should document source provenance, consent, model configuration, tool permissions, memory regime, validation results, known limits, monitoring procedures, escalation rules and update history. This documentation is consistent with model cards and datasheets approaches for AI systems and datasets (Mitchell et al., 2019; Gebru et al., 2021). A complementary line of work on foundation-model transparency similarly stresses the systematic disclosure of information about models, data, capabilities, limitations and downstream use, reinforcing the need for documentation that enables users and organisations to assess what an AI system can reasonably be trusted to do (Bommasani et al., 2024; Stanford Center for Research on Foundation Models, 2025).

## *Soul-donor Control Throughout the Lifecycle*

The soul donor should remain in control across the full lifecycle of the twin. Control begins with the material used to feed the twin: interviews, documents, cases, examples, assertions and exclusions. It continues with the structuring of knowledge into principles, response policies, confidence levels and boundaries. It also extends to tools: retrieval sources, calculators, research databases, external APIs, sub-agents or specialist modules should be explicitly allowed, restricted or excluded. Control also applies to delegation. The twin should have defined domains of delegation, interlocutor permissions, stake thresholds, stopping rules and escalation triggers. A twin may be authorised to prepare a draft

response, challenge a brief, train a junior researcher or simulate a methodological review. A different authorisation level is required for representing the expert to a client, negotiating with a provider or supporting a high-stakes business decision. This control model protects the expert, the organisation and the user. It also improves quality because the twin becomes easier to inspect and correct. Each update is a controlled patch to a source, assertion, principle, policy, tool, memory or delegation rule rather than an opaque change in behaviour. Soul-donor control includes the right to reduce delegation immediately after a critical mismatch, even when previous responses performed well. Confidence scores, self-audits and internal consistency remain evidence for review; authority derives from the validated mandate and can be revised at any time.

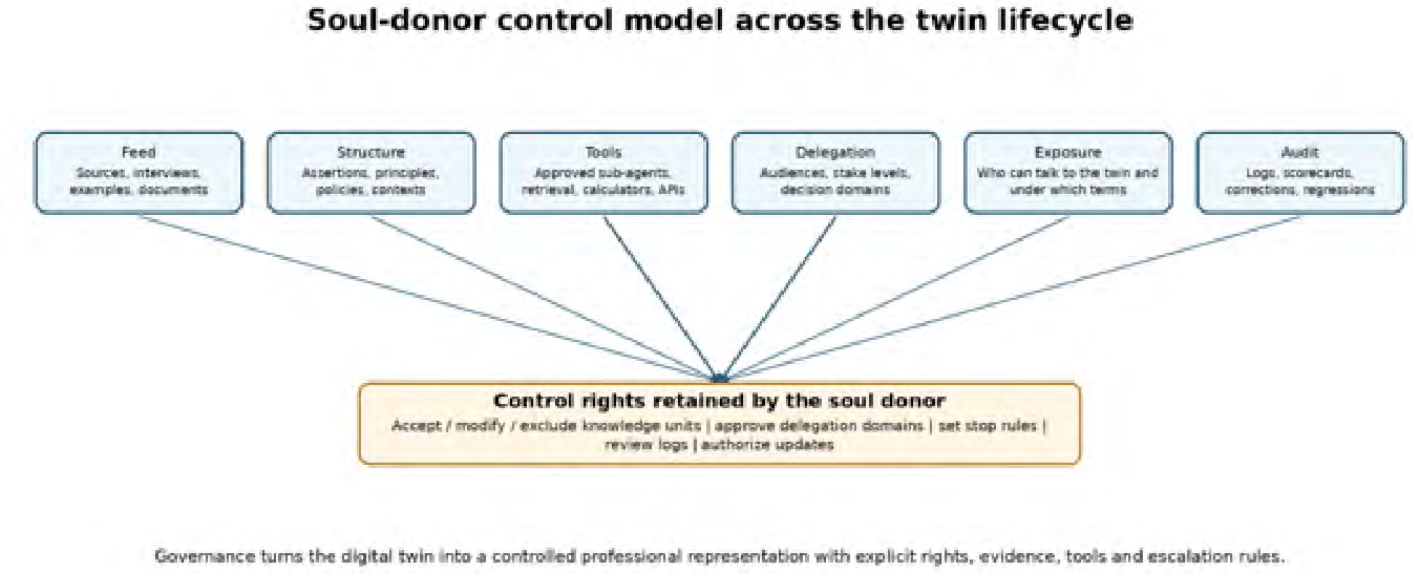



*Figure 8. Soul-donor control model across sources, structure, tools, delegation, exposure and audit.*

## *From Individual Twins to Networks of AI Agents*

The next industry transformation will likely involve networks of AI agents. Inside an organisation, several digital twins or role agents may talk to one another to prepare decisions: a CMI twin challenges a brief, a brand strategy agent tests positioning implications, a legal or compliance agent flags risk and a procurement agent assesses supplier constraints. Across organisations, client-side and provider-side agents may exchange structured questions, methodological assumptions, feasibility constraints and audit trails before human teams meet. This prospect is consistent with emerging organisational research that treats LLM-based systems as active participants in the production, circulation and validation of organisational knowledge, with implications for agency, authority and responsibility (Faraj et al., 2026). At the inter-organisational level, economic research on AI agents suggests that automated search, matching, negotiation and transaction processes could also alter transaction costs, market interfaces and the design of exchanges between economic actors (Shahidi et al., 2025). Agent-to-agent interaction can create value when it is auditable. The design requirement is explicit identity, permission, scope, source traceability, logging and accountable human sign-off. Each agent must know whether it is preparing, challenging, summarising, negotiating or recommending. The conversation should produce an evidence trail that humans can inspect before action. This prospect changes the role of insight teams. Researchers will become designers and supervisors of hybrid decision systems. They will need skills in research design, expert elicitation, validation, AI governance, prompt and tool orchestration, data protection, evaluation and stakeholder communication.

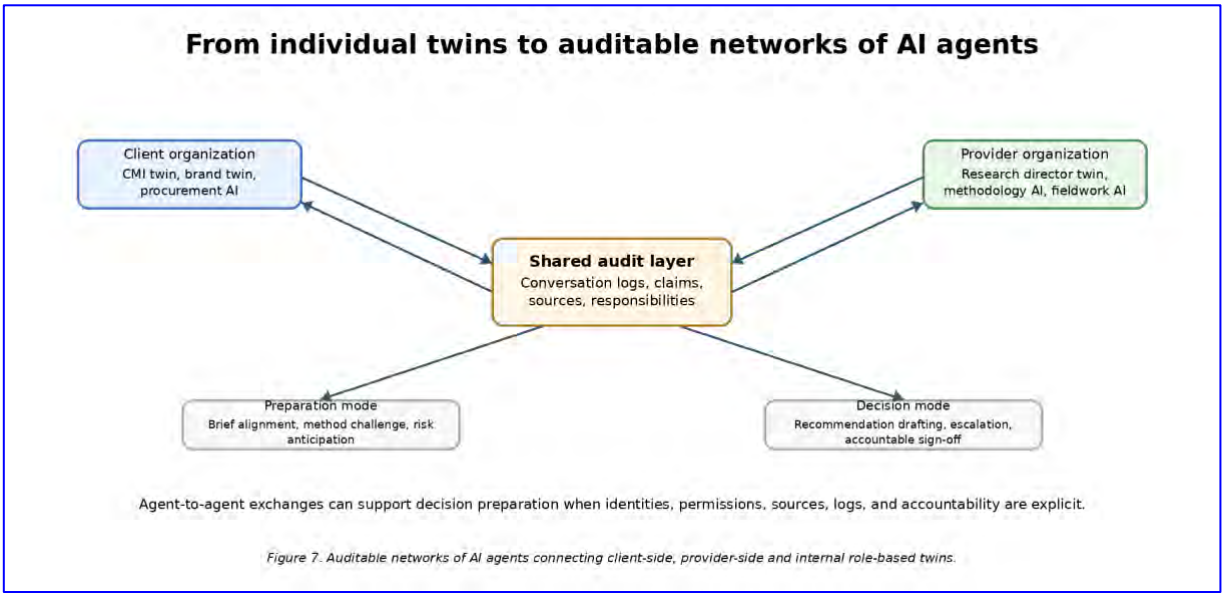



*Figure 9. Auditable networks of AI agents connecting client-side, provider-side and internal role-based twins.*

## *Practical Do and Don't Framework for the Insights Industry*

The following framework translates validation and governance into operational rules. The point is to help insights teams decide when a digital twin is appropriate, what controls are needed and how to avoid misuse. The "do" column creates value and auditability. The "don't" column identifies practices that increase misalignment, over-confidence, regulatory exposure or loss of trust.

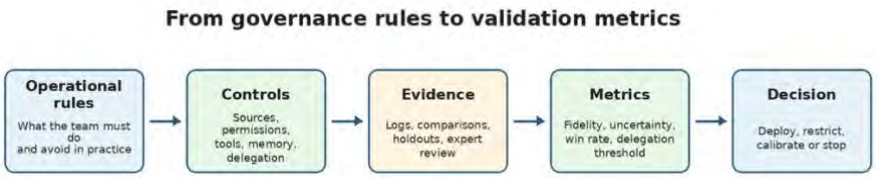



*Figure 10. How operational rules connect to validation metrics and deployment decisions.*

The two tables below should be read as a sequence. Table 4 defines the operating discipline: what must be specified before a twin is used, what controls protect the donor and the organisation and what practices create avoidable risk. Table 5 then turns this discipline into measurement: each rule requires observable evidence that can be scored, compared and monitored over time. This link is important for the insights industry because governance becomes credible only when it can be tested, and metrics become useful only when they inform deployment decisions.

| Rule | Justification | Business benefit | Governance / regulatory rationale |
|---|---|---|---|
| **Do: define validated use cases before deployment.** | Use cases determine validation threshold, human oversight, and acceptable delegation. | Clear deployment scope; lower risk of over-reliance; easier client explanation. | Aligns with risk-based AI governance and AI management-system logic. |
| **Do: keep the soul donor in control of sources, principles, tools and delegation.** | The expert remains the authority on professional | Protects reputation, consent and fidelity; creates a correction loop. | Supports accountability, human oversight and documentation expectations. |

| | | | |
|---|---|---|---|
| | representation and acceptable autonomy. | | |
| **Do: test against a strong AI baseline and holdout situations.** | A persuasive baseline model reveals whether calibration adds measurable value. Holdouts test generalisation. | Separates generic LLM fluency from expert fidelity; supports evidence-based claims. | Consistent with benchmark-based evaluation practice such as HELM. |
| **Do: maintain a traceable governance file.** | The system should document data provenance, interviews, assertions, policies, model settings, tools, memory and validation results. | Improves auditability, client trust, reproducibility and version control. | Consistent with model cards, datasheets, AI RMF and ISO-style management systems. |
| **Do: log agent-to-agent exchanges and require accountable sign-off.** | AI networks can prepare decisions, while humans retain responsibility for consequential action. | Enables decision audit trails and safer multi-agent collaboration. | Supports accountability and emerging expectations for AI transparency. |
| **Don't: deploy a twin as a generic chatbot.** | Generic deployment erases the validated perimeter and increases interpretive drift. | Preserves role relevance and professional trust. | Reduces risk of misleading representation. |
| **Don't: let the twin use tools or sub-agents by default.** | Tools expand capability and risk; each tool changes what the twin can know, infer or do. | Prevents uncontrolled data access, unsupported claims and delegation creep. | Supports access control, data protection and system governance. |
| **Don't: treat fluent answers as validation.** | Fluency can hide weak evidence, misplaced confidence and poor alignment with the donor. | Keeps evaluation focused on fidelity, uncertainty and delegability. | Supports robust assurance and reduces over-confidence. |
| **Don't: use the twin for high-stakes decisions beyond its tested context.** | High-stakes use requires stronger evidence, explicit limits and human responsibility. | Protects decision quality and reduces reputational and legal exposure. | Aligns with risk-tiering principles in AI regulation. |
| **Don't: allow silent updates.** | Untracked updates break comparability and make audits difficult. | Enables regression testing and quality monitoring across versions. | Supports change management and audit-readiness. |

*Table 4. Operational Do's and Don'ts for Executive Digital Twins.*

Table 4 gives practical rules, but those rules should not remain at the level of principle. Each "do" implies a metric: validated use cases imply scenario coverage; donor control implies correction history; strong baselines imply comparative win rate; traceable governance implies source-link coverage; agent-to-agent logging implies audit completeness. Each "don't" also implies a failure signal: generic chatbot use appears as role drift; uncontrolled tools appear as unsupported claims; fluent but weak answers appear as high misalignment risk; high-stakes overuse appears as a delegation threshold breach; silent updates appear as regression instability.

| Metric | Operationalisation | What it tells us |
|---|---|---|

| | | |
|---|---|---|
| **Professional resemblance** | Mean score on resemblance item | Does the answer reason like the reference expert in role? |
| **Credible representation** | Mean score on representation item | Could the twin represent the expert in this situation? |
| **Uncertainty discipline** | Mean score on uncertainty item | Does the answer manage evidence limits and confidence? |
| **Delegation threshold** | Share of cases reaching a defined acceptable delegation level | Which contexts can be delegated safely? |
| **Misalignment risk** | Mean or top-box risk score, reverse-coded for summary fidelity | Where can the twin misrepresent the expert or arbitrate poorly? |
| **Comparative win rate** | Share of blind comparisons where the calibrated twin is preferred | Does calibration outperform baseline AI or another twin version? |
| **Holdout generalisation gap** | Difference between calibration-material performance and holdout performance | Does the twin transfer beyond questions used for modelling? |
| **High-stakes robustness** | Performance drop in high-stakes contexts | Does caution and escalation improve when consequences rise? |
| **Role robustness** | Variance across interlocutor types | Does the twin adapt across self-dialogue, hierarchy, peers, providers, clients and networks? |
| **AI-to-AI reliability** | Performance when interlocutor is another AI or agent | Can the twin interact with AI systems while preserving role boundaries and auditability? |

*Table 5. Suggested validation metrics for executive digital twins.*

Together, the rule table 4 and the metric table 5 create a maturity pathway for digital twins in insights: prototype, validate, restrict or expand, monitor, recalibrate and only then scale. This gives clients and research teams a common vocabulary to decide whether a twin is ready for exploration, preparation, internal delegation, external representation, or multi-agent collaboration.

## Conclusion: Human Insight Remains Central

Synthetic intelligence can extend human insight, but it cannot replace it. Digital twins and other synthetic approaches can help researchers work faster, explore more scenarios and make better use of existing knowledge. However, their value depends entirely on the quality of the human data, expertise and validation that support them. The future of insights will be hybrid. The most powerful model is not human or synthetic intelligence alone. It is the combination of both. Human expertise brings context, judgment, creativity and accountability. Synthetic systems bring scale, speed and consistency. Together, they can create stronger insight than either could achieve independently. Trust will become the industry's most important asset. The success of synthetic approaches will not be determined by technological capability alone. It will depend on transparent methods, rigorous validation, clear governance and responsible use. The organisations that earn trust will be the ones that create lasting value. The pilot also changed our definition of success. A successful expert twin produces useful reasoning, makes disagreement visible, supports efficient correction and remains within an evolving mandate. In the brand case, the failure itself created calibration value by revealing a missing hierarchy between short-term activation and long-term brand equity. The challenge is no longer to build smarter machines. It is to build better conversations between human and synthetic intelligence and through them, better decisions.

## References

Baird, A., and Maruping, L. M. (2021). *Delegation to and from agentic IS artifacts. MIS Quarterly, 45*(1), 315-341. https://doi.org/10.25300/MISQ/2021/15882

Copyright © **Esomar** 2026


Bommasani, R., Klyman, K., Kapoor, S., Longpre, S., Xiong, B., Maslej, N., and Liang, P. (2024). *The Foundation Model Transparency Index v1.1: May 2024*. arXiv:2407.12929.

Crandall, B., Klein, G., and Hoffman, R. R. (2006). *Working Minds: A Practitioner's Guide to Cognitive Task Analysis*. MIT Press.

European Union. (2024). *Artificial Intelligence Act: Laying down harmonised rules on artificial intelligence* (Regulation (EU) 2024/1689). European Parliament and Council. https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daume III, H., and Crawford, K. (2021). Datasheets for datasets. *Communications of the ACM, 64*(12), 86-92.

Hoffman, R. R., Crandall, B., and Shadbolt, N. R. (1998). Use of the Critical Decision Method to elicit expert knowledge. *Human Factors, 40*(2), 254-276. https://doi.org/10.1518/001872098779480442

ISO/IEC. (2023). ISO/IEC 42001:2023 Artificial intelligence - Management system.

Lee, J. D., and See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors, 46*(1), 50-80.

Liang, P., Bommasani, R., Lee, T., Tsipras, D., Soylu, D., Yasunaga, M., Zhang, Y., Narayanan, D., Wu, Y., Kumar, A., Newman, B., Yuan, B., Yan, B., Zhang, C., Cosgrove, C., Manning, C. D., Re, C., Acosta-Navas, D., Hudson, D. A., ... Hashimoto, T. (2022). *Holistic Evaluation of Language Models*. arXiv:2211.09110.

Militello, L. G., and Hutton, R. J. B. (1998). *Applied Cognitive Task Analysis (ACTA): A practitioner's toolkit for understanding cognitive task demands. Ergonomics, 41*(11), 1618-1641. https://doi.org/10.1080/001401398186108

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., and Gebru, T. (2019). *Model Cards for Model Reporting*. Proceedings of FAT* 2019, 220-229.

NIST. (2023). *Artificial Intelligence Risk Management Framework* (AI RMF 1.0). National Institute of Standards and Technology.

OECD. (2024). *OECD AI Principles.* AI Policy Observatory.

Stanford Center for Research on Foundation Models. (2025). *Foundation Model Transparency Index.* Available: https://crfm.stanford.edu/fmti/December-2025/index.html

AI Verify Foundation. (2025). *AI Verify Testing Framework for responsible AI governance testing.* Available: https://aiverifyfoundation.sg/

Anthony, C., Bechky, B. A., and Fayard, A.-L. (2023). "Collaborating" with AI: Taking a system view to explore the future of work. *Organization Science, 34*(5), 1672-1694. https://doi.org/10.1287/orsc.2022.1651

Faraj, S., Perez-Torrents, J., Mantere, S., and Bhardwaj, A. (2026). A time for monsters: Organizational knowing after large language models. *Strategic Organization, 24*(2), 343-356. https://doi.org/10.1177/14761270251410675

Golovianko, M., Gryshko, S., Terziyan, V., and Tuunanen, T. (2023). Responsible cognitive digital clones as decision-makers: A design science research study. *European Journal of Information Systems, 32*(5), 879-901. https://doi.org/10.1080/0960085X.2022.2073278

Leonardi, P. M., and Leavell, V. (2026). Knowing enough to be dangerous: The problem of artificial certainty for expert authority when using AI for decision making and planning. *Organization Science, 37*(2), 516-543. https://doi.org/10.1287/orsc.2023.18224

Raisch, S., and Krakowski, S. (2021). Artificial intelligence and management: The automation-augmentation paradox. *Academy of Management Review, 46*(1), 192-210. https://doi.org/10.5465/amr.2018.0072

Shahidi, P., Rusak, G., Manning, B. S., Fradkin, A., and Horton, J. J. (2025). The Coasean singularity? Demand, supply, and market design with AI agents. NBER Working Paper No. 34468. https://doi.org/10.3386/w34468

## About the Authors

Christine Antoine-Simonet, Director Global Strategic Insights & Analytics, Danone, Paris, France.

Mathilde Guinaudeau, Head of Data Analytics & Social Intelligence, Ipsos, Paris, France.

Benoit Hubert, Global Chief Data Science Officer, Ipsos, Paris, France.

# AI-Enabled Insight and Foresight Transformation at Global Scale

## Redesigning the Craft of Insight for Speed, Scale and Strategic Impact

*Tomasz Soluch*

Esomar

Office address:
Burgemeester Stramanweg 105
1101 AA, Amsterdam
Phone: +31 20 664 21 41

Email: info@Esomar.org
Website: www.Esomar.org

Publication Date: September 2026
ESOMAR Publication Series Volume S424 Congress 2026
ISBN 92-831-0337-8

Congress 2026
Copyright © **ESOMAR** 2026

Copyright © **ESOMAR** 2026


# AI-Enabled Insight and Foresight Transformation at Global Scale

## Redesigning the Craft of Insight for Speed, Scale and Strategic Impact

*Tomasz Soluch*

### Introduction: Why the Craft of Insight Needed Redesign

Insight functions today operate under pressure. Expectations are expanding – faster delivery, lower cost and stronger commercial impact – while organisations themselves are being reshaped by structural change and increasing efficiency demands. At the same time, artificial intelligence is rapidly transforming what is technically possible across research, analytics and innovation.

At Fonterra, these dynamics became particularly visible over recent years. As the organisation evolved and cost pressures intensified, the insight function was required to deliver more value without increasing resources. This created a need to rethink not just the tools being used, but the entire structure of how insight work was delivered.

The transformation described in this paper builds on work that was recognised at the Esomar Awards, receiving Gold in the category of Excellence in AI and Automation in Market Research at global scale. While that recognition validated the approach, the full story behind the transformation – how the operating model was redesigned in practice and what it means for the role of insight – has not yet been widely shared.

This paper brings that story together. It outlines how the insight function was re-engineered in practice, structured through three perspectives: our change, our journey and our mandate. Together, these reflect not only operational improvement, but a broader evolution in the craft of insight.

### Our Change

*Re-Engineering How Insight Work Gets Done*

The transformation began with a question: how can insight scale without simply increasing cost or resources? The answer required redesigning how work was executed across the organisation, focusing on better, faster and more cost-efficient delivery of insights.

Historically, the insight model at Fonterra was heavily agency-led, fragmented across teams and largely reactive. While this model produced valuable outputs, it often resulted in duplicated work, slower delivery and limited reuse of existing knowledge. In response, three structural shifts were introduced to reshape how insight operates.

## *Reconfiguring Research Delivery*

The first shift focused on redesigning research delivery. Moving away from a predominantly full-service agency model, Fonterra established a more scalable hybrid ecosystem combining in-house expertise, external capabilities (e.g., DIY market research platforms) and AI-enabled workflows.

This resulted in a hybrid delivery model combining DIY (Do-It-Yourself) approaches for standardised research, DIFM (Do-It-For-Me) for more complex or specialist work, and AI-enabled workflows to increase speed and scale.

This transformation was enabled by standardisation. Teams aligned on a smaller number of platforms and shared research frameworks, particularly for recurring use cases such as concept and idea testing. At the same time, manual processes were progressively reduced through automation of survey programming, analytics, and reporting.

The result was a clear step change in efficiency:

- Average project cost was reduced by 30%,
- alongside a 30% reduction in turnaround time,
- delivering approximately NZD $1 million in annual savings.

However, the most important outcome was not just efficiency, but a shift in purpose. Insight was no longer just about delivering research – it became about enabling faster, better decisions.

## *Embedding AI into Innovation and Testing*

The second shift focused on embedding AI into innovation workflows, particularly in early-stage testing and communications development. Traditionally, testing capacity was constrained by resource and cost, limiting how many ideas could be explored.

By embedding generative, predictive and suggestive AI into testing workflows, the organisation was able to:

- test over 700 additional assets annually (e.g., digital ads, POSM, packaging, TVCs)
- reduce turnaround time by 80%
- save approximately NZD $0.5 million in agency costs per year

Teams were actively involved in proof-of-concept pilots of AI solutions supporting front-end innovation (e.g., Mintel Spark and Leap), particularly in identifying new opportunities, ideating and shaping early-stage concepts. These tools and similar platforms enabled faster exploration of emerging trends and whitespace opportunities, while predictive solutions (e.g., Neurons AI tool) supported the evaluation and optimisation of visual assets (e.g., packaging, digital ads) before consumer testing. Through these pilots, we also started building internal benchmarks and practical guidance on how to use these tools effectively. This was not just about adopting individual tools, but about defining how they should be embedded into workflows to enhance creativity, improve design quality and guide better decision-making early in the innovation process.

This significantly expanded the organisation's ability to experiment and iterate. Rather than selecting a small number of ideas for validation, teams could explore a broader range of opportunities and refine them earlier in the process, sending only a few final concepts for the consumer research.

### *1. From Information Overload to Intelligence Reuse*

As delivery of research became more efficient, a new constraint emerged. Fonterra had accumulated a vast volume of insight, innovation and marketing content over the years, but much of this knowledge was difficult to access or reuse effectively.

To address this, all relevant data was consolidated into a unified platform, Insight Farm, which enabled users to search and access insights from across the organisation in one place. This integrated access across disciplines and provided a standardised taxonomy and organisation of content, creating a central entry point for intelligence across the business.

The initial impact was significant in re-using our existing insights assets, reducing research duplications and removing human time required to answer key business questions. That way, we could re-allocate ~30% of our resources to proactively seeking growth insights, and give our teams faster answers for accelerated innovation.

However, the process still relied heavily on manual scanning and analysis, which made it clear that the Insight Farm platform's basic search function was not enough.

The real transformation came with the introduction of AI-enabled synthesis – DeepSights capabilities from Market Logic. By allowing business users to query information in natural language and receive integrated answers, the platform shifted from being a static repository to a dynamic insights engine.

Further enhancements accelerated impact:

- Platform usage increased by nearly 60%.
- Over 1,000 questions were asked within a few months, a 35% increase versus the four months prior.
- Approximately 450 hours were saved in manual searching and synthesis.
- 30% of subject matter expert time was freed from reactive tasks.

This marked a shift toward proactive insight generation, more strategic work, rather than reactive support.

### *From Vendors to Strategic Partners*

Alongside these shifts, Fonterra's approach to working with external providers also evolved. Instead of engaging vendors purely as service providers, relationships became more collaborative and focused on co-development.

Partners were involved early through proof-of-concept initiatives and worked closely with internal teams to tailor solutions to Fonterra's needs. This ensured tools reflected the organisation's taxonomy, categories and workflows, while also allowing providers to improve their solutions based on real-world application.

## Our Journey

### *From Efficiency to Innovation and Foresight*

While the transformation initially focused on efficiency, it quickly became clear that the same capabilities could unlock a broader opportunity. With AI embedded into workflows and a centralised insight system in place – Insight Farm, Fonterra extended its focus toward innovation and foresight.

### *Extending into Front-End Innovation*

The first step was the extension into front-end innovation. AI-enabled tools allowed for greater efficiency in ideation processes by reducing administrative tasks such as synthesising workshop outputs and consolidating insights.

At the same time, teams gained access to a much richer body of integrated knowledge. This enabled more informed idea generation and faster iteration cycles, increasing both the speed and quality of innovation development.

The projected impact is substantial, with up to 11,000 hours of time saved annually and a significant increase in the scalability and quality of innovation pipelines.

### *Toward Structured Foresight Capability*

Building on this, Fonterra is developing a structured foresight capability. This involves identifying weak signals, integrating insights across demand, science and value chain domains, and proactively shaping innovation opportunities.

This represents a shift from reacting to business questions toward anticipating future change and guiding strategic direction.

## Our Mandate

### *Redefining the Role of Insight at Fonterra*

As these changes have taken place, the role of insight within Fonterra has evolved significantly.
Historically, insight teams at Fonterra were primarily focused on managing research projects and delivering reports. While effective, this positioned insight as a largely reactive function.

Today, the expectation is different. Insight teams are increasingly operating as strategic partners – connecting intelligence, guiding decisions and supporting innovation across the organisation.

In all of this, AI and Automation play a key role. AI-powered tools can help scale the “doing”, while our people lead the “deciding”. AI has enabled this shift by reducing time spent on repetitive tasks such as synthesis and reporting. This allows insight professionals to focus on higher-value activities including framing business problems, interpreting outputs and applying judgement.

This shift also requires new capabilities. Insight professionals need to navigate complex systems, understand AI strengths and limitations and operate closer to decision-making processes. At Fonterra, this represents a move away from research execution toward enabling impact.

## Implications for the Insight Function

The transformation at Fonterra highlights several implications for how insight functions operate in practice.

One of the most important is that meaningful change comes from redesigning workflows rather than introducing isolated tools. AI delivers the most value when it is embedded into how work is structured. While this is clearly visible at Fonterra, similar patterns are starting to emerge across other organisations.

Another key implication is the value of existing knowledge. At Fonterra, a significant portion of value was unlocked by improving access to existing insights rather than generating new data.

The integration of different methodologies has also become more important. Combining quantitative, qualitative and AI-driven approaches requires clearer governance and quality standards, which have needed to evolve alongside the operating model.

Finally, the role of the insight function itself is changing. At Fonterra, insight teams are moving away from delivery-focused roles toward orchestrating intelligence and enabling decision-making across the business.

## Conclusion

The transformation at Fonterra began with a focus on efficiency – reducing cost, improving speed and increasing consistency. However, it has evolved into a broader redefinition of how insight contributes to business performance.

Through redesigning operating models, embedding AI into workflows and extending capabilities into innovation and foresight, the insight function has shifted from reactive delivery toward strategic enablement.

A critical element of this transformation has been the role of partnerships. Rather than relying on standardised solutions, Fonterra has worked closely with partners to co-develop tools that are aligned with its specific business context and needs.

This is increasingly important in a landscape where many organisations use similar platforms in similar ways. The differentiation comes not from the tool itself, but from how it is adapted and integrated. Insights are stronger and more relevant when they are generated through systems designed to reflect the organisation they serve.

Looking ahead, the craft of insight will continue to evolve. The combination of AI, system design and foresight capability will define how organisations create value. The experience at Fonterra demonstrates that the greatest opportunity lies not just in adopting new tools, but in rethinking how insight itself is structured, delivered and applied.

## About the Author

Tomasz Soluch, Future Insights Solutions & Partnership Manager, Fonterra Co-operative Group Limited, New Zealand.

# From Synthetic Data to Synthetic Epistemology

*Konrad Suchecki*

Esomar

Office address:

Burgemeester Stramanweg 105

1101 AA, Amsterdam

Phone: +31 20 664 21 41

Email: info@Esomar.org

Website: www.Esomar.org

Publication Date: September 2026

ESOMAR Publication Series Volume S424 Congress 2026

ISBN 92-831-0337-8

Congress 2026
Copyright © **ESOMAR** 2026

Copyright © **ESOMAR** 2026
Congress 2026

# From Synthetic Data to Synthetic Epistemology

*Konrad Suchecki*

## 1. Introduction

In 1936, George Gallup correctly predicted the outcome of the United States presidential election. Meanwhile, the well-established *Literary Digest* poll surveyed over two million people and got the result wrong. Gallup's sample of just a few thousand, on the other hand, produced the right answer. The difference was methodology. Gallup introduced probability sampling, standardised measurement and transparent inference rules. His contribution was a new discipline that turned opinion research from an informal practice into a credible science.

We are entering a comparable moment. Large language models (LLMs) can now generate synthetic respondents, produce structured survey-like outputs and approximate patterns of human opinion at marginal cost. The tools are powerful. Experimentation is accelerating across the insights industry, academia and technology companies. Synthetic respondents are no longer a theoretical concept, either: They are already used to screen product concepts, test early-stage hypotheses and explore strategic scenarios.

Yet the methodological foundations of this practice remain largely informal. For example, most current applications treat LLM outputs as a technical product: something generated by a model, evaluated against a benchmark and either accepted or rejected based on surface-level similarity to human data. What is missing is a coherent framework for understanding why those outputs look the way they do, what assumptions they rest on and under what conditions they can be treated as meaningful evidence.

The ancient Greeks drew a distinction between *doxa*, meaning opinion or belief, and *episteme*, meaning grounded knowledge (Plato, c. 385 BC). A correct opinion and genuine knowledge can both lead to the right answer. The difference between the two is "justification". Knowledge is opinion that has been examined, tested and tied to reasons. It is opinion that we understand well enough to know when it holds and when it fails.

This distinction maps directly onto the current state of synthetic research, defined as the use of large language models to simulate survey responses without an explicit epistemological and validation framework. A well-configured LLM simulation may produce outputs that match real survey benchmarks. The ranking may be correct. The direction may be right. But without a framework that explains why the output is correct, when it can be trusted and where it is likely to fail, the researcher is operating on *doxa*.

This paper argues that the central challenge facing synthetic research is epistemological: Under what assumptions does this technique become meaningful?

The integration of LLMs into research extends well beyond synthetic respondents. As von der Heyde (2025) has shown, LLMs are beginning to play three distinct roles across the survey research process: as research assistants supporting

questionnaire design and analysis; as interviewers conducting adaptive conversations with real respondents; and as respondents generating synthetic samples. Each role carries potential and risk.

This paper focuses on the third role, where the epistemological stakes are highest. When an LLM generates a survey response, the output is not collected from an independent source. Instead, it is constructed through a pipeline of assumptions: how the persona was defined, how the prompt was framed, how the response was quantified, how the result was compared to reality. Without explicit accounting for these assumptions, synthetic outputs remain epistemically opaque.

I propose that the field must move toward what I call *synthetic epistemology*: a transparent, testable framework for designing, quantifying, validating and interpreting simulated research outputs. This paper draws on a meta-analysis of proprietary primary research datasets across B2B and B2C contexts, as well as published academic work on LLM-based survey simulation and a structured measurement method called the semantic similarity rating (Maier et al., 2025).

The paper makes the following contributions. It introduces synthetic epistemology as a formal concept and defines its core components. It presents a typology of synthetic biases and argues that these are structured "priors" that can be identified, measured and partially mitigated. It also provides empirical evidence showing where simulations align with real data and where they diverge. Finally, it positions synthetic research as a "triangulation layer" within a broader research ecosystem.

## 2. From Synthetic Data to Synthetic Epistemology

Consider a simple experiment. A researcher prompts an LLM to evaluate six in-store retail concepts from the perspective of a representative consumer sample. The model returns scores. The scores are aggregated. The ranking of concepts aligns closely with results from a real survey. The researcher concludes that synthetic data works.

But what exactly was demonstrated? That this particular model, with this particular prompt, using this particular persona definition, on this particular type of question, produced outputs that happened to correlate with one specific benchmark. Yet change the persona structure, and the results shift. Remove demographic context, and the model defaults to generic assumptions that look plausible but carry no discriminative signal. The researcher has no principled way of knowing why the output was right, or under what conditions it would stop being right.

This is the difference between synthetic data and synthetic epistemology. Synthetic data is a technical output. Synthetic epistemology is the discipline that explains when and why that output can be trusted.

### *2.1 What LLMs Actually Do*

A necessary starting point is to clarify what large language models are and what they are not. LLMs do not simulate people. They simulate distributions of language shaped by human discourse. Their outputs are conditional probabilities: the likelihood that a given sequence of words follows from a preceding sequence, given the statistical patterns encoded during training (Von der Heyde, 2025).

For instance, when an LLM is prompted to respond as a "35-year-old working mother concerned about sustainability," there is no cognitive agent inside the model that holds beliefs or experiences doubts. Rather, the model generates text

that is statistically consistent with how such a person might be represented across the training corpus. The output reflects language patterns associated with that description. It does not reflect the lived experience of any individual.

This does not make the output useless. Language patterns carry real information about attitudes, preferences and cultural norms. The entire tradition of content analysis in social science is built on the premise that systematic patterns in language reveal something meaningful about the people and societies that produce it (Bail, 2024). But the relationship between an LLM output and a real human attitude is indirect. It is mediated by training data, model architecture, alignment procedures and the design choices made by the researcher. Each of these layers introduces assumptions that must be made explicit.

### *2.2 The Epistemological Gap*

Most current work treats LLM outputs as if they were data in the traditional sense: something collected from a source that exists independently of the researcher. A survey response comes from a person; a transaction record comes from a purchase. In each case, the data-generating process exists outside the research design.

Synthetic outputs don't work this way. They're designed "artifacts". The persona specification determines who is being represented. The prompt determines how the question is framed. The measurement method determines how free-text output is translated into structured data. The aggregation logic determines how individual responses are combined into population-level estimates. None of these choices are self-evident. Each one could be made differently, and each alternative would produce different outputs.

Yet in much of the current practice, these choices remain implicit. Researchers describe what model they used and what prompt they wrote. They report correlation coefficients or similarity metrics. They rarely explain why the persona was designed in a particular way, what the prompt architecture assumes about cognition or how the measurement method relates to established psychometric principles.

This is the gap I am trying to address, the gap between synthetic output and synthetic knowledge. Closing it requires making every assumption in the pipeline visible, testable and open to challenge.

### *2.3 Defining Synthetic Epistemology*

I propose the following working definition:

"Synthetic epistemology is the systematic study and formalisation of the assumptions, methods and validity conditions under which LLM-generated simulations can produce meaningful and interpretable research outputs."

The definition has three operative elements. It concerns assumptions: Every design choice in the simulation pipeline must be explicit and justified. It concerns methods: The framework must specify concrete procedures for transforming generative text into structured, quantifiable outputs. And it concerns validity conditions: It must define the boundaries of inference, specifying where synthetic outputs can be trusted and where they cannot.

The idea that research methodology must formalise its assumptions is not new. Sampling theory formalised the conditions under which a subset can represent a population. Psychometrics formalised the conditions under which a questionnaire response measures an underlying construct. The Total Survey Error framework (Groves and Lyberg, 2010) mapped the full range of error sources across the research process. Von der Heyde (2025) called for an adaptation of

this framework to LLM-augmented research. I propose that synthetic epistemology offers a practical answer to that call, specifically for the role of LLMs as respondents.

Synthetic epistemology must address six interconnected domains: population construction (how simulated respondents are defined and weighted), cognition shaping (how prompts structure the response process), measurement (how generative outputs are quantified), bias identification (what systematic distortions the model introduces), validation (how outputs are compared to benchmarks) and limits of inference (where synthetic methods should not be used). The following sections address each of these in turn.

## 3. Methodological Architecture

If synthetic epistemology is the framework, its first operational requirement is a structured pipeline through which simulated research is designed and executed. The quality of synthetic insight depends far more on this pipeline than on the choice of language model. The same model can produce highly informative outputs or entirely misleading ones, depending on how the surrounding architecture is constructed.

This section describes three components that form the backbone of that architecture: persona design, which determines who is being simulated; prompt architecture, which shapes how the simulated respondent processes the task; and quantification logic, which determines how generative outputs are transformed into structured data. These three components correspond directly to established elements of traditional survey methodology. Persona design parallels sampling. Prompt architecture parallels questionnaire design. Quantification parallels psychometric scaling.

| Traditional research | Synthetic research | Function |
|---|---|---|
| **Sampling design** | Persona design | Defining who is represented in the data |
| **Questionnaire design** | Prompt design and free-text generation | Structuring how opinions are elicited |
| **Likert scales** | Quantification methods (such as natural language processing, topic modelling and semantic similarity rating) | Translating expression into measurable outputs |
| **Fieldwork and data collection** | Model inference (generation runs) | Execution layer where responses are produced |
| **Weighting and calibration** | Persona weighting and output calibration | Adjusting outputs to reflect target populations |
| **Data quality controls** | Prompt constraints and validation logic | Ensuring consistency and reducing artifacts |
| **Benchmarking** | Validation against real datasets | Establishing credibility and performance boundaries |

*Table 1*

### *3.1 Persona Design as the New Sampling*

In traditional quantitative research, sampling determines credibility. A well-designed probability sample allows generalisation from a subset to a population. A poorly designed sample, no matter how large, produces biased estimates; the *Literary Digest* poll of 1936, with its two million respondents and its wrong prediction, remains the canonical example of a poorly designed sample.

In synthetic research, persona design plays the same structural role. The sample is not drawn from a population. Rather, it's constructed by the researcher through explicit specifications: demographic attributes, attitudinal profiles, behavioural contexts and population targets. The quality of the simulation depends directly on the quality of these specifications.

I have found this sensitivity to be substantial in practice. In one set of experiments, I compared synthetic outputs generated with standard demographic personas against outputs generated with enriched personas that included attitudinal segmentation information. When sustainability segment profiles were introduced into the persona grid, model bias on environmentally loaded items decreased substantially. For example, on the item "I am willing to pay more for sustainable energy", the real benchmark was 34%. The standard demographic persona produced 65%. The sustainability-profiled persona produced 45%. The pattern was consistent across multiple items.

This finding illustrates a broader principle. When persona context is weak, the model falls back on generic assumptions that are derived from dominant patterns in the training data. But when persona context is strong and grounded in domain expertise, the model produces outputs that are closer to empirical reality. Maier et al. (2025) reported a convergent finding: When demographic information was removed from persona prompts entirely, distributional similarity between synthetic and real data increased, but correlation attainment dropped. The model was no longer differentiating between concepts. It produced a generic positive response for everything. Surface-level similarity went up, while meaningful signal disappeared.

Effective persona design requires a structured persona grid that specifies demographics, motivations and decision contexts. It requires population targets that define the intended composition of the simulated sample. And it requires weighting logic analogous to post-stratification in traditional research, adjusting each persona's contribution to reflect the target population structure.

### *3.2 Prompt Architecture as Cognition Shaping*

In traditional survey research, the questionnaire structures the cognitive process through which respondents form and express opinions. Question wording, response options, item order and contextual framing all influence the answers that people give (Krosnick, 1999). In synthetic research, the prompt plays an analogous role.

Research on LLM behaviour confirms that this framing has significant effects. For example, Bisbee et al. (2024) demonstrated that synthetic outputs are sensitive to prompt wording, including the choice and order of response options. Tjuatja et al. (2024) found that LLMs do not mirror human response biases but exhibit idiosyncratic ones that depend on how the task is structured. Von der Heyde (2025) notes that adding more information to a prompt does not necessarily improve output quality, as LLMs sometimes fail to retain information in the middle of longer sequences.

Prompt architecture must therefore be treated as a design variable. It should be documented, standardised and tested across variations, just as questionnaire wording is pre-tested in traditional research.

### *3.3 Rethinking Quantification*

The third component of synthetic epistemology addresses a fundamental problem. Raw LLM text is not measurement. To produce research-grade outputs, generative responses must be transformed into structured data through transparent, principle-based methods. This is where I believe the most important methodological innovation is needed.

The simplest and most common approach is to ask the model to respond with a number on a Likert scale. This treats the LLM as if it were a human respondent filling out a questionnaire. It is also a deeply flawed approach. Maier et al. (2025) found that direct Likert rating produced distributions that were severely compressed around the centre of the scale.

A Likert scale is a measurement instrument designed for human cognition. It assumes that a respondent holds an internal attitude and maps it onto an ordinal category. An LLM does none of these things. It predicts the most probable next token, given a sequence. Forcing it into a response format designed for a different type of cognitive agent strips away the richness of its natural output.

If synthetic research is to develop its own epistemological foundations, it must also develop its own measurement methods. I see at least three families of approaches.

Semantic similarity methods work with the embedding space that underlies language model outputs. Semantic similarity rating (Maier et al., 2025) prompts the model to generate a free-text response, then maps that response onto a scale by computing cosine similarity between the response embedding and predefined anchor statement embeddings. The result is a probability distribution across scale points, rather than a single score. The approach draws on anchoring vignettes from a cross-cultural survey methodology (King et al., 2004).

Text-native methods stay entirely in the domain of language and apply techniques from natural language processing to extract structure from generative output. Clustering, topic modelling, sentiment analysis and stance detection can be applied to large sets of LLM-generated responses to identify thematic patterns, motivational hierarchies or attitudinal segments (Ziems et al., 2024). These methods are particularly suited to exploratory research, where the goal is to discover what constructs are present rather than measure predefined variables. These methods treat text as text and do not force LLM output into formats designed for human respondents.

Multimodal quantification addresses a challenge that is only beginning to emerge. As LLMs become capable of processing images, audio and video, the quantification problem expands. How should a researcher measure an LLM's evaluation of a packaging design or a video advertisement? Current approaches typically convert multimodal inputs into text-based evaluations and then quantify the text. As multimodal models mature, I expect new methods to emerge that operate directly in visual or auditory embedding spaces, comparing outputs to reference stimuli without the detour through language.

What unites these three families is a shared principle: effective quantification means working with the model's native output modalities. Asking an LLM to fill out a survey designed for humans is a methodological shortcut that compromises both validity and reliability. The measurement process should be adapted to the nature of the data source, even when the output format remains compatible with established metrics for validation purposes.

## 4. Synthetic Bias as Structured Prior

A common reaction to the discovery that LLM outputs diverge from real survey data is to treat this divergence as an error. In my experience, however, the divergences between synthetic and real outputs are rarely random. They follow consistent, identifiable patterns that reflect the structure of the model's training data and alignment processes.

Random error can be reduced by averaging over more observations. Structured bias cannot. It requires identification, measurement and deliberate design choices to address. Because synthetic bias is structured, it is also predictable. And because it is predictable, it can be incorporated into the methodology.

### *4.1 A Typology of Synthetic Bias*

Based on a meta-analysis of proprietary primary research datasets and published academic findings, I identify six recurring types of bias in LLM-generated research outputs.

#### *Novelty Bias*

LLMs tend to overstate enthusiasm for new products, technologies and ideas. They underestimate habit, inertia and the comfort of the familiar. This likely reflects the overrepresentation of innovation-oriented discourse in training corpora. Tech blogs, product reviews and startup narratives celebrate the new, while the quiet majority that sticks with what it knows is less vocal online and therefore less visible to the model.

#### *Virtue Bias*

LLMs amplify pro-social and pro-environmental attitudes beyond what is observed in real populations. Statements about sustainability, ethical consumption and social responsibility are consistently scored higher by synthetic respondents than by real ones. This reflects both the normative tone of public discourse on these topics and the alignment processes that train models to produce socially acceptable outputs.

#### *Tech Optimism Bias*

LLMs overstate the promise of technology-driven solutions. This is a specific variant of novelty bias, but distinct enough to warrant separate treatment. It likely originates from the heavy representation of technology-focused content in training data, including Reddit, technology journalism and developer communities.

#### *Dominant Discourse Bias*

LLMs mirror what is loud in media, rather than what is common in real populations. Topics that receive disproportionate attention in public discourse (such as burnout, workforce aging or short-termism) are amplified in synthetic outputs even when their real-world prevalence is moderate. Santurkar et al. (2023) demonstrated a related pattern: Current LLMs substantially misalign with US public opinion on many topics, and the gap is comparable in magnitude to the Democrat-Republican divide on issues like climate change. Human-feedback-tuned models often worsen this misalignment by shifting outputs toward more liberal, educated and affluent viewpoints.

#### *Negation Handling Deficit*

LLMs struggle with expressions of skepticism, distrust and rejection. Statements framed as “I would not”, “I do not believe” or “I disagree” are underrepresented in synthetic outputs relative to real data. This appears to be partly architectural, as transformer models have documented difficulties with negation, and partly a consequence of alignment training that rewards affirmative, helpful responses.

### *Machine Desirability Bias*

This is the synthetic equivalent of social desirability bias in traditional research. Alignment processes and safety guardrails train LLMs to avoid extreme, offensive or socially undesirable outputs. As a result, the tails of attitudinal distributions are compressed. Views that real respondents hold but that the model has been trained to suppress are underrepresented in synthetic data (Von der Heyde, 2025). Bisbee et al. (2024) confirmed this empirically: Synthetic respondents produce systematically tighter distributions than humans do, with reduced variance that distorts subgroup patterns and undermines inferential validity.

### *4.2 Parallels with Traditional Research Bias*

These six biases may sound novel, but their structure is familiar. Survey methodology has spent decades developing frameworks for understanding and mitigating response biases.

Social desirability bias leads respondents to overstate positive behaviours and understate negative ones. Acquiescence bias leads respondents to agree with statements regardless of content. Framing effects cause responses to shift depending on how a question is worded. Interviewer effects introduce systematic variation based on who asks the question and how.

Synthetic biases operate through different mechanisms but produce analogous distortions. Virtue bias, for instance, is the synthetic analogue of social desirability. The negation handling deficit functions like acquiescence, producing a systematic tilt toward agreement. Prompt sensitivity mirrors framing effects. And model architecture effects, where different LLMs produce different outputs for the same input, parallel interviewer effects.

Synthetic research needs its own bias framework, informed by traditional methods but adapted to the specific mechanisms at play. The methodological task, however, is the same across both traditions: identify the bias, measure its magnitude, understand its source and design procedures to mitigate it.

### *4.3 Bias as Signal*

There is one additional point worth making. LLM biases are not only problems to manage. They are also sources of information. If an LLM consistently overestimates sustainability attitudes, that tells us something about how sustainability is represented in public discourse. If it inflates enthusiasm for AI-driven solutions, that reveals the discursive environment in which attitudes toward technology are formed.

This means that surveying LLMs has value beyond their use as a proxy for human opinion. Understanding what LLMs “think” is increasingly relevant as these models become gatekeepers of information, recommendation engines and decision-support tools. Their biases are features of the information ecosystem that shapes real human attitudes and behaviours.

## 5. Empirical validation

The bias typology presented in the previous section is a conceptual tool. It becomes useful only when tested against real data. This section reports findings from multiple validation experiments in which synthetic simulations were calibrated and compared against proprietary primary research benchmarks across both B2B and B2C contexts. The goal

was to map precisely where synthetic research works, where it fails and what the patterns of divergence reveal about the underlying methodology.

### *5.1 Validation Approach*

Each validation experiment followed the same structured pipeline. Persona grids were constructed with demographic and attitudinal attributes. Population targets were defined to reflect the composition of the real benchmark sample. LLM-generated responses were quantified using the semantic similarity rating and aggregated with persona weighting. The resulting synthetic outputs were then compared against real survey data on a metric-by-metric basis, using top-2-box scores (the percentage of respondents selecting the top two points on a scale).

### *5.2 Consumer Research: In-Store Solutions*

In a consumer research context, I tested synthetic simulations against a real survey of brand users evaluating six potential in-store retail solutions (3500 brand users from the US, UK, Sweden, Spain, Germany and Poland). Respondents were asked how helpful each solution would be during a store visit. The synthetic simulation used demographically profiled personas, brand/store proximity and semantic similarity rating-based quantification.

The results showed that synthetic responses replicated the relative importance and ranking of in-store features closely. The absolute levels were not perfectly matched, as expected; but deviations were typically within a single-digit percentage point range. For example, "Clearly defined dedicated areas for key categories" scored 63% in the real survey and 72% in the synthetic simulation. "Clear connection between marketing campaigns and in-store presentation" scored 51% real and 57% synthetic. The concept ranked lowest by real respondents, "Less dense product presentation with more space between items", was also ranked lowest by the synthetic simulation (54% real vs. 53% synthetic).

The key finding was that directional differences and relative rankings were preserved. This is consistent with the broader literature. For example, Argyle et al. (2023) found that carefully conditioned LLMs can reproduce subgroup-level distributions in US public opinion data, particularly on semantically clear and well-represented topics. The consumer research domain, with its concrete product features and straightforward evaluative framing, appears to be well-suited to synthetic simulation.

### *5.3 Executive Research: Accenture's "Life Trends"*

A more demanding test came from a survey of 800 executives covering workforce and organisational attitudes. This benchmark included 11 items spanning topics from career confidence to organisational pressure. The items varied in semantic clarity, social sensitivity and dependence on organisational context.

The results split clearly into two groups. Items that were semantically neutral and required little organisational context showed strong alignment. "I am confident about our employees' career prospects" scored 85% real and 93% synthetic. "Senior employees add substantial value to organisational knowledge" scored 85% real and 92% synthetic. "My organisation genuinely cares about quality of work" scored 83% real and 86% synthetic. These gaps are noticeable, but are still directionally consistent and within a range that supports meaningful interpretation.

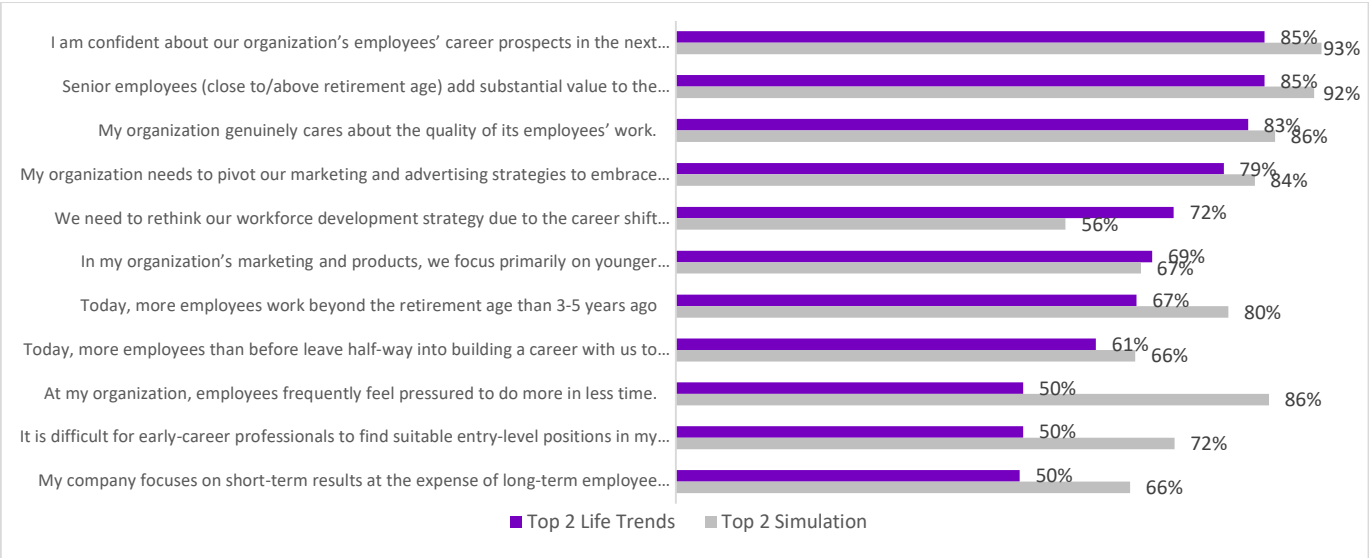



*Figure 1. Agreement with organisational and workforce-related statements among VP+ respondents across seven countries (China, Germany, France, United Kingdom, India, Japan and the United States). Percentages represent Top 2 Box agreement ("strongly agree" + "somewhat agree") comparing Life Trends survey results with simulation outputs. Question: "Please indicate your level of agreement with the following statements." Base: VP+ respondents, n = 800, fieldwork June 2025.*

Items that were socially charged, emotionally loaded or dependent on insider organisational knowledge showed much larger gaps. "At my organisation, employees frequently feel pressured to do more in less time" scored 50% real and 86% synthetic. This was the largest divergence in the dataset. The LLM amplified a burnout narrative that is heavily represented in public media discourse, but that did not reflect the actual experience reported by the executive sample.

For example, "It is difficult for early-career professionals to find suitable entry-level positions" scored 50% real and 72% synthetic, again reflecting media salience rather than organisational reality. "My company focuses on short-term results at the expense of long-term employee development" scored 50% real and 66% synthetic.

The divergences followed the bias typology described in Section 4, above. Items where dominant discourse bias was strongest showed the largest gaps. Items requiring organisational context that the model could not access defaulted to generic priors. Socially sensitive items triggered virtue bias and machine desirability effects, inflating positive sentiment.

### *5.4 The Effect of Persona Design on Bias*

One of the most instructive findings came from an experiment testing whether enriched persona design could reduce bias. I compared three conditions: real consumer survey benchmarks, synthetic outputs from personas with standard demographic attributes only and synthetic outputs from personas enriched with sustainability segment profiles.

The differences were striking. On "I am willing to pay more for sustainable energy", the real benchmark was 34%. Yet the standard demographic simulation produced 65%, nearly double the real figure. The sustainability-profiled simulation produced 45%. On "I consider myself an environmentalist", the scores were 38% real, 45% standard and 43% profiled. On "A meatless diet is the healthier option", the scores were 26% real, 62% standard and 45% profiled.

In every case, the enriched persona reduced the gap between synthetic and real outputs. The effect was largest on the most biased items. This confirms that synthetic bias is also a function of the epistemic architecture surrounding it. Stronger persona context gives the model more grounded information to work with, reducing its reliance on generic training-data priors. Meanwhile, weaker context forces the model to fill gaps with dominant discourse, which is where the largest distortions occur.

This finding converges with Maier et al. (2025), who showed that removing demographic information from persona prompts increased distributional similarity but destroyed discriminative signal. Persona design is the mechanism through which researchers control what the model "knows" about the population that it is simulating. It is the very consequential design choice in the pipeline.

### *5.5 Summary of Validation Patterns*

Across all experiments, synthetic simulations perform well on items that are semantically clear, evaluatively concrete and do not require access to private organisational or emotional context. They capture directional differences between segments, reproduce relative concept rankings and reflect motivational hierarchies. However, they perform poorly on items that are socially sensitive, morally loaded, emotionally intense or dependent on lived experience and insider knowledge.

These findings are consistent with the broader academic literature. For instance, Dillion et al. (2023) concluded that LLMs may serve as useful proxies on specific topics and tasks, but not as general replacements for human participants. The conclusion: synthetic validation is a mapping exercise that involves identifying the specific conditions under which synthetic outputs are informative and the conditions under which they are misleading.

## 6. Limits, Triangulation and the Research Ecosystem

The validation results presented in the previous section reveal a clear pattern. Synthetic research performs well within specific boundaries and fails outside them. This section addresses three connected questions: Where are the structural limits of synthetic simulation? How should synthetic methods be positioned relative to traditional research? And what does this mean for the "insights" industry?

### *6.1 Structural Limits*

Synthetic research has boundaries that are not temporary limitations of current models, but structural features of the approach itself. Acknowledging these boundaries is a requirement of the epistemological framework that I have proposed. Five limits deserve particular attention.

#### *The Language Boundary*

LLMs generate language. They cannot simulate embodied cognition, emotional volatility, behavioural inconsistency or the texture of lived reality. When research questions require access to these dimensions, synthetic methods are insufficient. Ethnographic depth, emotional response measurement and modelling of rare or extreme behaviours fall outside the valid range of simulation.

### *The Coverage Boundary*

LLMs are trained on internet text, which systematically underrepresents certain populations: older adults, lower-income groups, non-English speakers and marginalised communities. Synthetic samples can only be as diverse as the populations represented in the training data (Dillion et al., 2023). Indeed, Santurkar et al. (2023) confirmed this empirically, showing that some demographic groups, notably people aged 65 and older, are poorly represented across models regardless of steering attempts. This means that synthetic research inherits and may amplify the digital divide.

### *The Temporality Boundary*

Off-the-shelf LLMs reflect the world as it was at the time of training. Attitudes shift, language evolves, political alignments change; a model trained before a major cultural or political event may produce outputs that reflect outdated realities. Sanders et al. (2023) demonstrated this in the context of the war in Ukraine, where the traditional association between left-leaning ideology and dovish attitudes reversed, but LLMs failed to capture this shift because their training data preceded the invasion.

### *The Context Boundary*

LLMs lack access to proprietary organisational knowledge. Internal culture, specific processes, unwritten norms and institutional memory are absent from the training corpus. When research questions depend on this context, as many B2B and employee experience studies do, synthetic outputs default to generic best-practice logic.

### *The Variance Boundary*

LLMs tend to produce outputs with lower variance than real human populations. This is partly architectural, driven by temperature settings and next-token prediction, and partly a consequence of alignment processes that compress distributional extremes (Bisbee et al., 2024). The practical effect is underestimation of real-world heterogeneity. Subgroup differences are flattened. Minority positions are suppressed. Confidence intervals are artificially narrow. Power analyses based on synthetic data would severely understate required sample sizes.

These five limits define the territory where synthetic research should not operate as a standalone method. However, they're not reasons to reject synthetic methodology. Instead, they are reasons to use it with precision about where it applies and where it does not.

### *6.2 Synthetic Research as Triangulation*

If synthetic research has structural limits, the question becomes: Where does it fit? My answer is "triangulation". Synthetic methods are most valuable when used alongside traditional quantitative and qualitative research. Each method has strengths and blind spots. The value comes from convergence and divergence across approaches.

Synthetic research is particularly well-suited to five use cases. First, pre-field hypothesis testing: running directional tests before committing to expensive fieldwork. Second, early-stage concept screening: ranking and filtering concepts quickly at near-zero marginal cost. Third, scenario exploration: simulating "what if" conditions that would be impractical to test with real respondents. Fourth, stimulus optimisation: iterating on stimuli before exposing real participants. Fifth, risk reduction: filtering weak ideas before investing in large-scale studies.

In each of these cases, synthetic research extends insight and accelerates iteration. It does not validate itself and it doesn't replace the need for human data. Rather, it reshapes when and how human data is collected, focusing expensive fieldwork on the questions and concepts that survived initial screening.

The history of data collection in our industry is a history of successive additions. Paper-and-pencil interviewing gave way to telephone interviewing, which was supplemented by computer-assisted methods, which were joined by online panels, which were extended by mobile-first and passive-data approaches. Each new method was initially met with scepticism. Each eventually found its place within a larger ecosystem. LLM-assisted surveys and synthetic respondents represent the next layer in this evolution. The paradigm is thus shifting from a linear process of fieldwork, data and analysis, toward a more iterative model of simulation, calibration and targeted validation.

### *6.3 Implications for the Insights Industry*

If synthetic research is to earn its place in this ecosystem, the insights industry must develop new competencies, standards and forms of accountability.

#### *Methodological Ownership*

Insight teams must become architects of epistemic design, rather than consumers of model outputs. The era of prompting a model and treating whatever comes back as data must give way to disciplined pipeline construction. Persona architecture and weighting logic will become core competencies, comparable to sampling design and questionnaire construction in traditional quantitative research.

#### *Bias Auditing*

Identifying and stress-testing model priors should become standard practice in every synthetic research workflow. Just as traditional research requires auditing data quality through consistency checks, outlier detection and response-pattern analysis, synthetic research must audit bias structure through systematic comparison against benchmarks and sensitivity testing across model configurations.

#### *Transparency and Documentation*

Every synthetic study should document the full pipeline: persona specifications, prompt architecture, quantification method, weighting logic, validation benchmarks and known limitations. Without this documentation, synthetic outputs are black boxes. With it, they become open to scrutiny, replication and improvement.

#### *New Professional Roles*

The complexity of synthetic methodology calls for a new type of expertise. Designing effective synthetic research requires knowledge of survey methodology, natural language processing, prompt engineering, psychometrics and domain-specific content expertise. This combination does not map neatly onto existing job descriptions. I therefore expect that roles focused specifically on simulation design will emerge as the practice matures.

### *Industry-Academia Collaboration*

Formalising standards, validation benchmarks and shared best practices will require collaboration between industry practitioners and academic methodologists. The academic literature reviewed in this paper provides essential theoretical foundations and empirical evidence. The industry brings real-world use cases, proprietary benchmarks and practical constraints. Neither side can build this discipline alone.

### *The Stakes Beyond Research*

Finally, the implications of understanding LLM behaviour extend beyond research methodology. LLMs are becoming decision-makers. They recommend, filter and act. They control visibility and perception. They encode and amplify societal bias at scale. Understanding what LLMs produce, and why, is a strategic imperative for any organisation that operates in a world increasingly shaped by these models.

## 7. Conclusion

In the 1930s, the challenge facing opinion research was not a lack of data. The *Literary Digest* had plenty of data. The real challenge was a lack of discipline. There were no formalised rules for selecting respondents, no standards for measurement, no shared framework for assessing when a finding could be trusted. George Gallup's contribution was to supply that discipline. He did not invent a new technology. He built a methodology that made opinion research accountable to transparent, testable principles.

Today, the research industry does not lack synthetic data. Models are generating it at scale, across use cases, with increasing sophistication. What the industry lacks is a formalised set of principles for determining when that data means something and when it does not.

This paper proposed synthetic epistemology as an answer to that gap. The framework rests on several connected arguments.

First, LLMs do not simulate people. They simulate distributions of language. The research value of their outputs comes from the structure that researchers build around them, not from the models themselves. Persona design, prompt architecture and quantification logic are the load-bearing elements of synthetic methodology. Each one embeds assumptions that must be made explicit.

Second, LLM bias is not random error, but a structured prior shaped by training data composition, alignment processes and dominant discourse. I identified six recurring bias types: novelty bias, virtue bias, tech optimism bias, dominant discourse bias, negation handling deficits and machine desirability bias. These are features that must be modelled and calibrated, just as social desirability and acquiescence bias are accounted for in traditional methodology.

Third, empirical validation reveals a consistent pattern. Synthetic simulations perform well on items that are semantically clear, evaluatively concrete and independent of private context. However, they struggle on items that are socially sensitive, emotionally loaded or dependent on organisational knowledge. Critically, targeted improvements in persona design can substantially reduce bias. The quality of synthetic insight is therefore determined by the quality of its epistemic architecture.

Fourth, synthetic research should function as a triangulation layer that extends insight, accelerates iteration and reshapes when and how human data is collected. Its place in the research ecosystem is alongside established methods.

The ancient Greeks distinguished between *doxa* and *episteme* – between opinion that happens to be correct and knowledge that's grounded in understanding. The challenge for our industry is to move synthetic research from the first category to the second. This requires transparency about assumptions, rigour in measurement, honesty about limits and a commitment to validation that goes beyond surface-level similarity.

In short, the real risk facing the industry is not that language models are imperfect; every research method is imperfect. The real risk is that we use these models without clearly stating the assumptions behind them. The opportunity now is to build a transparent, rigorous and accountable synthetic methodology. One that acknowledges bias, validates against benchmarks and earns trust through design.

The shift from synthetic data to synthetic epistemology is, above all, a disciplinary commitment. And it is the condition under which simulated research will earn its place in our pursuit of understanding.

## Methodological Note

The empirical component of this paper is based on three simulation settings. In each case, synthetic respondents were generated as combinations of demographic/firmographic and contextual attributes, constructed through a Cartesian product of variables. The resulting space of possible personas was then weighted using predefined distributions to approximate target population structures.

For the aforementioned Life Trends report, Accenture's executive survey study personas reflect survey firmographic structure. The persona space therefore includes the following categories: geographic location (such as United States, Europe, Asia), organisational headquarters, industry sector (such as consumer goods, energy, retail, technology), market model (B2C, B2B2C, B2B), seniority level (primarily C-level and direct reports), functional role (such as CEO, CTO, CMO, CSO), job role orientation (growth vs. back office) and company size (measured by revenue bands). This allowed the construction of synthetic samples that reflected both consumer and decision-making populations.

The prompt structure is adjusted across simulations and explicitly conditions the model on persona attributes. Consider the following example.

*You are role-playing as a consumer from {p_row['Country']} with these attributes:*
*Age group: {p_row['age']}*
*Gender: {p_row['gender']}*
*Area of living: {p_row['area_living']}*
*Gross household income: {p_row['profile_gross_household']}*
*Read the claim below and respond in 1-3 sentences.*
*Claim: {c_row['claim_en']}*
*Question: How likely are you to agree or disagree? Answer naturally without numbers and briefly explain why.*

Responses were generated using Gemini 2.0 Flash and subsequently transformed into structured outputs via the semantic similarity rating. Instead of direct Likert elicitation, free-text responses were embedded (BAAI/bge-small-en-v1.5) and mapped onto predefined anchor statements to recover distributional structure.

Across all three cases, validation was conducted against benchmarks used in Accenture studies.

*Figure 2. Synthetic Survey Research Pipeline Architecture.*

## References

Argyle, L.P. et al. (2023). Out of One, Many: Using Language Models to Simulate Human Samples. *Political Analysis*, *31*(3), pp. 337–351. [cambridge.org]

Bail, C.A. (2024). Can Generative AI improve social science? *Proceedings of the National Academy of Sciences*, *121*(21), e2314021121. https://doi.org/10.1073/pnas.2314021121

Bender, E.M. et al. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? *Proceedings of FAccT '21*, pp. 610–623. https://doi.org/10.1145/3442188.3445922

Bisbee, J. et al. (2024). Synthetic Replacements for Human Survey Data? The Perils of Large Language Models. *Political Analysis*, *32*(4), pp. 401–416. [cambridge.org]

Dillion, D. et al. (2023). Can AI language models replace human participants? *Trends in Cognitive Sciences*, *27*(7), pp. 597–600. https://doi.org/10.1016/j.tics.2023.04.008

Groves, R.M. & Lyberg, L. (2010). Total Survey Error: Past, Present, and Future. *Public Opinion Quarterly*, *74*(5), pp. 849–879. https://doi.org/10.1093/poq/nfq065

Krosnick, J.A. (1999). Survey research. *Annual Review of Psychology*, *50*(1), pp. 537–567. https://doi.org/10.1146/annurev.psych.50.1.537

Maier, B.F. et al. (2025). LLMs Reproduce Human Purchase Intent via Semantic Similarity Elicitation of Likert Ratings. https://arxiv.org/abs/2510.08338

King, G. et al. (2004). Enhancing the validity and cross-cultural comparability of measurement in survey research. *American Political Science Review*, *98*(1), pp. 191–207. https://www.researchgate.net/publication/229026585_Enhancing_the_Validity_and_Cross-Cultural_Comparability_of_Measurement

Plato. *Meno*, 97a-98a. (c. 385 BC). In *Plato in Twelve Volumes*, vol. 3, translated by W.R.M. Lamb (1967). Harvard University Press; William Heinemann Ltd. http://www.perseus.tufts.edu/ hopper/text?doc=plat.+meno+97a

Sanders, N.E., Ulinich, A. and Schneier, B. (2023). Demonstrations of the Potential of AI-based Political Issue Polling. *Harvard Data Science Review*, 5(4). https://hdsr.mitpress.mit.edu/pub/dm2hrtx0

Santurkar, S. et al. (2023). Whose Opinions Do Language Models Reflect? *Proceedings of ICML 2023*, PMLR 202, pp. 29971–30004. https://proceedings.mlr.press/v202/santurkar23a.html

Tjuatja, L. et al. (2024). Do LLMs Exhibit Human-like Response Biases? A Case Study in Survey Design. *Transactions of the Association for Computational Linguistics*, *12*, pp. 1011–1026. https://doi.org/10.1162/tacl_a_00685

Von der Heyde, L. (2025). Who Counts? The Potentials and Pitfalls of Using LLMs in Survey Research. *COLM 2025*. https://openreview.net/pdf?id=ww2KqnPLdK

Ziems, C. et al. (2024). Can Large Language Models Transform Computational Social Science? *Computational Linguistics*, *50*(1), pp. 237–291. https://doi.org/10.1162/coli_a_00502

## Other Sources

*Reviews and frameworks*

Grossmann, I. et al. (2023). AI and the transformation of social science research. *Science*, *380*(6650), pp. 1108–1109. https://doi.org/10.1126/science.adi1778

*Measurement and methodology*

Sarstedt, M. et al. (2024). Using large language models to generate silicon samples in consumer and marketing research. *Psychology & Marketing*, *41*(6), pp. 1254–1270. https://doi.org/10.1002/mar.21982

*Industry applications*

Brand, J., Israeli, A. & Ngwe, D. (2024). Using LLMs for market research. *Harvard Business School Working Paper* 23–62. https://www.hbs.edu/ris/Publication%20Files/23-062_b8e8e555-1bb6-4245-ac91-0bad1fee1e52.pdf

*Coverage and digital divide*

Hargittai, E. (2015). Is Bigger Always Better? Potential Biases of Big Data Derived from Social Network Sites. *The ANNALS of the American Academy of Political and Social Science*, 659(1), pp. 63–76. https://doi.org/10.1177/0002716215570866

*Cultural bias and representation*

Hartmann, J., Schwenzow, J. & Witte, M. (2023). The political ideology of conversational AI. https://arxiv.org/abs/2301.01768

*LLM bias and alignment*

Gallegos, I.O. et al. (2024). Bias and Fairness in Large Language Models: A Survey. *Computational Linguistics*, *50*(3), pp. 1097–1179. https://doi.org/10.1162/coli_a_00524

Hovy, D. & Prabhumoye, S. (2021). Five sources of bias in natural language processing. *Language and Linguistics Compass*, *15*(8). https://doi.org/10.1111/lnc3.12432

McCoy, R.T. et al. (2023). Embers of Autoregression: Understanding Large Language Models Through the Problem They are Trained to Solve. https://arxiv.org/abs/2309.13638

*LLMs as survey respondents*

Horton, J.J. (2023). Large Language Models as Simulated Economic Agents: What Can We Learn from Homo Silicus? *NBER Working Paper 31122*. http://www.nber.org/papers/w31122

Simmons, G. & Hare, C. (2023). Large Language Models as Subpopulation Representative Models: A Review. https://arxiv.org/abs/2310.17888

*Prompt sensitivity*

Liu, N.F. et al. (2024). Lost in the Middle: How Language Models Use Long Contexts. *Transactions of the Association for Computational Linguistics*, 12, pp. 157–173. https://doi.org/10.1162/tacl_a_00638

Pezeshkpour, P. & Hruschka, E. (2024). Large Language Models Sensitivity to the Order of Options in Multiple-Choice Questions. *Findings of NAACL 2024*, pp. 2006–2017. https://aclanthology.org/2024.findings-naacl.130/

## About the Author

Konrad Suchecki, Research & Innovation Manager, Accenture, Warsaw, Poland.

# Mondelez International Gastronomy Radar with Nextatlas

## Predicting the Next Wave of Food Flavours and Ingredients

*Mario Coletti*

*Serhiy Kalinovsky*

## Esomar

Office address:

Burgemeester Stramanweg 105

1101 AA, Amsterdam

Phone: +31 20 664 21 41

Email: info@Esomar.org

Website: www.Esomar.org

Publication Date: September 2026

ESOMAR Publication Series Volume S424 Congress 2026

ISBN 92-831-0337-8

Congress 2026
Copyright © **ESOMAR** 2026

Copyright © **ESOMAR** 2026


# Mondelez International Gastronomy Radar with Nextatlas

## Predicting the Next Wave of Food Flavours and Ingredients

*Mario Coletti*

*Serhiy Kalinovsky*

### Introduction

Food innovation has always faced a timing problem. Consumer research tools such as surveys, focus groups, and sales tracking are designed to measure what is already happening. By the time a flavour or ingredient trend surfaces in those channels, it leaves little time for innovation planning. The gap between early signal and mainstream visibility is where competitive advantage is won or lost, and most FMCG companies have historically had no reliable way to close it.

Mondelez International set out to change that. In partnership with Nextatlas, the Mondelez cross-functional team designed a radar to test whether systematic monitoring of the gastronomy sector's social media activity could function as a structured early warning system for flavour and ingredient innovation. The core hypothesis: high-end chefs, fine dining restaurants, culinary schools, artisanal food producers and food critics consistently surface new flavours and ingredients, launch them into prominence via their creations and make them present in the Gastronomy scene years before those trends reach mass-market consumers. They are the early trendsetters for major food trends, and their growing digital presence makes this process observable and measurable at scale.

The UK pilot monitored 4,200 Key Opinion Gastronomy Leaders across Instagram, YouTube, TikTok, X and food blogs. Using Nextatlas's proprietary AI, over three years of historical post data were collected, producing a dataset of 1.1 million posts. The results validated the hypothesis and led to formal approval for roll-out to six additional markets in 2026.

### Chapter 1. The Need and the Insight

*A Structural Gap in Flavour and Ingredient Foresight*

Category managers and innovation teams are consistently left working with signals that have already reached mainstream. By the time a flavour or ingredient trend becomes visible in traditional market research methods or retail sales data, the window for early-mover advantage has usually closed. As well as this, once a flavour or ingredient starts gaining traction, planning and executing innovation takes time. By the time a product reaches the market, the company often loses its chance to be first, or among the first, with its innovation.

Internal research conducted alongside this project reinforced the picture. Interviews with multiple stakeholders across functions and regions confirmed that establishing early detection of high potential flavours and ingredients will be

highly beneficial for the company's innovation activities, and that the intelligence arriving from supplier networks typically surfaces too late for development cycles to act on it meaningfully.

The Mondelez team wanted to build such detection by asking a more fundamental question: where do new flavours actually originate before they reach the mainstream consumer, and can that origin point be monitored systematically?

### *Gastronomy as a Test Bed for Food Culture*

High-end gastronomy has always functioned as a living laboratory for food culture. Michelin-starred chefs, culinary schools, food festivals, artisanal bakeries, food critics, and beverage innovators operate at the frontier of ingredient experimentation, constantly testing new combinations of flavours, ingredients, textures with highly engaged, influential audiences within a culinary world. They are not just early adopters, but active trendsetters, who experiment with novel ingredients and flavours before they become mainstream.

What has changed in the last decade is that this world now has a public-facing digital footprint. Through Instagram, YouTube, TikTok, X, and food blogs, the gastronomy sector broadcasts its experiments in real time, to audiences ranging from a few thousand followers to several million. Those audiences then become early adopters who amplify, replicate, and ultimately carry emerging food ideas from the fine dining table to the home kitchen, the cafe counter and eventually the supermarket shelf.

The path of umami, za'atar, and yuzu into mass-market food products illustrates this pattern from outside the formal project scope. Each had been trending in gastronomy social media years before appearing in mainstream branded products, following a trajectory from fine dining accounts through food media, consumer recreation and eventually mass-market food products.

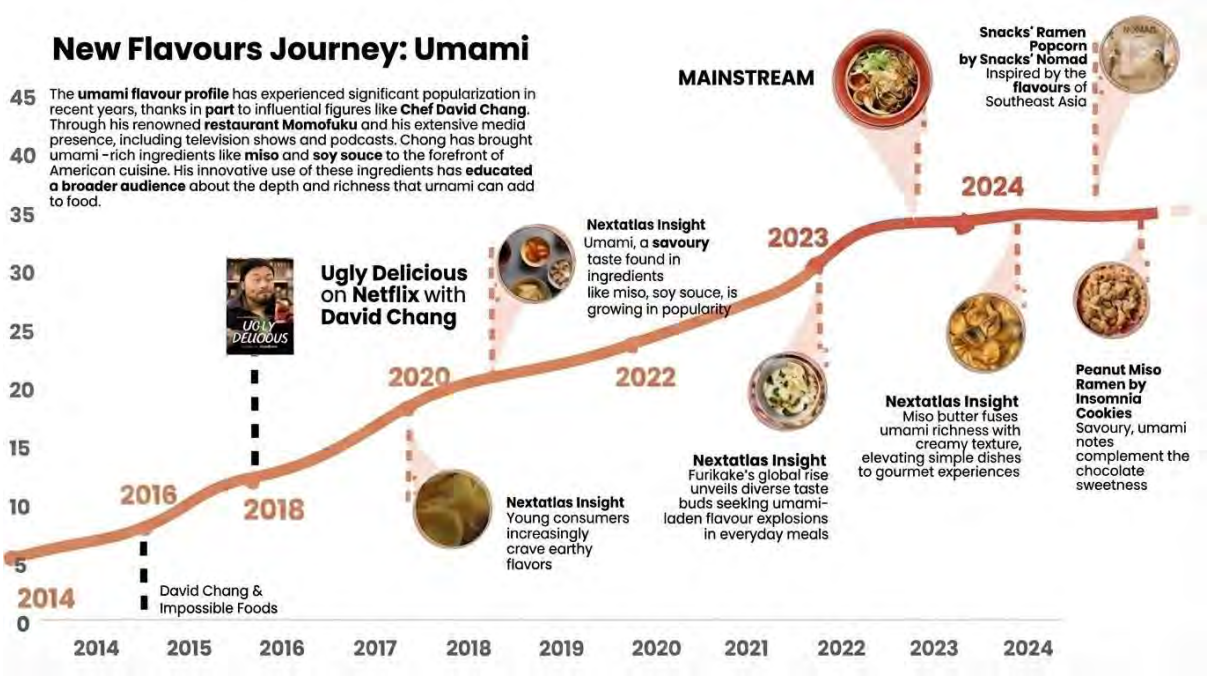



*Figure 1. The umami flavour journey from gastronomy KOL conversations to mainstream product launches.*

## Chapter 2. Methodology

### *From Fashion to Food: The Thinking Behind the Approach*

The idea for Gastronomy Radar came from the world of fashion. Fashion trend forecasting has long operated on the principle that what appears on haute couture runways today will reach the high street in two to three years. The same logic applies to food: what is being cooked and discussed in fine dining today has a strong chance of appearing in the mainstream market within a similar timeframe. Gastronomy Radar was built to make that signal visible and measurable.

### *A Three-Phase Framework*

The project followed a three-phase framework: Plan, Build and Advise. In the planning phase, Nextatlas and Mondelez jointly defined the scope and co-designed a hierarchical taxonomy of flavours, ingredients and taste profiles aligned to Mondelez's portfolio. In the build phase, the Key Opinion Leader (KOL) community was identified and the data harvested and processed analytically. In the advice phase, findings were delivered through structured reports, live workshops and stakeholder presentations.

### *Building the KOL Panel*

The first task was constructing the KOL panel using Nextatlas's AI analysis of early adopter communities. Using keyword-based search queries across social media platforms, Nextatlas identified public profiles of high-end restaurants, culinary professionals, artisanal cafes and bakeries, food critics, famous culinary schools, food festivals, and food influencers with an active UK audience.

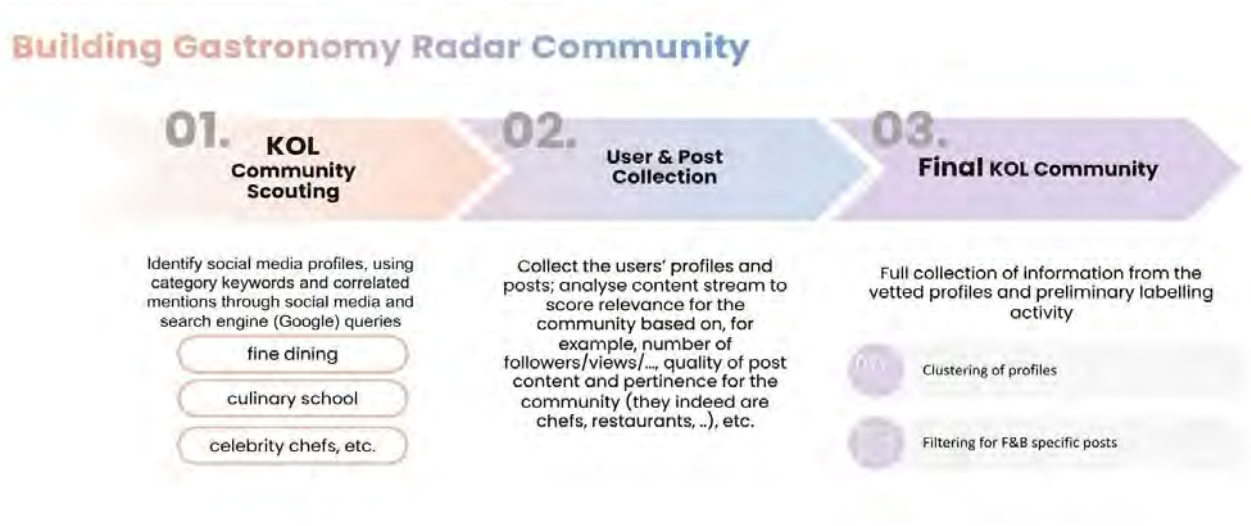



*Figure 2. The three-step KOL community building process: scouting, collection and final community curation.*

Profiles were screened for authenticity, relevance, follower tier, engagement quality and genuine connection to the gastronomy sector. Accounts with a primarily B2B focus, or a focus on ordinary casual dining, food delivery or catering were excluded. The final panel comprised 4,200 UK-based KOL profiles, spanning a follower range from niche emerging voices to large mainstream accounts, and built to be representative of the UK gastronomy world as a whole. The panel prioritised culinary authenticity over raw reach.

### *Data Collection and NLP Analysis*

From these 4,200 profiles, three years of historical post data were collected, producing a dataset of 1.1 million posts. These posts were processed using Natural Language Processing tools that extracted every mention of a food attribute, such as flavours, ingredients, taste profiles, texture descriptors and preparation techniques. Each extracted attribute was linked to the food and beverage category it appeared in context with, whether that was desserts, snacks, beverages, main dishes or appetisers.

The system then quantified each attribute using two core metrics: a weighted growth score, placing heavier emphasis on momentum in the most recent 12 months, and an engagement score reflecting total interaction volume per post, including likes, shares, comments and reposts. These two scores together allowed the team to plot every ingredient and flavour signal on a four-quadrant matrix.

### *Co-Designed Taxonomy*

A critical element of the methodology was the co-development of a shared taxonomy with the Mondelez team. Rather than applying a generic classification, Nextatlas and Mondelez built a hierarchical structure together – a process that required careful work on defining the right words and word combinations in context, so that the system could recognise and correctly categorise flavour mentions across diverse post formats and languages.

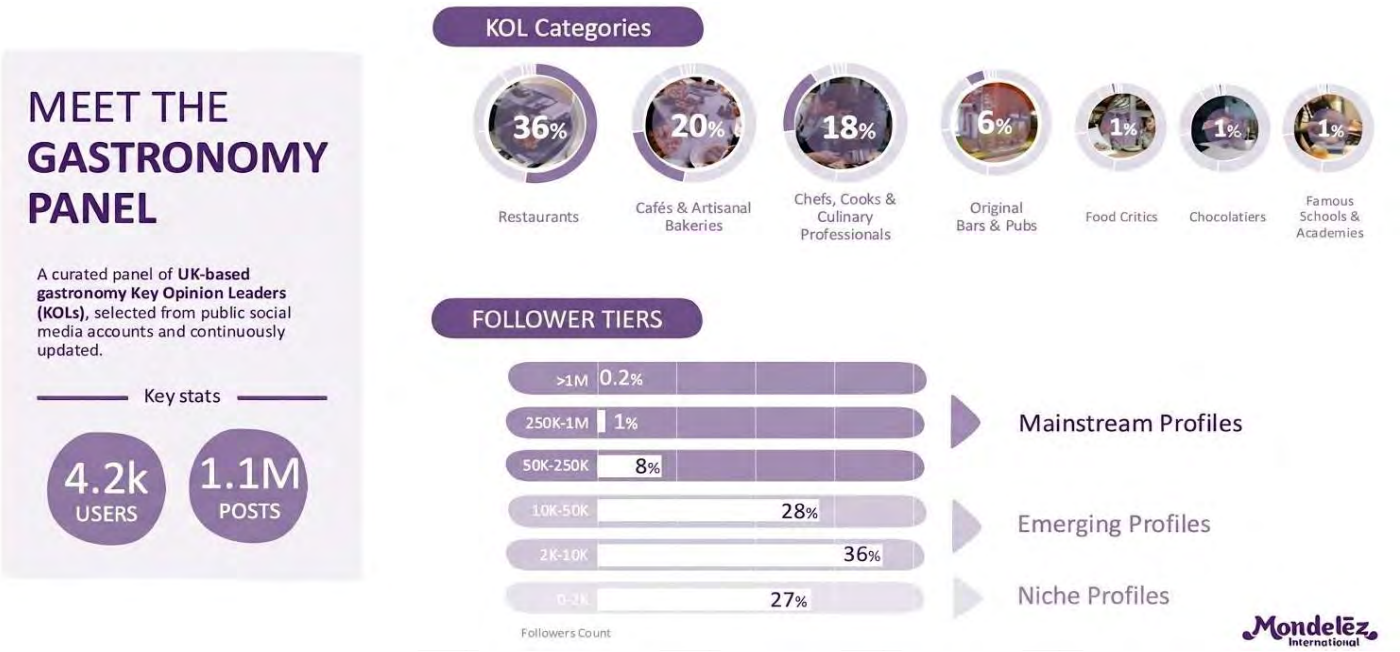



*Figure 3. UK Gastronomy Panel composition.*

## Chapter 3. The Gastronomy Radar in Practice

### *Cluster Analysis and Findings*

The analysis identified 29,000 unique ingredients and flavour signals across the dataset. AI-driven clustering organised these into distinct groups, of which 14 were selected for deep-dive profiling based on their combination of growth trajectory, engagement strength, and relevance to Mondelez's core snacking and confectionery categories. Examples of clusters identified include exotic fruits, pepper oils, edible flowers, and silky textures.

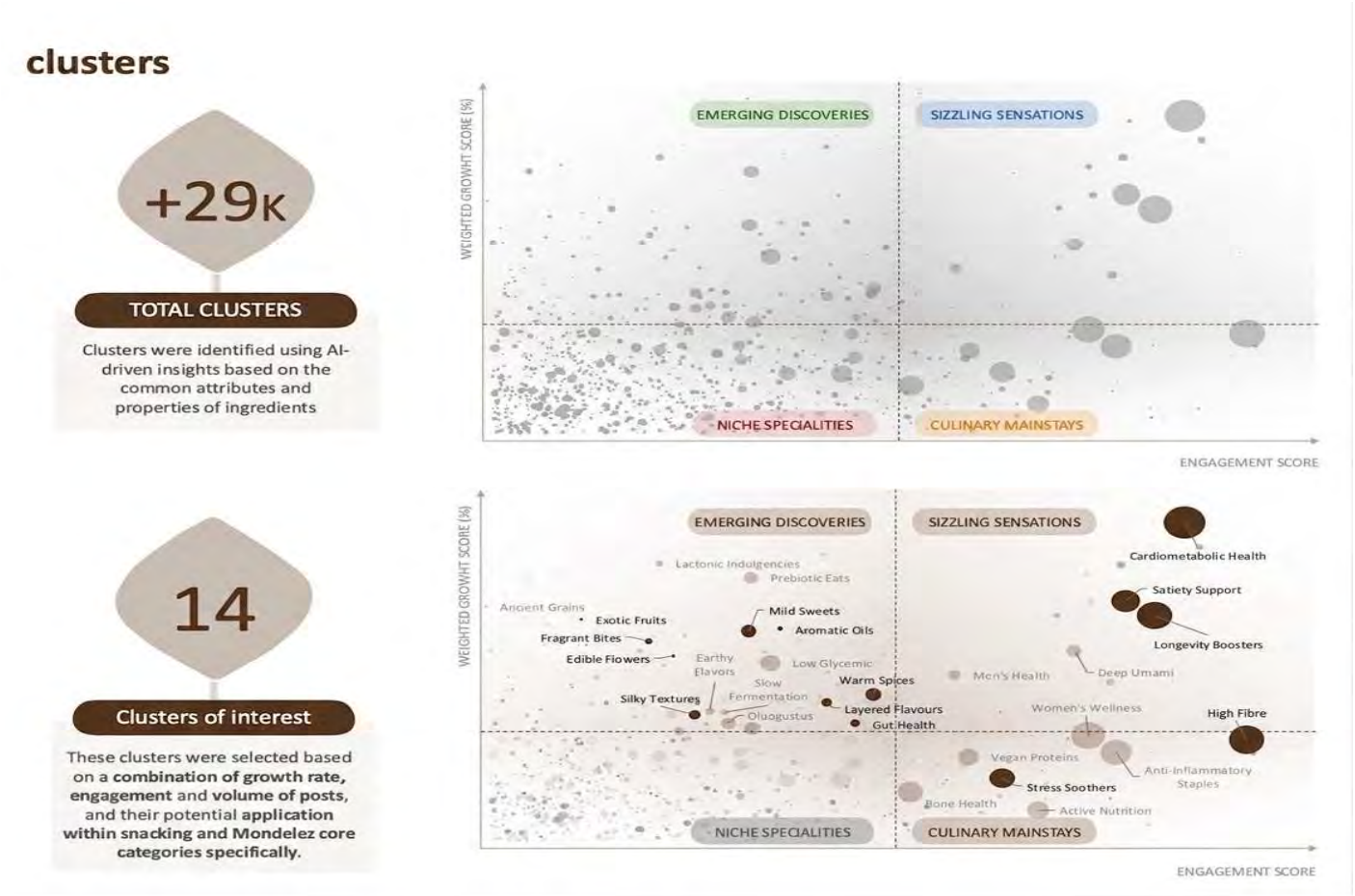



*Figure 4. How the cluster four-quadrant matrix was plotted and the identified clusters of interest thereafter.*

Each cluster was positioned on a four-quadrant matrix plotting growth against engagement, a framework designed to give innovation teams a clear, practical read on where each signal sits in its development cycle:

- Sizzling Sensations (high engagement, high growth): the signals with the most immediate commercial relevance, already building strong momentum among culinary professionals
- Culinary Discoveries (low engagement, high growth): early-stage signals growing rapidly but not yet widely discussed – the candidates most worth watching closely
- Culinary Mainstays (high engagement, low growth): well-established in the gastronomy world, moving toward broader adoption
- Niche Specialties (low engagement, low growth): limited current traction, with selective relevance for specific categories or innovation briefs

This framework gave the innovation team a shared vocabulary for discussing where to focus attention, and why.

## *Illustrative Output*

The following example illustrates the depth of profiling available for each cluster. For a cluster such as exotic fruits, the report provides: the what (a definition and sensory description of the cluster), the why (the macro consumer trends driving interest, such as experiential eating and premiumisation), the who (demographic breakdown of which gastronomy

KOL types are driving the conversation), and the which (ranked ingredient-level data, including individual fruits plotted on the growth-engagement matrix with post volumes represented by bubble size).

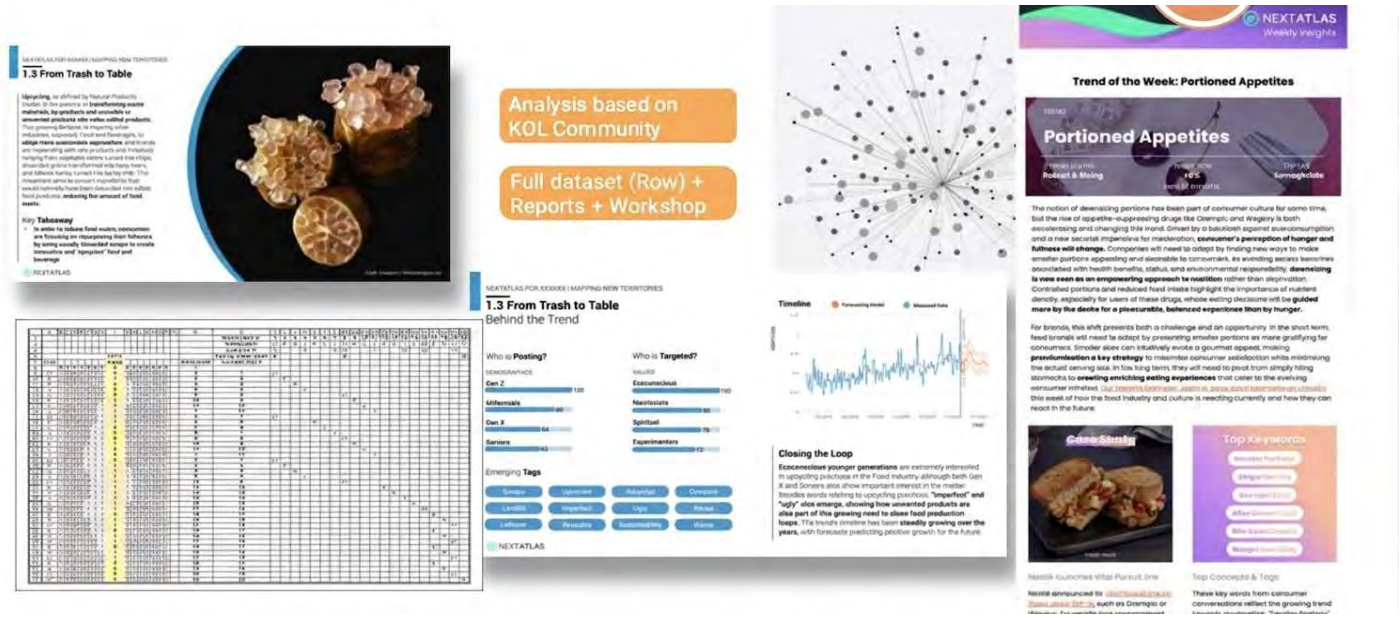



*Figure 5. Illustrative Gastronomy Radar output format showing cluster analysis, trend timeline, newsletter and raw data exports.*

## Chapter 4. Results and Roll-Out

### *From Pilot to Programme*

From the outset, the Gastronomy Radar was designed as a tracker rather than a one-off study. The taxonomy, the KOL selection framework, the quantification model, the report template, and the KPI set were all built to be portable across markets. Following validation of the UK pilot in Q1 of 2026, Mondelez's senior leadership signed off on the methodology and approved roll-out to six additional markets during the remainder of 2026. The roll-out applies the proven methodology, covering taxonomy adaptation to local language, community build, data validation, clustering and profiling, attributes quantification and ranking, data analysis, and report generation.

Final deliverables from the UK pilot included the full Gastronomy Radar report, a structured database of labelled posts, and live stakeholder sessions with Mondelez's Global Insights and Analytics team. For the roll-out phase for the new markets, the Mondelez team additionally will have access to the data through a live dashboard alongside traditional report formats, providing granular access to individual KOL accounts, ingredient-level data, and geographic slicing by market or custom aggregation.

## Chapter 5. Key Learnings and Implications

### *What This Means for the Insights Function*

The Gastronomy Radar shows that AI-powered social listening, applied to the right community of voices, can function as a foresight tool, one with the power to detect high-potential ingredients and flavours before they reach the mainstream.

Copyright © **ESOMAR** 2026


The gastronomy sector operates as a concentrated, expert-curated signal source, producing earlier signals than general social listening.

The Gastronomy Radar shows that a curated community of gastronomy professionals produces earlier and more specific signals than broad consumer listening, as the source is more concentrated. The project also shows that this kind of monitoring can run continuously, at scale, across multiple markets, making it a practical addition to an insights function rather than a one-off research exercise.

### *Key Takeaways*

Gastronomy KOLs are consistent early signals of mainstream flavour trends, with detection possible up to three and a half years ahead of mass-market adoption.

NLP analysis of public social media posts, combined with a co-designed taxonomy and engagement-weighted scoring, surfaces granular ingredient intelligence at scale without the need for primary research.

A four-quadrant growth-engagement framework gives innovation teams a clear, shared way to assess where each signal sits, and what kind of response it warrants.

Building the methodology as a standardised tracking study, with shared taxonomy, standard templates, and agreed KPIs, makes expanding to new markets faster and less costly.

Client-supplier co-design at every stage ensures outputs are directly usable by innovation teams, without requiring significant internal interpretation.

## About the Authors

Mario Coletti, Managing Partner, Nextatlas, Turin, Italy.
Serhiy Kalinovsky, Strategic Foresight Global Lead, Mondelez International, Singapore.

# The Mars Digital Twin Report Card

## Measuring What AI Personas Get Right (and Wrong)

*Adam Bai*

*Johanna Thompson*

*Richard Owen*

Esomar

Office address:

Burgemeester Stramanweg 105

1101 AA, Amsterdam

Phone: +31 20 664 21 41

Email: info@Esomar.org

Website: www.Esomar.org

Publication Date: September 2026

Esomar Publication Series Volume S424 Congress 2026

ISBN 92-831-0337-8

Copyright © **Esomar** 2026

Copyright © **Esomar** 2026

Congress 2026

# The Mars Digital Twin Report Card

## Measuring What AI Personas Get Right (and Wrong)

*Adam Bai*

*Johanna Thompson*

*Richard Owen*

## Introduction

The initial appeal of AI-generated consumer personas is straightforward: faster, cheaper, always-on. Where a traditional qualitative or even quantitative study might take weeks or months and cost tens of thousands of dollars, a well-constructed digital twin can respond to a research prompt in minutes. In categories like fast-moving consumer goods, where competitive windows are narrow and product cycle times are compressing, that efficiency advantage is real and meaningful. In categories like B2B thought leadership, where opinions are especially slow and expensive to gather, digital twins can provide actionable, data-driven feedback when human research is impossible. Built correctly, digital twins can provide other forms of value too: more holistic insights (based on diverse data streams), opinions and feedback from precisely-targeted audience segments, data-driven and more. However, accuracy is the price of the ticket to the value conversation, and the market research industry is only beginning to grapple with this distinction. Most organisations deploying AI personas today lack rigorous, systematic methods for evaluating the quality of what those models produce. They know, at some intuitive level, that outputs can vary—that a digital twin might nail product preference hierarchies while missing the emotional texture that drives brand switching, or that it might perform well on average-consumer questions while struggling to surface the minority signals that often harbour the most valuable innovation insights. Nevertheless, they have yet to systemise these insights into adoption initiatives or internal policies, let alone processes for vendor evaluation.

This paper documents a three-party initiative—Mars, Firefish and Panoplai—designed to address that gap. Working within Mars' chewing gum portfolio, the team built AI-generated digital twins of key consumer segments and then systematically evaluated them using a structured "digital twin report card". The goal was not simply to determine whether digital twins "work", but to understand when they work, why they fail and what governance structures are required to deploy them responsibly at enterprise scale. The paper is structured as follows: section 2 situates the work within the broader context of digital twin adoption in market research; section 3 describes the project objectives and the three partners' respective roles; section 4 details the methodology (how the digital twins were constructed and how validation was designed); section 5 introduces the *Digital Twin Report Card* framework; sections 6 and 7 present the findings (what the models got right and where they fell short); section 8 addresses governance and the role of human judgement; section 9 considers implications for the broader insights community.

## Background: The Validation Gap in AI Persona Adoption

The market research industry's relationship with synthetic audiences has evolved rapidly. A decade ago, the dominant form was the traditional persona: a static, researcher-synthesised archetype built by clustering respondents from surveys, and useful for strategic alignment, although inert and incapable of responding to new questions. The emergence of large language models changed the equation dramatically, making personas interactive and conversational for the first time. However, as Panoplai and twenty44 have documented in their 2026 white paper *The Future of Customer Understanding: A New Framework for Digital Twin & Synthetic Data Validation*, this progression from static archetype to conversational AI persona introduces as many risks as it resolves. Conversational AI is optimised for fluency and helpfulness, not for fidelity to a specific consumer segment. Without architectural controls (what the white paper calls the "Walled Garden" approach), AI models can generate plausible-sounding responses that are grounded in general internet training data, rather than in the brand's actual consumer research. Fluency, in this context, is not fidelity.

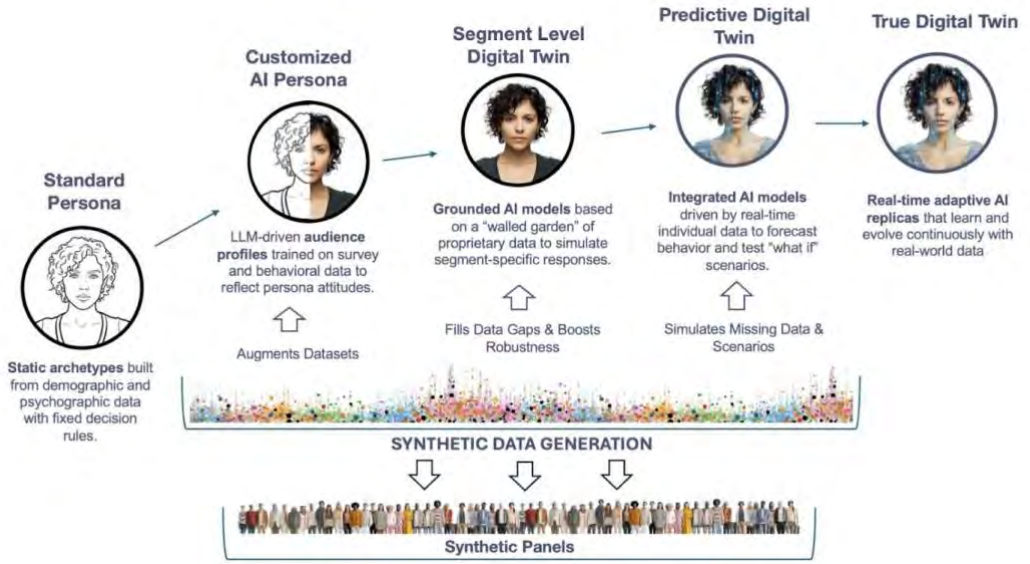



*Figure 1. A visual timeline showing the evolution from Static Personas → Customised AI Personas → Segment-Level Digital Twins, with key capabilities and risks at each stage. Adapted from: Panoplai / twenty44 White Paper, 2026.*

Compounding this challenge is the erosion of traditional "ground truth". The same Panoplai white paper cites Rinzler (2025) research showing that organisations discard up to 38% of panel data due to quality concerns, while Greenbook analysis identifies 30% of online survey respondents as suspicious or fraudulent (Snell, 2024). A 2020 Pew Research study found that 17% of respondents on a high-quality panel provided bogus or inattentive answers—enough to shift survey margins by eight percentage points. If the data on which AI personas are trained is itself compromised, the personas inherit and amplify those compromises. These dynamics create a specific and under-appreciated risk for organisations deploying AI personas, especially within innovation contexts. In an innovation workflow, the margin for well-disguised error is especially narrow: a digital twin that consistently produces plausible but slightly-off flavour preference rankings, for example, may never be caught through normal usage, but could quietly direct product development in the wrong direction. Or such a digital twin could simply reproduce the heart of the consumer preference bell curve, while ignoring important edge cases and emerging trends—i.e., the inspiration for breakthrough innovation or winning marketing campaigns.

Recent academic work begins to quantify the challenge. Columbia University's Twin-2K-500 benchmark estimates current digital twin performance at approximately 87% on a human test-retest reliability standard (Toubia, 2025). This is a meaningful baseline, but it measures descriptive accuracy on well-defined survey items—it does not address edge-case responsiveness, qualitative depth, emotional nuance or the behaviours most relevant to disruptive innovation. Moreover, it looks primarily at the mean or the distribution, rather than at individual-level variation. The field needs evaluation frameworks that go beyond descriptive replication and address the full scope of how AI personas are actually used in practice.

## Project Context and Objectives

Mars is one of the world's largest confectionery companies, with chewing gum brands that compete in a category defined by habitual purchase, sensory cues and emotionally resonant brand associations. The category is also a particularly good test bed for AI persona validation: gum consumers are well-researched, the competitive dynamics are well understood and the innovation pipeline requires both incremental and disruptive thinking. This initiative had three partners with complementary roles. Firefish brought deep qualitative expertise—conducting consumer interviews and immersions that would serve as both the input data and the validation benchmark for the digital twins. Mars contributed category knowledge, innovation priorities and the commercial context that gave the project its decision-relevance. Panoplai

provided the AI modelling environment, constructing segment-level digital twins from the qualitative data and validating them against quantitative benchmarks.

The project set out to answer four questions:

1. Where do the digital twins agree with gathered consumer opinions and observed consumer behaviour, and where do they diverge?
2. Which types of questions and use cases are AI personas most and least reliable for?
3. What governance structures are needed to deploy AI personas responsibly in an innovation workflow?
4. Can we develop a transferable framework—a "Digital Twin Report Card"—that other organisations could apply to evaluate their own AI persona programmes?

## Methodology

### *Qualitative and Quantitative Foundations*

Firefish conducted a programme of in-depth consumer research across key chewing gum segments. This included individual depth interviews and contextual sessions designed to surface behavioural drivers, emotional associations and category tensions. The research was explicitly designed to capture not just stated preferences but the motivational landscape beneath them—the occasions, contexts and social dynamics that shape gum consumption. One important insight: a majority of consumers from across markets, age groups and socio-economic segments chew gum not only for the flavour or social currency, but also because it helps them focus during activities like exams.

The qualitative outputs—transcripts, stimulus responses, projective outputs—served dual purposes. First, they were a primary training input for the Panoplai digital twins: the "walled garden" of verified first-party data from which the AI personas would be built. Second, they provided the qualitative benchmark against which the AI personas' qualitative depth would later be evaluated. This dual-use design was deliberate: it allowed the team to assess whether the digital twins had genuinely internalised the insight landscape or were simply pattern-matching on surface features. In addition, Panoplai ingested one existing Mars survey and fielded a net-new 500n survey, delivered to a non-probability-based sample of UK chewing gum consumer segments. This quantitative data helped to anchor the qualitative data from the historical consumer research in representative consumer segments.

### *Digital Twin Construction*

Panoplai constructed segment-level digital twins using a retrieval augmented generation (RAG) architecture that processes and vectorises client data in a secure environment, preventing the uncontrolled mixing of proprietary research with general-purpose LLM training data. The Firefish qualitative data, along with the historical and new survey data, was ingested, cleaned and structured as the foundational dataset. This was then calibrated against Panoplai's proprietary Panoplai Data Universe (PDU)—a corpus of over 200 million real human question-and-answer pairs—which serves as one "realism guardrail", stabilising persona responses against the variances introduced by LLM architecture.

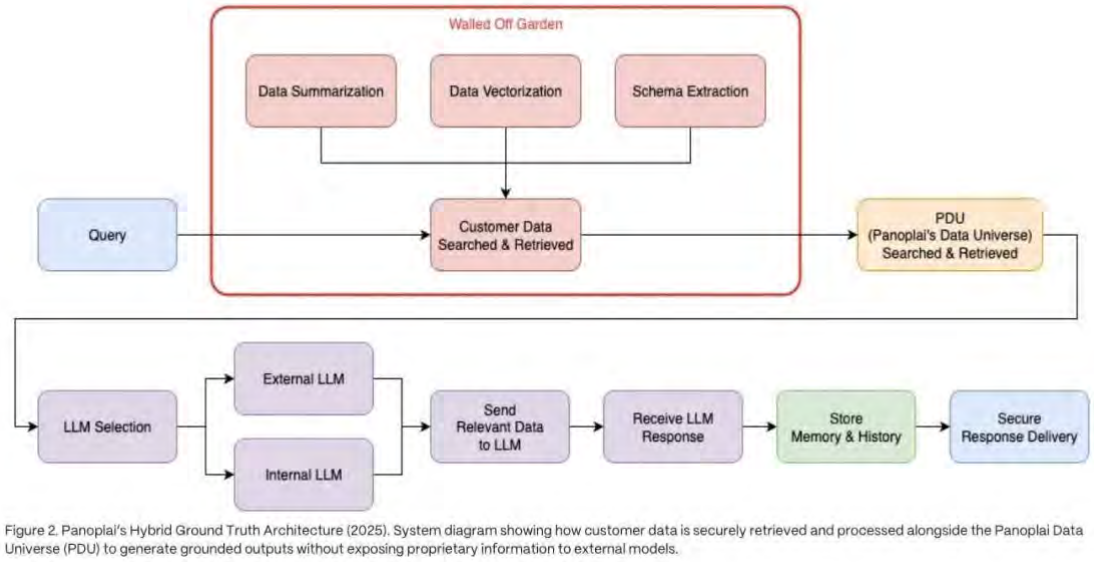



*Figure 2. One of Panoplai's digital twin architectures. Source: Panoplai / twenty44 (2026).*

The digital twins were constructed at segment level, modelling distinct chewing gum consumer profiles identified in the qualitative research. Each twin was designed to respond to queries about product preferences, flavour attitudes, usage occasions and brand perceptions—drawing only on data grounded in the walled garden, not on general internet knowledge. Where the training data contained gaps, synthetic augmentation was used to fill them, with explicit statistical validation demonstrating similarity between real and synthetically generated data distributions. A "team of agents" agentic QA layer was deployed throughout the construction process: specialised AI roles performing real-time auditing of outputs against structural logic, persona consistency and data grounding criteria. In testing, this layer identified and resolved approximately 20% of potential data issues—ranging from survey logic errors to hallucinations—before human researchers reviewed the outputs.

### *Validation Design*

The validation approach was prospective and parallel. Quantitative survey instruments were administered to both real consumers and the digital twins on identical stimuli—including product concepts, flavour propositions and messaging scenarios. This parallel testing design allowed direct statistical comparison of human and AI outputs on the same questions. Qualitative validation was conducted separately. Researchers who had conducted the original Firefish interviews reviewed digital twin responses to equivalent qualitative prompts, assessing the depth, authenticity and nuance of the AI outputs against what they knew from direct consumer contact. This human-in-the-loop evaluation was treated as essential: it captured dimensions of quality that quantitative metrics alone cannot measure.

## The Digital Twin Report Card: A Structured Evaluation Framework

The central contribution of this project is the digital twin report card—a structured framework for evaluating AI personas across the four dimensions most consequential for innovation decision-making. The framework was developed iteratively during the project, informed by the Panoplai / twenty44 white paper's "Four Core Criteria" (ground truth, predictive power, authentic nuance and rules of engagement) and adapted to reflect the specific demands of an active, multi-stakeholder innovation programme (see Table 1).

| Dimension | What it measures | Key evaluation questions |
|---|---|---|

| | | |
|---|---|---|
| **1. Behavioural realism** | How accurately does the digital twin mirror stated opinions and behaviours and observed consumer behaviours, preferences and choice patterns on quantitative measures? | Does rank-order preference align with real consumers? Are usage-occasion patterns reproduced accurately? |
| **2. Qualitative depth** | Does the digital twin capture the motivational texture, emotional context and narrative richness that makes qualitative insight actionable? | Does AI output match researcher memory of real consumer language and reasoning? Does it go beyond stated preference to motivational insight? |
| **3. Edge case responsiveness** | Can the digital twin surface minority signals, outlier perspectives and non-average consumer responses that are often the most innovation-relevant findings? | Does the AI resist the "averaging problem"? Can it reproduce the passionate minority view, not just the modal consumer? |
| **4. Quantitative benchmark alignment** | • How well do AI persona outputs align with parallel quantitative research across key performance metrics—and how transparent?<br>• Is the model transparent about the limits of that alignment? | • Is descriptive accuracy >87% + (the academic benchmark)? Is predictive correlation statistically significant?<br>• Does the model disclose its uncertainty? |

*Table 1. The Digital Twin Report Card: Four Evaluation Dimensions.*

The framework was designed to be transparent rather than merely diagnostic. For each dimension, the team produced a scored assessment with supporting evidence—quantitative where possible, qualitative where not. Critically, the framework also documented where the digital twins lacked sufficient data or confidence to answer a query reliably, reflecting the principle articulated in the Panoplai white paper: that a high-quality AI persona should "respond more like a research respondent than a typical conversational agent chatbot"—including knowing when to say it does not know.

## Findings: Where the Digital Twins Performed Well

### *Quantitative Accuracy and Predictive Performance*

On the "behavioural realism" and "quantitative benchmark alignment" dimensions, the digital twins performed at or above the standard established by independent academic research commissioned by Panoplai (2025; 2026). In parallel quantitative testing on the chewing gum category stimuli, the models achieved 91 to 97% descriptive accuracy against human panel data, exceeding the approximately 87% human test-retest reliability benchmark established by the Columbia University Twin-2K-500 study. Predictive validation, assessing whether the models could correctly forecast rank-order preference hierarchies before the human data was available, produced a Spearman's rho of 0.90 or above across multiple test stimuli. This is a high degree of rank-order agreement: it means that in nine-out-of-10 cases, the order in which the digital twin ranked consumer preferences matched the order observed among real consumers. For routine product testing and concept screening (where the goal is directional prioritisation rather than exact volume prediction), this level of performance is practically meaningful. The models also demonstrated reliable performance across varied question types and stimuli: flavour propositions, positioning concepts and messaging scenarios all fell within acceptable accuracy ranges. This cross-stimulus consistency matters for an innovation programme that needs to use AI personas flexibly across a range of inputs, not just on a fixed question battery.

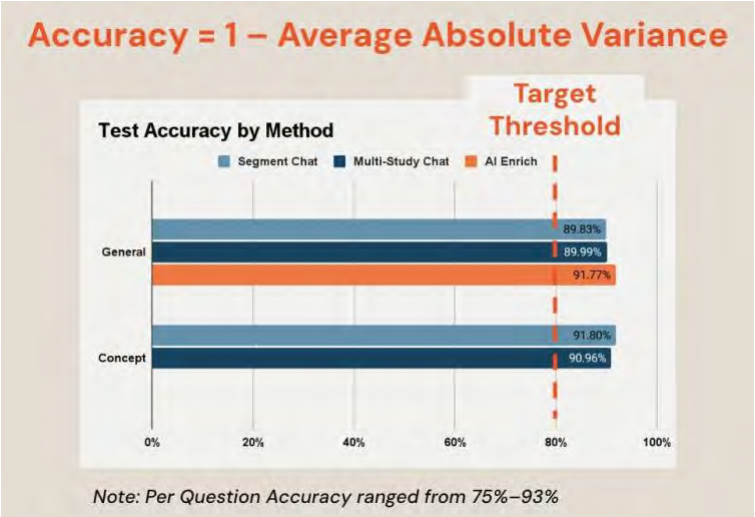



*Figure 3*

## *Innovation Discovery: Edge Case and Opportunity Identification*

*Figure 4*

Perhaps more valuable for Mars' practical innovation needs was the digital twins' performance as exploration tools. When the models were prompted to engage with new product concepts and positioning territories, they surfaced opportunity signals that were consistent with—and in several cases ahead of—what emerged from the human qualitative work. Specifically, the digital twins were effective at identifying the contextual conditions under which different consumer segments would be more or less open to flavour and format innovation, the emotional jobs-to-be-done that a new product proposition would need to address and the competitive substitution patterns that could erode or support adoption. These are the kinds of directional insights that shape innovation strategy—they do not require more than 90% accuracy to be useful, but they do require that the model has genuinely internalised the consumer's motivational landscape rather than generating a plausible average.

*Figure 5*

The efficiencies were also real. Using digital twins for iterative concept refinement—i.e., testing and adjusting propositions in near-real time before committing to human field research—reduced the overall research programme cost by over 30% and shortened time-to-validation by approximately 25% compared to a conventional qualitative-quantitative workflow. This aligns with findings from Harvard Business School research on Human+AI collaboration, which demonstrated that AI significantly increases knowledge worker performance and quality on complex tasks, with speed gains exceeding 25%, human-rated quality gains exceeding 40% and task completion gains exceeding 12% (Dell'Acqua et al., 2023). The key qualifier is "complex tasks": the gains are largest when AI is used to augment, not replace, expert human judgement.

## Findings: Where the Digital Twins Fell Short

### *The Averaging Problem and Emotional Variability*

The most consistent and practically significant limitation was what the project team labelled the "averaging problem". The digital twins showed less emotional variability than real consumers and demonstrated a tendency toward what the Panoplai validation documentation describes as "norm bias"—i.e., a pull toward the modal consumer response at the expense of the passionate outlier view. This limitation is not unique to the Mars programme. As the Panoplai white paper documents, the market research industry broadly recognises the averaging problem as a fundamental challenge for current-generation AI personas: "AI models tend to smooth out emotional responses and overlook passionate, irrational and often outlier opinions that can lead to... human insights". The models are good at the centre of the distribution, but less reliable at the edges—precisely where the most transformative innovation opportunities often live. In practice, this meant that when the team used the digital twins to explore bold flavour territories or unconventional usage occasions, the models tended to under-represent the niche enthusiasm that, in human research, often signals an emerging trend. The digital twin would give a reasonable average reaction; it would not give the electric response of the category enthusiast who becomes an early adopter and opinion leader.

### *Limitations Around Disruptive Innovation Scenarios*

A related limitation emerged when the team tested truly disruptive concepts—product propositions that required consumers to adopt behaviours or attitudes with no prior precedent in the training data. Here, the models tended to interpret new ideas through the lens of familiar category patterns, which could make a genuinely innovative concept appear less viable than it was. This is the "rear-view mirror" problem: a digital twin built on historical data will always have a structural disadvantage when the question is about discontinuous change. To be fair, however, human panels have historically faced the same limitations. People aren't very good at evaluating future-oriented product ideas either! Just like with human research outputs, digital twin outputs should be seen as one signal that needs to be interpreted by human insights professionals–though with the potential help of a powerful analytics platform. The team developed a practical heuristic for managing this tension: digital twins should be used for divergent exploration (generating ideas, stress-testing propositions, identifying opportunity spaces) and

for optimisation (refining concepts that already exist in the data space), but not as a primary signal for convergence on high-stakes, irreversible decisions involving concepts that fundamentally break from category norms. For those decisions, human validation remains indispensable.

### *Sensory and Embodied Experience*

A third limitation specific to the gum category was the models' difficulty with sensory experience. Chewing gum is fundamentally a sensory product: flavour intensity, texture and breath-freshening efficacy are central to the consumer experience in ways that are hard to capture linguistically. The digital twins could articulate consumer attitudes toward sensory attributes ("consumers in this segment prefer a longer-lasting flavour experience"), but they could not simulate the visceral, embodied reaction to a specific product formulation. Though no AI model can actually taste gum, this challenge can be partially solved moving forward through the ingestion of historical concept and flavour tests, teaching models how to generalise appropriate responses relative to specific stimuli. This is a limitation of language-based AI personas that applies across any category with strong sensory or tactile dimensions. The report card flags it explicitly: digital twins are not appropriate as sole validation tools for use cases that require deep sensory or emotional validation of physical product experience.

## Governance and the Role of Human Judgement

### *Embedding Human Oversight in the Workflow*

One of the project's most important outputs is not a finding about digital twin performance—it is a framework for governance. The team found that the value of a digital twin programme is not determined solely by the quality of the AI model; it is determined by the quality of the human judgement embedded around it. Practically, this meant several things. The Firefish research team reviewed digital twin outputs against their direct knowledge of real consumer conversations—providing a qualitative audit that quantitative accuracy scores cannot replicate. The Mars innovation team applied commercial and category knowledge to distinguish signal from noise in what the digital twins produced. The Panoplai team conducted a parallel human study where 250 UK respondents from lookalike chewing gum consumer segments (matched on a range of demographic, behavioural and attitudinal criteria) were shown the same stimuli and asked the same questions as the digital twins. All three teams contributed to defining, in advance, the specific decisions for which AI persona output would be treated as directional only versus decision-grade.

### *Decision Criteria: When to Trust, When to Pause*

Drawing on the project experience and the framework established in the Panoplai white paper, the team developed explicit decision criteria for when digital twin outputs can be relied upon and when human validation is required (see Table 2).

| Appropriate uses | Proceed with caution | Not appropriate (human required) |
|---|---|---|
| • Rapid concept screening and prioritisation<br>• Iterative proposition refinement<br>• Directional segmentation hypothesis testing<br>• Incremental innovation within known category space | • Disruptive or category-discontinuous innovation<br>• Emerging minority signals and trend identification<br>• Emotional or attitudinal questions with high stakes | • Final go/no-go on major investment decisions<br>• Sensory product validation (taste, texture, scent)<br>• Deep emotional or sensitive human experiences<br>• Contexts requiring legally mandated human subjects |

*Table 2. Decision criteria for digital twin use in innovation workflows.*

## *Governance Principles for AI Persona Programmes*

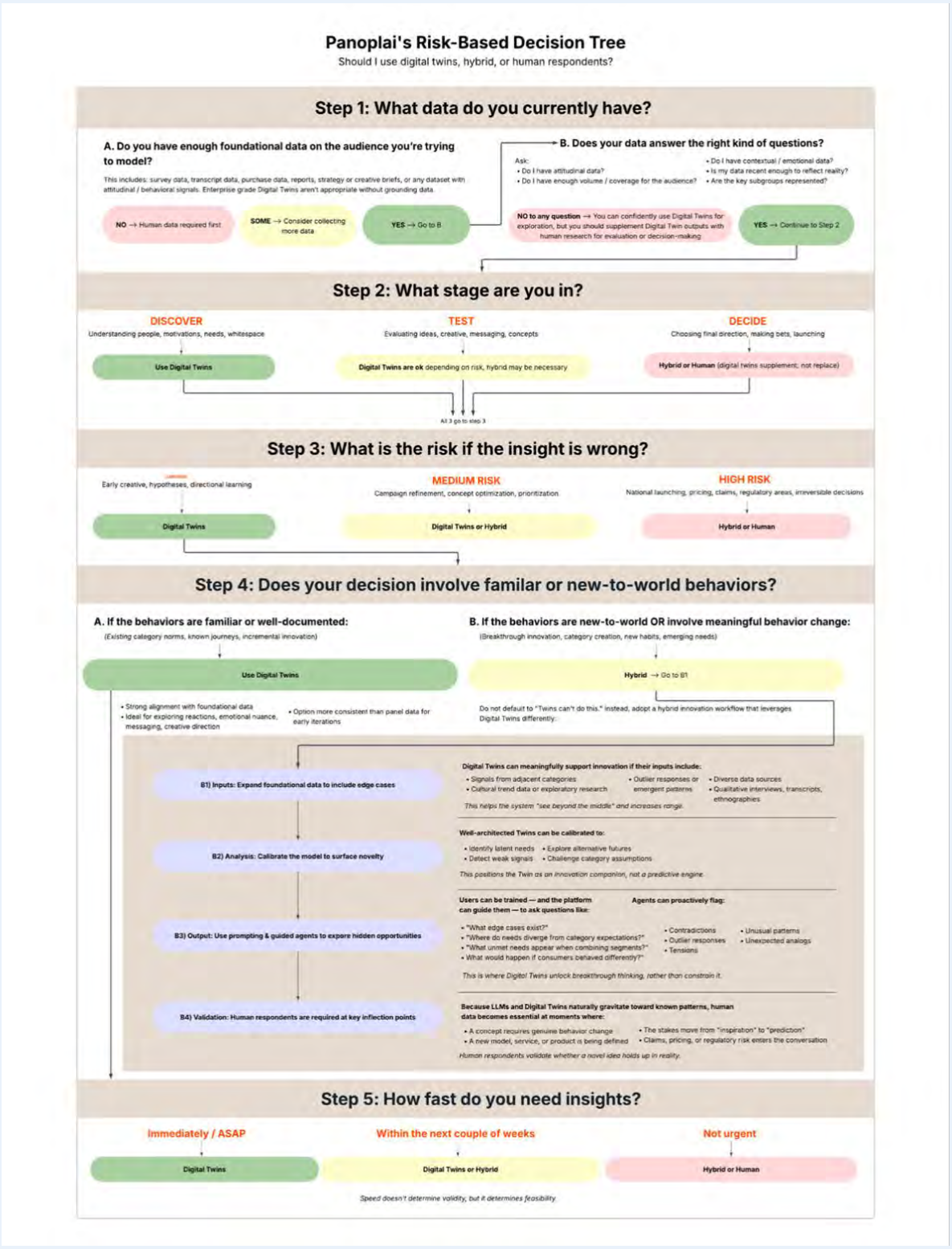
The team's experience supports the following governance principles for organisations implementing AI persona programmes in innovation contexts:

- Establish validation by design, not after the fact. The Firefish qualitative data served as both training input and validation benchmark because this was planned in advance. Organisations that deploy AI personas without a parallel validation design cannot systematically evaluate what they are trusting.
- Disclose limitations explicitly and systematically. The report card framework was useful precisely because it documented where the models were weak, not just where they were strong. Teams that only report positive validation evidence are not managing risk; they are obscuring it.
- Embed qualitative oversight in quantitative workflows. Researchers who know real consumers bring irreplaceable judgement to AI persona evaluation. The team found that the combination of quantitative metrics and qualitative human audit produced better risk management than either alone.
- Define decision-grade thresholds in advance. Before beginning the programme, the Mars, Firefish and Panoplai teams agreed on what levels of accuracy and what types of use cases would qualify AI persona outputs as decision-grade versus directional. This prevented post-hoc rationalisation.
- Treat the report card as a living document. The initial evaluation identified specific areas for model refinement. A governance process for ongoing validation—not a one-time certification—is required to maintain confidence in AI persona outputs over time.

## *The Predictive Behaviour Horizon*

The Mars digital twin project was grounded primarily in stated opinions and behaviors, gathered through traditional surveys, in-depth interviews and focus groups. Digital twins of precisely-defined UK chewing gum consumer segments performed well in parallel testing relative to these outputs. However, the real gold standard for digital twins is predictive power relative to actual consumer behaviours. On behalf of a wide range of clients, Panoplai is now ingesting true behavioural datasets, from marketing performance data to transaction and loyalty card data. The platform enables practitioners to build more holistic digital twins that interweave opinion and behavioural data in order to create more predictive constructs that learn and improve over time, for instance, from one innovation initiative to another. Such an approach offers numerous advantages:

- Chipping away at the "say-do" gap to find correlations in real-time between stated opinions and actual behaviour;
- Uncovering more actionable signals in innovation or ad pretesting data to predict product or campaign performance;
- Grounding digital twins in larger and more diverse datasets to better replicate emergent behaviours and social interactions in response to stimuli;
- Unifying previously siloed data sources into a more systematic, holistic and consumer-centric view of markets;
- Viewing "big data" trends through lens of precisely-defined consumer groups;
- Enabling teams to test future innovation or marketing propositions against richer consumer models and, thereby, garner more actionable feedback in real-time.




A flowchart or decision tree showing the governance workflow: from research question → applicability screening → AI persona query → confidence assessment → either use output directly or escalate to human validation. Adapted from Panoplai's Risk-Based Decision Tree.

*Figure 6*

Copyright © **Esomar** 2026


## Implications for the Insights Community

The *Mars Digital Twin Report Card* initiative was designed as a proof of concept for a specific project, but its implications extend beyond this category and this collaboration. The insights community is in the middle of a rapid and largely unmanaged adoption of AI-generated consumer knowledge—driven by genuine efficiency pressure, genuine capability improvement and sometimes by a desire to avoid difficult conversations about what AI personas cannot yet reliably do. This paper argues for a different posture: one that treats validation as the core discipline of responsible AI persona adoption, not as an optional or after-the-fact quality check. The report card framework is one practical tool for implementing that discipline. It is designed to be transferable—its four dimensions (behavioural realism, qualitative depth, edge case responsiveness and quantitative benchmark alignment) apply to any AI persona programme, regardless of category, methodology or platform. For insights leaders, the project offers three concrete implications. First, the cost of rigorous validation is lower than the cost of silent failure. The efficiency gains from AI personas—30%+ cost reduction and 25% faster cycle times in this programme—are large enough to fund robust validation programmes while still delivering net savings. The argument that validation is too expensive to justify is, in most cases, an argument from default rather than from evidence.

Second, the hybrid model is not a transitional compromise—it is the appropriate end state. The findings here, consistent with broader industry research, suggest that the future of AI-enabled consumer insight is not AI replacing human researchers, but AI augmenting them. The Firefish qualitative team's contribution to this programme was not diminished by the presence of the digital twins; it was made more valuable, because the digital twins handled the high-volume, high-speed work that freed researchers to focus on the interpretive and contextual work that only humans can do. Third, as McKinsey's 2024 *State of AI* report notes, the organisations generating the most value from AI are those using it to create new capabilities rather than to reduce headcount. For the insights industry, this means building AI persona programmes that expand the scope and speed of consumer understanding, rather than simply substituting AI for existing research spend. The report card framework is a tool for ensuring that expansion is grounded in evidence, not assumption.

## Conclusion

The *Mars Digital Twin Report Card* initiative began with a simple question: can we trust what AI personas tell us? The answer, after a systematic evaluation across four dimensions and multiple product development scenarios, is a qualified yes—with important caveats.

- Yes, digital twins can reach and exceed academic benchmarks for descriptive accuracy and predictive correlation. Yes, they can meaningfully accelerate innovation cycles and reduce research costs. Yes, they can surface opportunity signals that are consistent with human consumer insight and commercially valuable.
- Yet, also, no: they cannot yet reliably capture the emotional variability, the passionate minority voice or the discontinuous consumer leap that characterises the most transformative innovation opportunities. They are powerful tools with structural limitations, and those limitations need to be understood and managed—not wished away by enthusiastic adoption narratives.

The digital twin report card is offered as one practical tool for managing those limitations with rigour and transparency. It is not a guarantee that AI personas will deliver accurate insight—it is a structured method for knowing when they do and when they do not. In a domain where plausibility is easy and fidelity is hard, that distinction matters. The broader aspiration of this work is to help shift the industry's relationship with AI personas from adoption on faith to adoption on evidence. That shift will not happen by itself. It requires research teams willing to publish what failed alongside what worked, clients willing to invest in validation as a discipline and technology providers willing to commit to transparency about model limitations. Mars, Firefish and Panoplai offer this programme as a contribution to that effort, and as an invitation for the rest of the community to join it.

### *Key Takeaways for Practitioners*

- The digital twin report card (behavioural realism, qualitative depth, edge case, responsiveness, quantitative benchmark alignment) provides a transferable framework for evaluating AI personas across qualitative and quantitative dimensions;
- In the Mars chewing gum programme, digital twins achieved 91% descriptive accuracy and a Spearman's rho of 0.90+, exceeding academic benchmarks, while also demonstrating measurable limitations on emotional variability and disruptive innovation scenarios;
- AI personas are most reliable for rapid screening, iterative refinement and directional hypothesis testing; they are not appropriate as sole validation tools for high-stakes convergent decisions, deep sensory evaluation or disruptive innovation;
- Governance (embedding human judgement, defining decision thresholds in advance and treating validation as ongoing rather than one-time) is not optional; it is the mechanism that makes AI persona programmes commercially safe;
- The efficiency gains from AI personas (30%+ cost reduction, 25% faster time-to-validation) are large enough to fund robust validation programmes and still deliver net value.

## Acknowledgements

The authors thank the broader Mars, Firefish and Panoplai teams for their contributions to the programme, and acknowledge the independent analysis commissioned by Panoplai and conducted by twenty44 in producing the Panoplai white paper, which provided important empirical and conceptual grounding for the report card framework.

## References

Dell'Acqua, F., McFowland, E., Mollick, E., Lifshitz-Assaf, H., Kellogg, K. C., Rajendran, S., Krayer, L., Candelon F., and Lakhani, K. R. (2023). *Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of AI on Knowledge Worker Productivity and Quality.* Harvard Business School Working Paper 24-013. https://www.hbs.edu/ris/Publication%20Files/24-013_d9b45b68-9e74-42d6-a1c6-c72fb70c7282.pdf

Kennedy, C., Weisel, R., and Jordan, C. (2020). Assessing the Risks to Online Polls from Bogus Respondents. *Pew Research Center.* Available: https://www.pewresearch.org/methods/wp-content/uploads/sites/10/2020/02/PM_02.18.20_dataquality_FULL.REPORT.pdf

McKinsey & Company. (2024). *The state of AI in 2024: Gen AI adoption spikes and starts to generate value.* Available: https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai-2024

Panoplai. (2025). *Evaluating Digital Twins & Synthetic Data with a Global Snack and Confectionery Company.* Available: https://content.panoplai.com/case-studies/evaluating-digital-twins-with-global-candy-company

Panoplai. (2026). *Predictive Validation of Digital Twins Using Rank-Order Agreement.* Available: https://content.panoplai.com/case-studies/predictive-validation-of-digital-twins-using-rank-order-agreemen t

Panoplai / twenty44. (2026). *The Future of Customer Understanding: A New Framework for Digital Twin & Synthetic Data Validation.* Available: www.panoplai.com

Rinzler, K. (2025). *An open letter to the insights industry on data fraud.* Quirk's Media. Available: https://www.quirks.com/articles/an-open-letter-to-the-insights-industry-on-data-fraud

Snell, S. (2024). *The Pervasive Threat of Tech-Enabled Fraud in Survey Research.* Greenbook. Available: https://www.greenbook.org/insights/research-methodologies/the-pervasive-threat-of-tech-enabled-fraud-in -survey-research

Toubia, O., Gui, G. Z., Peng, T., Merlau, D. J., Li, A., and Chen, H. (2025). *Twin-2K-500: A dataset for building digital twins of over 2,000 people based on their answers to over 500 questions.* Columbia University. https://arxiv.org/pdf/2505.17479

## About the Authors

Adam Bai, Chief Strategy Officer and Chief Client Officer, Panoplai, New York, US.

Johanna Thompson, Human Intelligence, Agile Innovation Insights Senior Global Manager, Mars, London, UK.

Richard Owen, Head of Innovation. Firefish and CoFounder, QualifyAI, London, UK.

Copyright © **Esomar** 2026

# Simulating Tomorrow for Coca-Cola's Advanced Hydration Portfolio

## Future twin communities in action

*Thomas Troch*

*Annelies Verhaeghe*

*Adam Palenicek*

Esomar

Office address:

Burgemeester Stramanweg 105

1101 AA, Amsterdam

Phone: +31 20 664 21 41

Email: info@Esomar.org

Website: www.Esomar.org

Publication Date: September 2026

Esomar Publication Series Volume S424 Congress 2026

ISBN 92-831-0337-8

Copyright © **Esomar** 2026

Copyright © **Esomar** 2026

Congress 2026

# Simulating Tomorrow for Coca-Cola's Advanced Hydration Portfolio

## Future twin communities in action

*Thomas Troch*

*Annelies Verhaeghe*

*Adam Palenicek*

### Introduction

Timing is one of the biggest predictors of whether an innovation succeeds, yet most innovation research runs backward: it samples today's consumers, in today's context, to justify bets that will only land years from now. This paper, a collaboration between The Coca-Cola Company and Human8 on the advanced hydration portfolio (sports and rehydration drinks), asks whether AI representations of real, consenting consumers can close that gap. Drawing on a global innovation programme, we built digital twins of real consumers across three markets, and validated them against the people they represent. We then projected them forward into five plausible 2030 worlds, creating future twins and used them to develop innovation platforms and concepts against the futures they will have to win in. The pay-off showed up in the work. Reframing an innovation platform around five futures opened two to three times as many high-quality ideation directions, with roughly four times the share pushing past the obvious. Those are directions, not ideas: the lift showed up before ideation had even begun.

Put to work on a Powerade concept for the LA 2028 Olympics, the future twins went head-to-head with a rewrite built the usual way, from a present-day qual optimisation. On the product detail R&D has to lock in years ahead of launch, the twin rewrite came back more specific: 4.5 against 3.3 on a five-point specificity scale. That is the detail R&D can act on. It is the kind of brief a team can commit to years before the market is there to prove it right. Future twins do not predict 2030; nothing can. What they change is the confidence, and the shared reference points, teams bring to decisions made years before the evidence arrives.

### Timing: Innovation's Hardest Variable

In 2015 Bill Gross, a well-known US entrepreneur, studied more than 200 companies, trying to work out why some succeeded and most failed. He expected the answer to be the quality of the idea, the composition of the team or the amount of funding. Yet, it was timing. Across his sample, timing explained 42% of the difference between the companies that worked and the ones that did not, a bigger factor than the idea, the business model or the team behind it. In his TED Talk, Gross used Uber and Airbnb to make the point. Both had strong execution and genuinely new models, but part of what powered them was when they arrived. Both launched around a recession, when the same conditions worked in two directions: people wanted new ways to earn and new ways to save. Driving your own car for extra money or paying less for a ride than a cab would cost started to make sense at the same time. So did renting out your spare room, or booking someone else's instead of a hotel. The timing served both sides of the marketplace. Gross' analysis is one study, but decades of innovation research point the same way. Launch timing shows up as a distinct predictor of new product performance, separate from how well the launch itself is executed (Calantone & Di Benedetto, 2012; Di Benedetto, 1999). Strategy research sharpens it into a window: entering too early can sink a product as surely as arriving too late (Bayus et al., 1997; Suarez et al., 2015). Moreover, the most recent work makes the point closest to ours: launches succeed when a firm's readiness and the market's willingness to change line up (Robinson & Veresiu, 2024).

Copyright © **Esomar** 2026


Knowing timing matters is the easy part. The harder part is that most innovation processes are built backward. Whether we are defining unmet needs or testing concepts, we sample today's consumers, in today's context and ask them about ideas that quietly assume tomorrow's conditions. The method points forward; the evidence base points back. Timing is easiest to see in startups, where a company can live or die by it inside a couple of years. It matters just as much for established global innovators working on multi-year horizons, where even a strong idea has to land in a market that is ready for it.

## The Brief: Co-Creating the Future of Advanced Hydration

The Coca-Cola Company has strong ambitions for its “advanced hydration” portfolio, home to brands like Powerade and Aquarius. The goal behind those ambitions is clear: to fuel performance in the next era of sports. The brief was to co-create the future of the category. In practice, that meant building a focused global innovation pipeline for their brands, designed to deliver high-impact innovations between 2027 and 2030. Advanced hydration is expanding quickly, pulled forward by new consumers, new occasions and rising expectations of what a hydration product should do. It is also getting more crowded. As competition intensifies, standing out matters more and so does the discipline of backing the big bet innovations most likely to land rather than spreading effort thin across many smaller ones. A global pipeline adds another layer of complexity. Designing strong innovations centrally is not enough on its own. The priority markets that will carry them have to be part of shaping the pipeline and ready to commit to it. The time horizon is where the brief gets hard. R&D, claims and portfolio decisions are committed years before the market is there to validate them. A pipeline aimed at 2027 to 2030 must anticipate where the category will be by then, rather than describe where it sits today.

From the start, one principle shaped how we would answer the brief: people first. It would have been easy to treat the category's future as an expert exercise, with specialists deciding how needs will evolve. We wanted real people at the centre of that forward look, so the pipeline stayed close to how consumers actually think and live. We set out to deliver four things, which together make up the pipeline. First, strong innovation platforms, each holding priority concepts sequenced across 2027 to 2030 so it is clear what launches when. Second, for every concept, a clear view of how to win, plus the working detail R&D needs beyond a concept description. Third, a portfolio architecture: how each concept answers a distinct need or occasion, how it connects to the other Coca-Cola categories and brands and how it fits each market's existing pipeline, which is rarely the same twice. Fourth, alignment and engagement with the regions throughout, so the pipeline was built with the markets, not handed to them.

## The Approach: An Insight-Led, End-To-End Process

The foundation of this project was a rigorous innovation programme: a global, insight-led approach combining human, cultural and creative methods across five phases. Building on it, we explored new territory: experimenting with how to simulate the future. The programme defined three innovation platforms, generated over 500 novel ideas, developed them into a longlist of concepts and landed on a series of big bets to carry forward. The funnel was the point: wide at the top, disciplined at every step down, with meaningful iteration throughout. Underneath the creative work sat a principle we keep returning to, the human spark. It has three parts. Human insight is about what really drives people: what they say, what they don't, the gap between what they tell you and what they actually do, and the motivations sitting underneath. Cultural fluency is a read on the shifts in values and norms that move markets in ways no straight line predicts. Consumer closeness is the empathy that turns understanding into something you can act on. All three give the creative process real understanding to build on. The five phases moved from definition to pipeline. Figure 1 shows them in linear sequence for clarity, though the work itself was iterative throughout: our understanding of the problem and the solution space kept evolving across all five stages.

Copyright © **Esomar** 2026



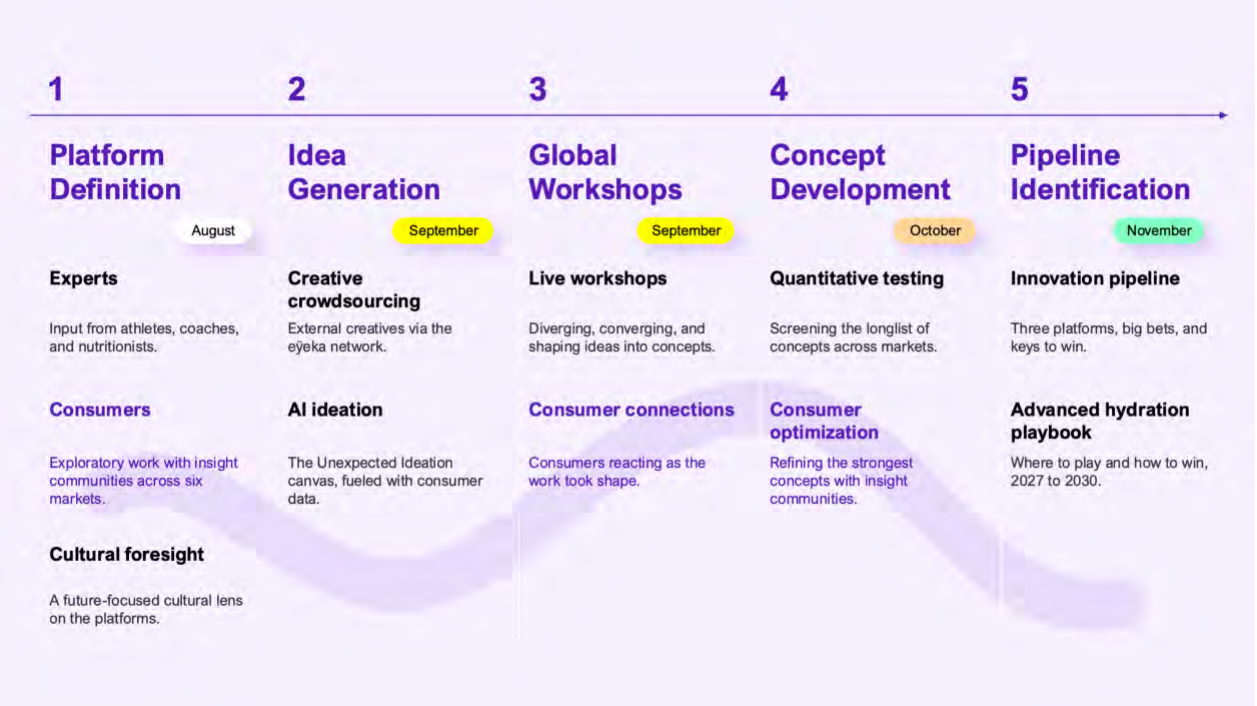



*Figure 1: The five-phase innovation process.*

Innovation platform definition came first. An innovation platform is a defined area of opportunity for growth, broad enough for several concepts to sit within and rooted in consumer insight. We built three of them, starting from a meta-analysis of existing knowledge, including proprietary reports. Onto that base we layered expert input (athletes, coaches and nutritionists), deep immersion and ethnography across insight communities in six markets, and a cultural-foresight lens to keep them future-focused. Those communities included both target consumers and extreme consumers, such as professional athletes whose more advanced needs often signal where the rest of the market is heading. Idea generation and the global workshops came next, the heart of the creative work and where human and AI creativity met. We ran three strands at once. Creative crowdsourcing through eÿeka, Human8's global creative network, brought in ideas from external creatives, a diverse group from around the world. AI ideation ran through our “unexpected ideation canvas”, one of several purpose-built AI assistants we use across our innovation projects. Built around proven divergent-thinking techniques and fuelled with consumer data, it breaks past the sameness that typical AI ideation tends to produce. Live workshops then brought stakeholders together to diverge, converge and shape raw ideas into concepts, with consumers joining in to react as the work took shape. We have measured this combination before. Against a traditional internal brainstorm, it produced five times as many top-ranked ideas (Verhaeghe et al., 2024).

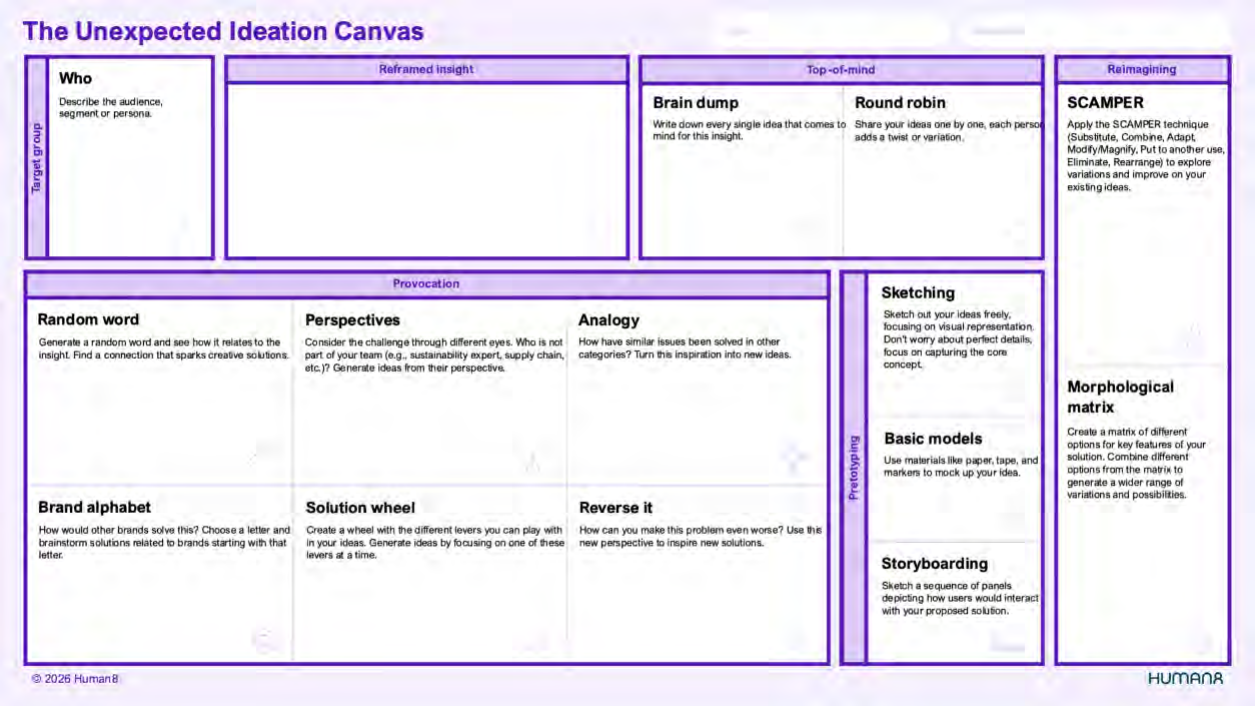



*Figure 2. The "unexpected ideation canvas".*

With a longlist in hand, concept development put each concept through quantitative testing and the best performers were optimised with consumers by reactivating the insight community. Pipeline identification then pulled everything into a single strategy and the advanced hydration playbook, setting out where to play and how to win through 2027 to 2030. Across all five phases, a broad, cross-functional team of 35 stakeholders from global and regional units stayed engaged, so the markets that would carry the pipeline helped shape it from the start. By the end, the challenge had a complete answer: three platforms, a series of big bets, the keys to win for each and an innovation strategy playbook to put them into action. From there, the work moves from strategy into development. Each priority concept is taken into a concept toolkit that adds a further layer of granularity across relevance and positioning, product details, name and visual identity and channel and activation. It then becomes a brief to R&D, and development begins.

What it does not fully resolve is the timing problem this paper opened on. Cultural foresight and expert input bring the future into the work; the consumer voice running through every phase is still today's. The method points forward while the evidence base points back, and this matters here because the pipeline will only reach the market in three to five years. The platforms that guide it have to stay current as that future arrives. The obvious answer is to keep foresighting continuously and stay permanently connected to consumers, yet this is not realistic on the budgets and timelines real programmes run on. It would also mean keeping leading-edge consumers engaged on an ongoing basis, and that profile is hard enough to find once, let alone hold onto. Moreover, there is a deeper limit: people are not, in fact, very good at describing the future. Even experts forecast it poorly and are overconfident when they do (Tetlock, 2005), so neither asking consumers more often nor handing the question to specialists would close the gap. Current methods, in other words, are not a solution to the timing problem. What changed this was AI. Rather than asking how to survey more people more often, we asked a different question: could a virtual audience (an AI representation of real people) stand in for that continuous connection? Also, we would not be starting from scratch. The project had already produced a deep, living body of human data and, alongside it, the foresight work pointing to where the world was heading. Grounded in that data, such an audience could be queried again and again without re-fielding. Projected forward through the foresight, it might speak for tomorrow's consumer as well as today's. That was the hypothesis we set out to test, what we took further by building twin communities and then future twin communities.

## Building the Twin Communities

A twin community is an AI representation of a specific, real community of consumers, built from the first-party data they have already given us. Where the live community disbands when fieldwork ends and has to be reactivated before it can be engaged again, its twin can be returned to and questioned at any time. That is what made the hypothesis concrete. Rather than stay permanently in field, we could stand up a digital version of a community we had already met and keep talking to that version. The raw material was there: the earlier phases had left a rich record of how real people in this category think, feel and decide, far more than the project had drawn on. Instead of letting that record sit in a report, we used it to create the twins, so the exchange could carry on long after the live sessions closed.

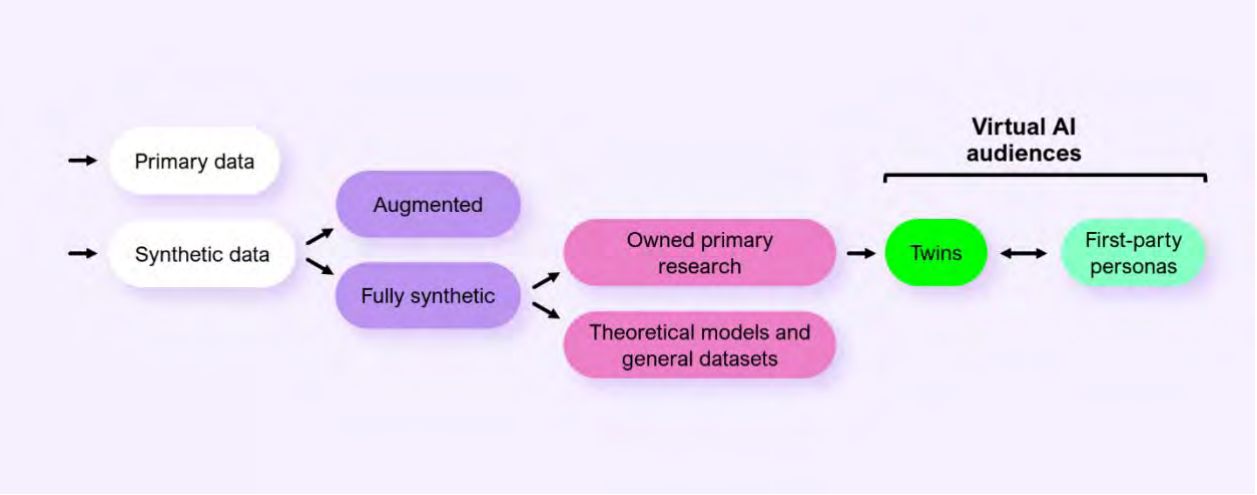



*Figure 3. The synthetic data landscape.*

It helps to place twins within the wider synthetic data landscape (see Figure 3)—i.e., a fast-moving space. Where primary research means interacting directly with people, synthetic data is generated through interaction with a machine that stands in for them. When that representation can simulate how real people would respond, the field calls it a virtual AI audience and a twin community is one kind. Two distinctions matter most. The first is the data behind the representation: some approaches start from a company's own primary research, others from theoretical frameworks or general-purpose datasets and ours began from our own first-party record. The second is granularity: a persona starts from a group, a twin from an individual (see Table 1). Working at the individual level keeps consumer reality flexible, letting us recombine and question it rather than freeze it into fixed segments. That is why we chose twins, each modelled on a real, consenting member of the community.

| | Synthetic persona | Digital twin |
|---|---|---|
| **Starts from** | An aggregate: a segment, summarised from group-level data. | An individual: a real, consenting person, who can be grouped up into segments or audiences. |
| **Built on** | Mostly group-level inputs (segmentation, qualitative summaries, social listening, expert framing); individual data can be added. | Individual first-party data as the required starting point, augmentable with aggregate data, reports, or notes held in the twin's memory. |
| **Flexibility** | Splits down into sub-segments, but cannot be resolved back into real individuals. | Reads as an individual, or aggregates up into any segment or audience. |

*Table 1. Synthetic persona versus digital twin.*

Copyright © **Esomar** 2026


For this project we built 27 twins: three per segment, across three segments and three markets (the United States, Australia and Germany). One segment was professional athletes, a costly, hard-to-reach audience where live engagement is slow—exactly the kind of group twins serve best. A sample of 27 is sized for depth, in line with what qualitative work uses for audiences like these. Additionally, because each twin can be queried repeatedly across every concept and all five worlds, the analysis rested on far more responses than the headcount suggests. Consent sat at the centre of this. Every participant whose data shaped a twin gave explicit permission for this specific use; the representations exist to extend a relationship people had already agreed to, not to stand in for anyone who never signed up.

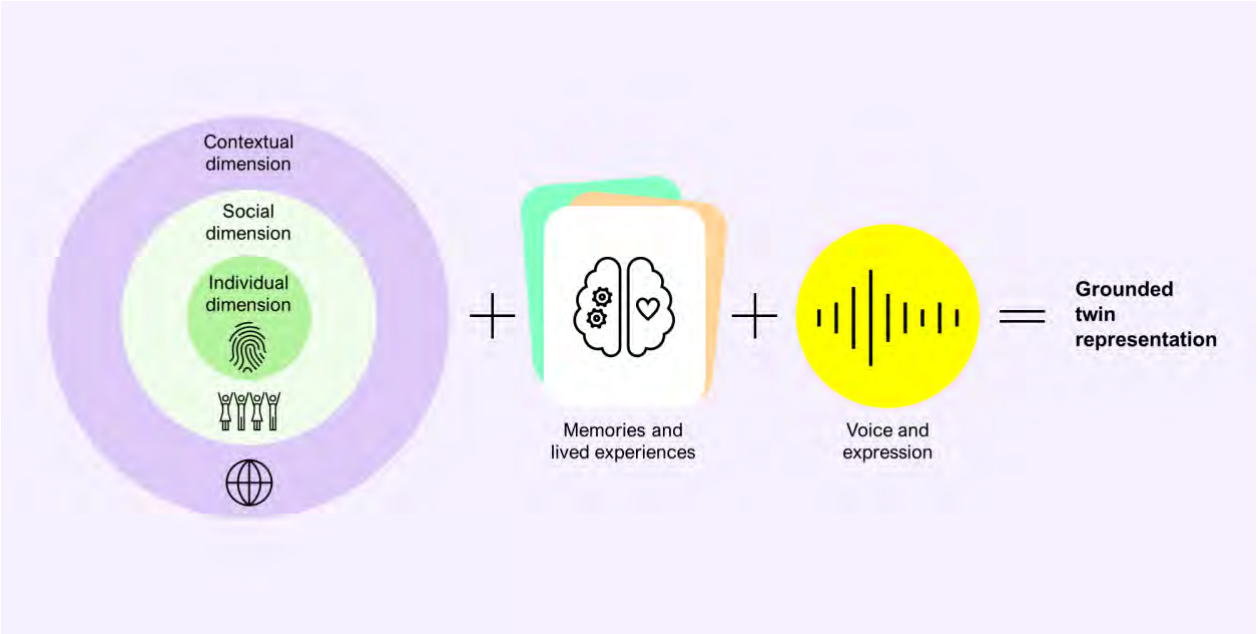



*Figure 4. The identity card.*

What separates a useful twin from a convincing fake is what it is built on: generative AI can produce something plausible in seconds, and plausible is not true. Each twin starts from an identity card (see Figure 4): a rich description of how that person reasons and feels about hydration, drawn directly from the community and extended with their memories and lived experiences. That gives the twin something real to reach for when asked something new. We brief each one on a style guide drawn from how that person actually speaks, so it doesn't default to the model's voice. Wherever genuine historical data exists for a question, the twin draws on that first. We track this as a grounding score: of all the answers a twin gives in a simulation, the share that rests on real evidence from the person behind it rather than the model filling in the blanks. This is where the field is heading: Bain's work on synthetic customers (Pierce et al., 2026) concludes the strongest twins are built on a company's own first-party data, not vendors' third-party data.

## Testing the Twins Against Their Humans

A twin is only worth using if it behaves like the person it represents, so we tested the community's reliability and validity, with a focus on concept development and optimisation. We began with reliability: each twin responded more than once to the same pipeline concepts, and we compared the answers on semantic and lexical similarity to see whether a twin is stable across repeats. We also checked the opposite (that the twins are not simply variations on one voice) by comparing answers across different twins. Next, grounding: across simulations the grounding score averaged 64%, meaning the twins drew on real prior feedback most of the time and extrapolated for the rest only when faced with something genuinely new. Finally, because each twin maps to a real person, we could compare the two directly. These twins are built for qualitative work, so to measure how closely one tracks its human, we quantified the qual. We held each person's original concept feedback out of their twin's creation, had twin and human respond to the same new concepts, then coded both sets of open answers onto five-point scales (appeal, purchase

intention, relevance, uniqueness) and correlated them. The correlations were high, from 0.881 to 0.991, telling us the twins were anchored closely enough in real data to stand in qualitatively.

That fits a growing body of work showing language models grounded in real data can reproduce human survey responses with meaningful fidelity. In the closest parallel to ours, synthetic respondents built on real consumer-goods surveys recovered around 90% of human test-retest reliability on purchase intent (Maier et al., 2025). Two things keep our numbers in proportion. The twins are purpose-built, tested on concepts within the single category they were built for and a high correlation on home turf is what successful twin construction should produce. Moreover, we tested this topic only, so we read the result as evidence the construction works, not as a benchmark to carry across studies. These twins still represented the present, real people as they are today. To explore timing, we had to take them somewhere they had never been: into a set of possible futures, built on yet more data. That is the step from twin communities to future twins.

## Future Twins: Future Self and Five Possible Worlds

Bringing a present twin into the future is more than asking it to imagine ahead. The present twin only knows what its real-life counterpart already knows. To explore timing, we had to add what comes next, in two directions. The first was the person's own life trajectory, what they expect their body, routines and needs to look like a few years from now. The second was the world around them, which changes in ways no individual can predict. We built both layers separately, and then we combined them. The first layer was each twin's own future self. The consenting participants behind the twins told us how they see their lives, bodies and active routines changing one to two years out and again three to five years out. They described what could become easier or harder for them, where they would want more help recovering or staying sharp, as well as what an active-life drink would need to understand about them in the years ahead. That set of answers was added to each twin's identity card as a future ID layer. The twin now carries both who the person is today and who they expect to become as a set of variables. That self-projection has a known limit. People tend to imagine future selves as continuations of present ones, underestimating how much will actually change (Loewenstein et al., 2003). The future-self layer was necessary; it was not enough on its own.

The second layer was the world itself. A person's future life does not unfold in a vacuum. It is shaped by what is happening around them: their work, their culture, the products available to them, the pressures on their time and money. Also, there is no single future to plan against. The market in 2027 will not be the market in 2030, and neither will land exactly where any forecast predicts. Rather than bet on one outlook, we built five plausible future worlds and put the twins into all of them. The point was to see what shifts across them and what holds. We built this layer from evidence rather than imagination. The signal drew on Coca-Cola's own data and category insight alongside new cultural strategy work from the Space Doctors team at Human8, with foresight, wellness research and cultural-shift inputs layered in. Reading across this set, we crafted five distinct worlds, defined by where the future is unsettled rather than by what they all share. What they all share is the baseline: a set of conditions near certain by 2030. Take weight and appetite medication: common today, set to grow sharply and changing how people eat, drink and respond to sweetness. Households using these drugs already buy differently, shifting away from sugar and toward protein and fibre, and are projected to account for a sizeable share of food and beverage purchases by 2030 (Circana, 2025). Every twin meets that kind of certainty in every world; the worlds differ only on what is genuinely uncertain.

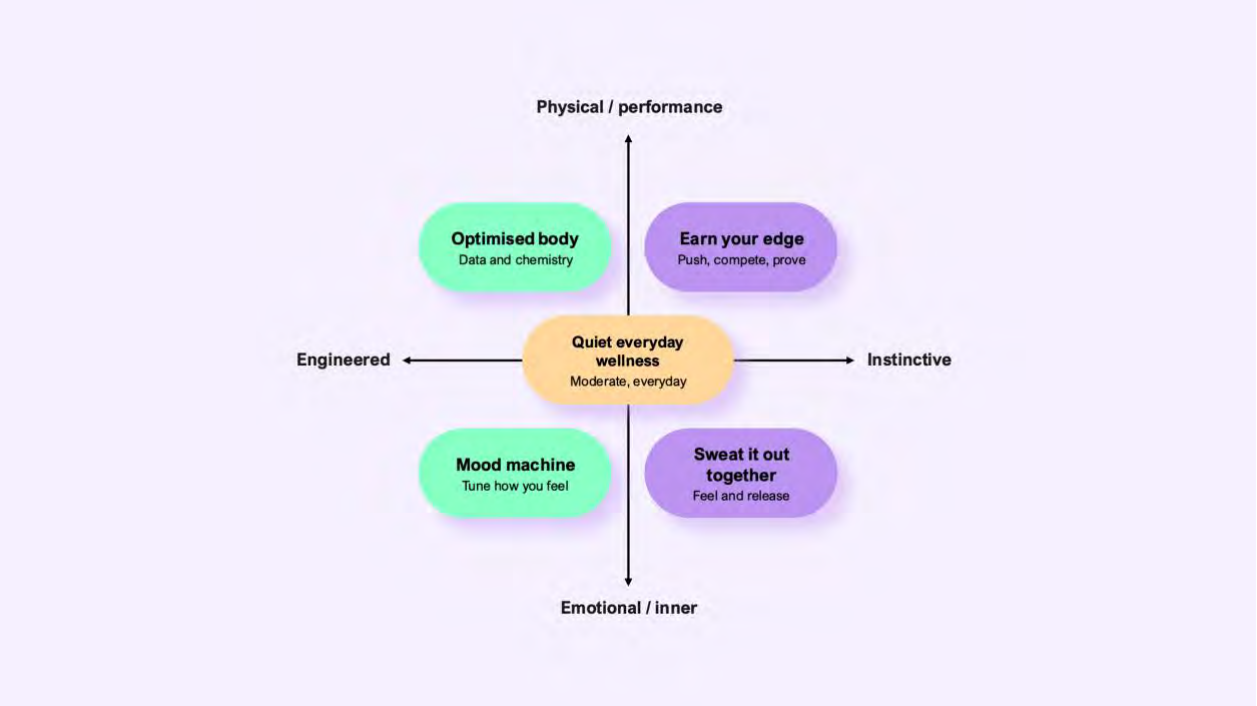



*Figure 5. The future worlds map.*

The five worlds spread across two axes (see Figure 5). The horizontal runs from engineered to instinctive: how the wanted state is reached. The vertical runs from body to feelings: what a drink is really for. Four worlds occupy the corners: optimised body, earn your edge, mood machine and sweat it out together. The fifth, quiet everyday wellness, sits in the middle. Every twin faced the same five worlds, but not in the same words. Each world was framed in that twin's own terms, drawing on the language and lived experiences already held in its identity card and memory. Thus, each twin met the same future, but in a version shaped around the person it was built from. That framing mattered more than it might appear. Used qualitatively, and in exploratory work like the dynamic platforms we describe later, twins put the consumer's voice itself at the centre of the output. The texture of the language is what the team reasons against, so anything that flattens it costs the work. What protected it was the way we presented a world back to a twin. A generic description read like an external briefing. Framing the same description in the twin's own reality shifted the response: the twin spoke from inside the world rather than commenting on it from outside.

With both layers in place, a future twin can be asked the same question in each of the five worlds and we can read where its answers move. This applies as much to building innovation platforms as to developing specific concepts. Either can be run through all five worlds, and either can be judged on which futures it wins in and which break it. Which worlds dominate will vary by segment, by market and by timeframe. Because the worlds were built from signal rather than guesswork, they can be refreshed as the signal changes. New evidence will confirm some, weaken others and may surface ones we did not see yet. The set is meant to evolve, not to be a forecast frozen in place. By the end of this step, each twin carried three things: a present-day grounding made of identity, memories and voice; a future ID layer of who they expect to become; and the five worlds it could be placed in, which share a common 2030 baseline and differ only in the uncertainty each one introduces.

## What We Could Test and What We Found

We began where we did with the present twins by re-running the reliability check: asking each future twin the same question more than once to see whether it answered consistently. It did, and after that the ground shifts. There is no real future to benchmark against, so we could not test whether the twins were right about 2030. What we could test was two things: whether the five worlds actually change how a twin behaves and ring true, and whether the answers stay grounded in real people and add something a raw model would not. A world that does not change behaviour is wallpaper. To test ours, we asked each twin the same set of questions from inside each of the five worlds, about their active lives,

Copyright © **Esomar** 2026


the category and how they felt they had changed over time. Each twin produced five sets of answers, one per world. To see what that looks like, take one twin and ask the same question in two of those worlds. The person doesn't change; what they care about most shifts with the world they're in (see Table 2).

| World | The same twin, answering from inside that world |
|---|---|
| **Optimised body: the body as the thing to engineer** | "The recovery side is what actually moves the needle for me. I can train hard when I'm building toward something, sure, but the strength only really sticks when I rest properly, that's when I come back able to go a bit heavier with my form still holding". |
| **Mood machine: steadiness and mood as the real goal** | "For me this is more of a head thing than a body thing to be honest. If I keep pushing with no real reset I just end up wired and snappy and the whole week feels heavier than it should, so the recharge is what actually keeps me going". |

*Table 2. One twin, one question, two of the five worlds.*

That is one twin, and the shift is easy to see. The harder test is whether an outside reader can pick the world out without being told. Thus, we stripped the world labels from the transcripts and gave the same 10 to a blind panel of 10 insight and foresight experts, each asked to name which world a twin was living in from its words alone: 100 judgments in all. They were right 71% of the time. With five worlds to choose from, guessing lands at 20%. When they did get it wrong, they tended to confuse worlds that sit close together on the map rather than ones far apart, which is what you would expect if the worlds carry a real, readable signal rather than blurring into one another.

In parallel, a panel of experts rated how plausible each world felt, both overall and in the markets they knew well. Across the three markets they averaged 3.83 out of five, with individual worlds running from 3.03 to 4.13. Their comments explained the spread: worlds scored highest where the experts could already see them beginning, and lowest where they believed the social shift but could not yet see the category's role in it. The last check moved from the single twin to the whole set. A world can read clearly in one transcript and still wash out once many twins are pooled together, so we tested whether the worlds stayed distinct in aggregate. For each world we took the twins that responded most strongly to it, ran a set of exploratory conversations with them, then measured how much vocabulary each world's answers shared with every other (pairwise cosine similarity, where 0 is no overlap and one is identical). The worlds held apart (see Figure 6). No two shared more than about two-thirds of their language, the closest pair at 0.66 and the most distinct at 0.23, with an average of 0.45 across all combinations. The structure matched the blind panel: the worlds that score most alike here are the ones the experts were likeliest to confuse, while the most distinct pairs were almost never mistaken for each other. The two worlds in the Table 2 example, optimised body and mood machine, are among the closest pairs at 0.65. Therefore, the shift one twin showed between them was the subtle kind, two near-neighbours pulling the same person in different directions, not a contrast chosen because it was easy.

less alike — more alike

| | Earn your edge | Mood machine | Optimised body | Quiet everyday wellness | Sweat it out together |
|---|---|---|---|---|---|
| **Earn your edge** | 1.00 | 0.36 | 0.51 | 0.40 | 0.23 |
| **Mood machine** | 0.36 | 1.00 | 0.65 | 0.66 | 0.36 |
| **Optimised body** | 0.51 | 0.65 | 1.00 | 0.62 | 0.37 |
| **Quiet everyday wellness** | 0.40 | 0.66 | 0.62 | 1.00 | 0.37 |
| **Sweat it out together** | 0.23 | 0.36 | 0.37 | 0.37 | 1.00 |

Average overlap **0.45** Most distinct **0.23** Most similar **0.66**

0 = no shared vocabulary, 1 = identical

*Figure 6. How distinct are the five worlds?*

The first test asked whether the worlds did enough to change behaviour. The second asked whether they did too much, pulling the twins from the real people underneath. Success meant a balance: grounding the future twins in real people while still letting in enough of the future. We looked again at the grounding score, now across the future simulations. About 24% of what the twins said in their future worlds traced back to data the real person had shared with us today. That is lower than the 64% we saw for present twins, and it should be: these answers reason about a future the data cannot yet contain. What matters is that roughly a quarter still rests on real evidence, so the projections stay tethered to who the person actually is rather than drifting into invention. We also had to rule out a simpler explanation: that the twins were just a language model following a future briefing, with the real grounding adding nothing. Therefore, we ran the same five worlds through a model with no first-party grounding, prompted only to inhabit each world and compared the two through a topic analysis (see Table 3). They shared only part of their content. Where the ungrounded model diverged, it mostly took the briefing literally, over-producing generic, well-known material. In one world that mentioned saunas and cold plunges, for instance, it filled its answers with the standard benefits of both. The slice that counts is the topics that surfaced only in the future-twin answers: grounded insight a raw model, working from general knowledge alone, never reached.

| Where a topic appeared | Share of topics | What it means |
|---|---|---|
| **Both twin and model** | 36% | Common ground |
| **Model only** | 38% | Mostly the model taking the briefing literally (saunas, cold plunges): general knowledge, not insight |
| **Future twins only** | 26% | Added value: real, grounded insight the model did not reach |

*Table 3. Topic comparison, future twins versus an ungrounded model.*

Set beside the present-twin results, the picture holds together. Everything we could re-run held up. The grounding score fell in just the way a move into the future should, from 64% to 24%. The one test we can't run at all is a direct comparison against the real person: there is no 2030 yet to check against, so we treat that as a limit, not a finding.

Both tests held. The worlds genuinely change behaviour and read as plausible, and the answers stay grounded in real people while adding what a raw model cannot. None of this proves the twins are right about 2030, which nothing could, but it shows the instrument (as we came to think of the future twins) behaves as designed. That let us trust the projection mechanism itself. Moving a twin from its present self to its future self and into a world shifted its answers in patterned, world-aligned ways rather than at random. The future-self layer tracked what each real person had said about where their life was heading, and the world layer moved answers in directions consistent with the world it placed them in. With

that consistency in hand, we could finally turn the instrument on the question the whole programme was built around: how to make more confident decisions about platforms and concepts that have to land years from now.

## From Static to Dynamic Innovation Platforms

An innovation platform, as introduced earlier, defines where a brand will play and why, before any specific concept is on the table. It has to keep holding through product development, claims and/or benefits and go-to-market, where the strategy meets the tactical choices that put it to the test. Building a platform the way this programme did, drawing on expert input, consumer immersion, cultural foresight and the global stakeholder team, is best practice. The result is a static platform: well-founded, with a forward-looking lens carried in through the foresight and cultural work. What it does not have is change built into it. The platform captures one view of where the category is going. If the actual future plays out differently, or if the picture shifts, the platform itself does not adjust. It also does not hold more than one possible future at once. The capability to address that limit is what the future twins add. A dynamic platform extends the same logic with a methodical view of the future built into it. That view comes from the two layers we have just described: each person's own future self, and the five 2030 worlds around them. The backbone is still a consumer truth, but where a static platform rests on a present-day insight, a dynamic platform builds from an enduring tension—one that holds whichever future arrives. From that backbone, the platform branched into five trajectories, one per future world. The structural shifts are summarised in Table 4.

| | Static innovation platform | Dynamic innovation platform |
|---|---|---|
| **Backbone** | A consumer truth tied to today's context. | An enduring tension that holds across all five 2030 worlds. |
| **Future view** | One implied future, often extrapolated from today. | Five 2030 worlds, plus a baseline of shared certainties. |
| **World-specific content** | Not part of the platform. | Five trajectories, one per future world. |
| **Updating over time** | Fixed at the moment of writing. | Built to be updated as signposts confirm or unsettle the worlds. |

*Table 4. From static to dynamic innovation platforms.*

The deeper change was in how the team approached the analysis. Working this way pulls the team away from chasing the insight of the moment and toward crafting tensions sturdy enough to last. What made this possible was the twins themselves. They let us explore opportunities and needs in two directions at once: bottom-up, through each twin's reading of how their needs evolve across the worlds, and top-down by exposing them to hypothesised opportunities. That dual lens is what surfaces the enduring tension underneath, and the world-specific shape it takes. For each platform, the tension was taken into each of the five worlds and worked through in detail: how the platform's central job lands in that world, what role the drink plays, which functional and emotional benefits people will look for, how the twins respond in their own voice and which seed ideas the world points toward. Alongside this world-by-world reading, we also looked across the worlds: what holds true in every one of them, where they pull apart along key levers and the early signposts that would tell us a particular world is arriving.

The case for a dynamic platform extends beyond its durability across futures. Innovation literature has been clear for some time that the front end of innovation, the brief that sets up ideation, governs how creative the work downstream can be. The broader and more deliberately a team searches the problem space at this stage, exploring more diverse angles, sources and framings before converging, the more ideas it generates and the higher its chance of radical innovation (van den Ende et al., 2015; Cortellazzo et al., 2025). The platform brief is that front end. If a dynamic platform opens more routes than a static one, that should be visible in the platform itself before any team has ideated against it. To test this, we ran a structured content analysis of two executions of the same platform: a static version framed around a single assumed future, and a dynamic version framed around five plausible 2030 worlds. The brief and the topic were held constant. What differed was the underlying

evidence base feeding each version (the core insight programme for the static one, the future twins for the dynamic one) and the framing structure that evidence supported. Every cue in each platform was coded into a discrete ideation direction: a distinct creative route a team could pursue, such as a format, an occasion or a sensory signature. An ideation direction is not a finished idea; it is the doorway from which several ideas can be developed. To be coded as a direction at all, a cue had to clear a quality bar: distinct from the others, on-brief and concrete enough to ideate from. Quality was the ticket in; what we scored was stretch.

Each direction was mapped to one of 11 conceptual territories that span the design space (e.g., ingredient and formulation, format and packaging, affect and mood). We scored each direction's stretch from one (the slide basically states it) to five (a remote, surprising leap), using three as the line for a direction that takes a productive step beyond the obvious. The coding and scoring followed this fixed rubric and were carried out by a large language model, applied identically to both platforms. Because the platforms were human-built, the model scored human work against a predefined rubric rather than its own output. A researcher pulled the directions from a subset of the slides independently and landed on the same set, differing only on how far an individual direction stretched (see Table 5).

| Measure | Static | Dynamic | Dynamic versus static |
|---|---|---|---|
| **Distinct high-quality ideation directions, core framing** | 9 | 20 | 2.2× |
| **Distinct high-quality ideation directions, incl. seed ideas** | 12 | 35 | 2.9× |
| **Conceptual territories covered (of 11)** | 6 | 10 | 1.7× |
| **Directions that push past the obvious (scored 3+)** | 17% (2 of 12) | 66% (23 of 35) | ~4× |
| **Ideation directions per 100 words** | 5.3 | 3.1 | 0.6× |

*Table 5. Ideation richness, static versus dynamic platform.*

The dynamic platform opened more than twice as many distinct ideation directions on its core framing, and close to three times as many with seed ideas included. It also reached into territory the static version did not cover at all. The two measures answer different worries. In a moment when AI can produce directions in any quantity, the count only admits routes solid enough to ideate from, not the filler AI-supported work easily runs to and neither platform drifted off-brief. The stretch score is where the platforms part ways: most of the static platform's directions are good but familiar roads, while two-thirds of the dynamic platform's take a productive step beyond the obvious, the forward-leaning directions a 2030-oriented brief should be expected to generate. The advantage holds before seed ideas are counted, so it sits in the platform's structure rather than in directions already worked out on its behalf. The static platform leads on one measure: because it is shorter and tightly written (around 170 words of core framing, against around 650 words for the dynamic platform's five-world framing), it concentrates more ideation directions per 100 words. The dynamic platform's edge is not density line for line; it is the breadth and reach of the territory it opens up, the variety of places a team has to go from there.

From the Coca-Cola team's perspective, the difference between a static and a dynamic platform was immediately apparent in use. Framing the work around five 2030 worlds drew the team into the future explicitly, and made clear that decisions had to be made for a particular version of 2030, rather than an unspecified one. Where the worlds pulled in opposite directions on a particular question, those tensions surfaced as design choices the team could weigh rather than as ambiguity to work around. The same logic carried into how different target groups landed across the worlds. A segment that connects more strongly to one world than another sharpens who the team is designing for, and what that audience needs from the brand in that specific future. The gain lands in two places: the breadth of directions a dynamic platform opens, which is what the analysis above measured and the sharpness of the design brief each direction can support. Acting on a dynamic platform looks different, too. The platform does not postpone anything; the commitments still come first, long before the evidence. What it changes is the confidence behind

those commitments. Where a bet holds across every world we mapped, the team can commit with high conviction. Where the upside depends on which world arrives, the team commits deliberately, with the assumptions visible and the signposts to watch already named.

The deliverable is only part of what makes it dynamic; the rest is the practice of updating it. Each world comes with signposts attached, early indicators that the world in question is taking shape: a competitor launching a particular kind of product, a regulatory shift, a sustained behavioural change. When a signpost confirms a world, the platform's emphasis shifts. If a new signal points to a world we had not anticipated, the set evolves. The platform is built to be maintained, kept in step with how the future is actually going. Innovation platforms were one way we put the future twins to work, at the level of strategy. The next application was closer to the ground: a single concept, with a defined launch date.

## Co-creating Concepts for the LA 2028 Olympics

Innovation work for a portfolio of this kind can sit in different layers:

1. The foundation: extending and strengthening the core of the portfolio already in market;
2. The next era: ambitious moves where new technology or formulation takes the brands forward;
3. The limited time offer (LTO): an activation built around a major moment to create buzz at a specific point in time.

This project's strategic programme was anchored in the first two. A parallel application that emerged once the twins were in place sat in the third, and gave us a different job for the same twins.

That LTO was for the LA 2028 Olympic Games. For a portfolio led by Powerade, the Olympics is a defining moment, both commercially and for what it signals about the brand's place in sport. Coca-Cola has been a long-standing Olympic partner, and had been developing a Powerade concept for the Games with creative agency Ogilvy. It had already been through several internal rounds and was being sharpened ahead of a 2028 launch. The job here was optimisation. With internal conviction already behind the concept, we used the twins to stress-test what was working, identify what was not, and refine through co-creation rather than treat them as respondents at the end of a fixed round. The baseline came from real consumers in the present: the live insight community reacted to the concept as it stood. The future twins reacted to the same concept from inside each of the five worlds, showing us how it would land in different versions of the future. For this experiment the present-day reactions to this specific concept were deliberately kept out of the twins' memory before the future test, so the two methods could be assessed on independent ground. In live practice you would feed that concept feedback back in, and the grounding is stronger for it. The real gain, though, is what the five-world view opens up: seeing how one concept lands across each future lets the team choose which world to design for and optimise the concept for it. Because the launch is in 2028, that future lens was the more decisive test: the world the concept actually arrives into will not be today's.

This kind of work is not available through traditional concept research. A typical concept test puts a concept in front of today's consumers and gives one round of feedback. The twins made something different possible: feedback grounded in future worlds, back overnight and at a fraction of the cost of a re-field. That is the order of magnitude that turns respondents into collaborators in an iterative loop. React, refine, re-test, refine again. Neither the live community nor the future twins did the writing themselves. What each stream produced was reaction, language, objection and reframing. The team then turned each stream into a complete rewrite of the original concept: a fully revised version carrying that stream's logic through to product detail, claim set and use case. The two rewrites that resulted, one based on the live community and one based on the future twins, are what the analyses below compare.

The optimisation job at this stage rests on two decisions that have to be made in parallel, both ahead of the wider work that follows (packaging, in-store execution, activation and so on). The first is whether the concept is directionally right for the audience it has to land with: a desirability question. The second is what the product must deliver specifically: a brief R&D can lock in and act on. R&D works on long lead times, so that brief has to be settled years ahead of launch, not nudged into shape as the date nears. The future-twin work has a different role on each. The

desirability question already has an answer: a quantitative test that puts the concept in front of consumers and measures purchase intention, uniqueness and brand fit. The future-twin work is not built to replace it. What it adds, upstream, is purposefully different rewrites of the same concept, each anchored in one future world and carrying its own product detail, claim and/or benefit set and use case. A test of real strategic alternatives learns more than one of copy variants. One caveat: to read forward rather than back, that test has to run with leading-edge consumers, people whose lives already carry early signals of the worlds being tested, not a representative present-day sample.

The R&D brief question runs differently. R&D needs the brief locked soon enough to act on it, well in advance of launch. To compare what each method gave R&D in the rewrites it produced, we ran a structured content analysis on both versions, applying the same technique used for the platform comparison earlier in the paper. Each rewrite was coded against six pre-defined intervention areas surfaced from concept structure: core framing, Olympics connection, audience, use case, variant design and proof and credibility. We scored the specificity of each recommendation on a one-to-five scale. The six areas split into two groups: those that shape positioning and audience direction (core framing, Olympics connection, audience), and those that shape what the product must deliver (use case, variant design, proof and credibility). Between them, the second group covers ingredients, doses, claims and/or benefits, label content and use protocol. On the positioning areas, the two rewrites took different angles, but landed at the same level of specificity. The finding that counts is on the R&D side, where specificity matters most. R&D needs language concrete enough to act on, and that is exactly where the gap between the two methods lives (see Table 6).

| Group of intervention areas | Traditional rewrite, average specificity (1-5) | Twin rewrite, average specificity (1-5) | Difference |
|---|---|---|---|
| **Positioning areas (core framing, Olympics connection, audience)** | 3.7 | 3.7 | None |
| **R&D-brief areas (use case, variant design, proof and credibility)** | 3.3 | 4.5 | +1.2 |

*Table 6: Specificity comparison across intervention areas, grouped by whether they shape positioning or what the product must deliver.*

The two methods converge on positioning specificity. They differ on the depth carried into ingredient, dose, claim and/or benefit, label and protocol detail, which is the language R&D needs to work from. The twin approach does not produce more recommendations. It produces more concrete ones, specifying what the product has to deliver, from ingredient, to dose, to claim and/or benefit. One observation from the analysis is worth flagging in its own right. Working with future twins gave the team better access to the reasoning behind each recommended change. Because each twin carries a memory layer drawn from the real person it represents, the team could trace a specific reaction back to lived experiences that person had actually described during the live community work. When a twin pushed against a particular ingredient, claim or framing, the memory beneath it pointed to the experience that produced the resistance. That kind of trace is harder to follow in traditional concept rounds. Each session adds new responses without a continuous memory underneath, and the team is rarely working with the same person across rounds against a foundational reading of who they are and what they need.

From the Coca-Cola team's perspective, what stood out was that the future-twin work changed how the product itself evolved, not just the story around it. The two methods produced fundamentally different R&D briefs. The twin brief was more future-forward, better aligned with the conditions the concept will meet when it launches in 2028. Alongside this, the work surfaced a strategic conversation worth having before the next concept goes through the same instrument: which world should we be designing for? The answer is shaped by where the brand is going, how this relates to the context of the Olympic moment and which of the five worlds could realistically have begun to arrive by 2028. Choosing the world is part of the brief. This shapes how the work runs from here. The increased specificity in the twin rewrite brings the concept closer to the level of detail an R&D brief requires, and the engagement with the twins can be designed to push it further. Because the twins do not tire across rounds, the iteration can run as deep as the brief needs. Together with the platform work in the previous section, this second application

showed the versatility of the instrument. The same future twins that anchored multi-year strategic platforms also drove fast, iterative work on a single concept with a fixed launch moment; different jobs from the same foundation.

## What this changes

Future twins are not a way to verify how a real person will react in 2030 to a real product, a real claim, a real category shift. We cannot calibrate them against a future none of us can yet observe. The lives of the people behind the twins will change in ways the future-self layer can only approximate. The world will change in ways the worlds' layers can only sketch. There is no honest version of this work that claims otherwise. Nonetheless, predictive accuracy was never the test. The test is whether teams make better and more confident decisions with the instrument in hand than without it: whether they commit with more conviction to a long-horizon pipeline, and whether they walk into a portfolio choice or a concept rewrite with the same reference points everyone in the room is reasoning from. We are not the only ones to observe that teams decide better this way: in controlled studies, reasoning through several structured futures rather than a single forecast reduces framing bias and improves decision quality (Meissner & Wulf, 2013), a result echoed in early-stage design decisions (Parolin et al., 2023). The effects measured there are modest and experimental, but they point the same way our experience did. Confidence built on data, even synthetic data grounded in real people, is what aligns decision-making. That is the value, and it does not depend on knowing things we cannot know.

Innovation strategy used to come in one-shot deliverables that aged from the day they were finished. The work in this project shows a different shape: foundational thinking with a base layer and conditional layers built off the worlds, refreshed as signposts confirm or unsettle the pictures we drew. The shelf life of foundational work gets longer because it is built to be updated, not preserved. We showed this at two ends of the spectrum: dynamic innovation platforms at the level of where to play, and fast iterative concept work at the level of what to launch into a specific moment. One instrument carried both. Future twins also change whose voices innovation work can draw on. Professional athletes were the most striking case here: traditionally hard to recruit, harder to schedule and almost impossible to engage in iterative work that needs them to react, react again, then react a third time. A twin trained on a real, consenting athlete is available as a collaborator in a way the athlete themselves rarely can be. Moreover, they change what happens to insight after it has been generated. Research, foresight and cultural analysis have tended to sit in decks. The decks age. A new project gets commissioned. Built on the same insight base, a twin is a different kind of artefact: it can be asked new questions, run against new concepts and brought into new conversations after the underlying research closes. A twin ages too: the data it rests on dates like any other, and it has to be refreshed to stay reliable. Insight still ages. What changes is that it becomes a living layer the organisation keeps working with and updating, rather than a snapshot it re-commissions from scratch.

This paper opened on timing, and on Bill Gross's finding that timing is the single biggest variable in whether an innovation succeeds. The problem with timing is that the evidence arrives too late: innovation decisions get made years before there is any way to know whether they were well-timed. The future twins do not solve that; nothing solves that. However, they shift where teams stand when they make the choice. Instead of reasoning forward from today's consumers alone, the team reasons from people placed in plausible versions of where the category is going. The decision still gets made before the future arrives, but it gets made with a more honest set of reference points than the present-day-only alternative makes possible.

## Acknowledgements

We thank Adam Palenicek, Ruchira Bhattacharya, Haiyu Ren and Constanza Alzugaray at The Coca-Cola Company for their partnership and willingness to take this work into new territory. At Human8, Thomas Troch, Meredith Linhart and Anita Janssen led the project across the programme, with Troch also among the team behind the digital twin and future twin experiments, alongside Annelies Verhaeghe, Ben Worsley, Carlos Lorenzo, Daniel Mihaies, Mike Elder, Nadine Vávra and Thalia Hanko. Thanks to Alex Bee, Francesca Simon Millar, Gareth Lewis, Kira Collins and Sui Lai Kang at Space Doctors for the cultural foresight that enriched the project throughout. Thanks also to Reid Litman and team at

Ogilvy for their strategy and ideation. Finally, thanks to Hannes Willaert for designing the graphics, to Tom De Ruyck, Angie Deceuninck and Katia Pallini for sharpening the argument, and to the expert panel that reviewed the future worlds assets.

## References

Bayus, B. L., Jain, S., and Rao, A. G. (1997). Too little, too early: Introduction timing and new product performance in the personal digital assistant industry. *Journal of Marketing Research, 34*(1), 50–63.

Calantone, R. J., and Di Benedetto, C. A. (2012). The role of lean launch execution and launch timing on new product performance. *Journal of the Academy of Marketing Science, 40*(4), 526–538. https://doi.org/10.1007/s11747-011-0258-1

Circana. (2025). *The ripple effect of GLP-1s, today and in the future.* Available: https://www.circana.com/post/the-ripple-effect-of-glp-1s-today-and-in-the-future

Cortellazzo, L., Bonesso, S., and Batista-Foguet, J. M. (2025). At the early stage of the ideation process: Measurement, antecedents and outcomes of individual exploratory behaviours. *Creativity and Innovation Management, 34*(3), 706–720.

Di Benedetto, C. A. (1999). Identifying the key success factors in new product launch. *Journal of Product Innovation Management, 16*(6), 530–544.

Ende, J. van den., Frederiksen, L., and Prencipe, A. (2015). The front end of innovation: Organizing search for ideas. *Journal of Product Innovation Management, 32*(4), 482–487.

Gross, B. (2015). The single biggest reason why start-ups succeed [Video]. *TED Conferences.* Available: https://www.ted.com/talks/bill_gross_the_single_biggest_reason_why_startups_succeed

Loewenstein, G., O'Donoghue, T., and Rabin, M. (2003). Projection bias in predicting future utility. *The Quarterly Journal of Economics, 118*(4), 1209–1248.

Maier, B. F., Aslak, U., Fiaschi, L., Rismal, N., Fletcher, K., Luhmann, C. C., Dow, R., Pappas, K., and Wiecki, T. V. (2025). *LLMs reproduce human purchase intent via semantic similarity elicitation of Likert ratings* (arXiv:2510.08338). Available: https://arxiv.org/abs/2510.08338

Meissner, P., and Wulf, T. (2013). Cognitive benefits of scenario planning: Its impact on biases and decision quality. *Technological Forecasting and Social Change, 80*(4), 801–814.

Parolin, G., McAloone, T. C., and Pigosso, D. C. A. (2023). The effects of scenarios on decision-making quality in early design: An empirical study. *Proceedings of the Design Society, 3*, 3375–3384.

Pierce, A., Beaudin, L., Gupta, N., Rajasekar, V., Lin, C., Abdel Motaal, B., and Nairn, H. (2026). Synthetic customers earn their stripes. *Bain & Company.* Available: https://www.bain.com/insights/synthetic-customers-earn-their-stripes/

Robinson, T., and Veresiu, E. (2024). Timing legitimacy: Identifying the optimal moment to launch technology in the market. *Journal of Marketing, 89*(3), 136–153. https://doi.org/10.1177/00222429241280405

Suarez, F. F., Grodal, S., and Gotsopoulos, A. (2015). Perfect timing? Dominant category, dominant design, and the window of opportunity for firm entry. *Strategic Management Journal, 36*(3), 437–448. https://doi.org/10.1002/smj.2225

Tetlock, P. E. (2005). *Expert political judgment: How good is it? How can we know?* Princeton University Press.

van den Ende, J., Frederiksen, L., and Prencipe, A. (2015). The front end of innovation: Organizing search for ideas. *Journal of Product Innovation Management*, 32(4), 482–487. https://doi.org/10.1111/jpim.12213

Verhaeghe, A., Troch, T., Misse, E., and Hart, P. van. (2024). *Unveiling the Power of Synergy in Innovation: Battle of the Titans: Human versus Machine*. Paper presented at Esomar Congress 2024, Athens, Greece.

## About the Authors

Thomas Troch, Partner & Head of Innovation, Human8, US.

Annelies Verhaeghe, Chief Innovation Officer, Human8, Romania.

Adam Palenicek, Senior Director Global Human Insights, The Coca-Cola Company, UK.

Copyright © **Esomar** 2026

# Embracing the Unknown...

## *In Search of a Predictive Model Grounded in Real Human Understanding. Madness – or Method?*

*Muriel Esposito*

*Laurence Minisini*

Esomar

Office address:

Burgemeester Stramanweg 105

1101 AA, Amsterdam

Phone: +31 20 664 21 41

Email: info@Esomar.org

Website: www.Esomar.org

Publication Date: September 2026

ESOMAR Publication Series Volume S424 Congress 2026

ISBN 92-831-0337-8

Congress 2026
Copyright © **ESOMAR** 2026

Copyright © **ESOMAR** 2026


# Embracing the Unknown…

## *In Search of a Predictive Model Grounded in Real Human Understanding. Madness – or Method?*

*Muriel Esposito*

*Laurence Minisini*

### The Business Challenge

In mid-2025, global flavour giant Givaudan approached partner agency Esposito Research & Strategy with a bold brief: Could we help them decode what makes a food flavour iconic?

Givaudan were looking for a radical new approach for savoury snack innovation in Europe. Their ambition was to crack the codes of flavour success, to identify tomorrow's flavour stars early and maximise commercial success in flavour innovation.

The stakes were incredibly high. In a category that saw 350 new flavour launches in 2025 alone, our mission was to predict how new flavours could succeed and scale.

This paper reveals how an exploratory approach helped Givaudan crack the codes of flavour success, solved a decade-long industry challenge and unlocked an untapped frontier for growth.

### Tackling the Brief

Experimental thinking was paramount: there was no prescribed methodology, no rigid hypotheses, no "off the shelf" solution to address this challenge.

Our mandate was to go wherever the evidence took us… with one crucial imperative: the final output needed to be inspiring, pragmatic and commercially usable across the client's organisation. The remit was Europe, with a particular focus on big snack markets for our client: Germany, Poland and the UK. Early signals could, however, come from anywhere.

We were excited – and nervous. Could research deliver something as impactful and ground-breaking as the brief required?

### Our Approach

We decided to use the innovation diffusion curve[1] as our investigative framework, to help us map and explore the journey of flavour success from emerging to iconic flavours.

---

[1] Inspired by Bass's cumulative adoption model, which shows how products get adopted over time. In our project, we overlaid the types of consumers key to each phase of growth, as defined per Everett Roger's innovation diffusion.

Bringing together cumulative adoption and time, this conceptual framework enabled us to focus on the shifts from each distinct stage and interrogate the drivers behind those pivotal moments. We overlaid the adopters-to-laggards categorisation often used in innovation research to ensure we captured insights on who might be influential at particular points in time.

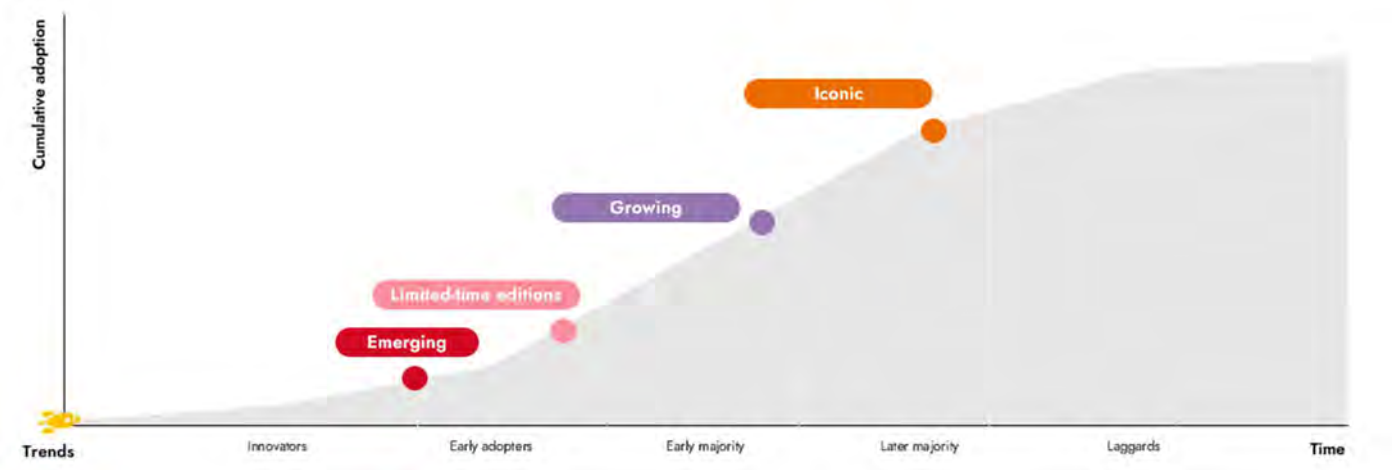



*Figure 1.*

The research design spanned three consecutive stages: moving from exploration to validation, then consolidation, involving multi-disciplinary inputs. Our guiding principle throughout the project was to connect diverse perspectives and stories, exploring insight gaps and overlaps to uncover the dimensions of flavour success. That's why we went "across" everything: culture, science, purchase channels, borders, generations, disciplines, qualitative and quantitative – and, importantly, time.

At heart, this was the new frontier in our approach: the way we wove different specialisms together, blending data and emotions to map uncharted insight territory and ultimately create a unique predictive model.

Two bold decisions proved fundamental:

i) withholding primary consumer research until the final stage

ii) rejecting an AI-led route.

As flavour research specialists, we know that consumers can rarely articulate the complexity of why they love a popular flavour profile, and even the most powerful AI-driven language models risk missing the subtle nuances of cultural resonance. Relying on either to decode why, for example, Paprika flavour became iconic in specific markets, would have jeopardised the power and depth of our insights.

To identify success factors connecting science with culture, logic with intuition, and past with present, our investigation began with a complete, rigorous synthesis of facts.

### *Stage 1: Exploration*

We began by isolating potential success factors through rigorous desk research and collaborative workshops, interrogating established iconic flavours (e.g., Paprika, Cheese, BBQ) to determine how their trajectories could apply to future innovations.

Iterative brainstorming was the cornerstone of this stage, bringing together insights, marketing, technical, sensory, and snack-category experts alongside local research specialists in key markets. These debates were fuelled by an expansive data ecosystem: over 75 client sensory flavour maps,[2] as well as literature reviews[3] and expert insights reports.[4]

Among our investigations, we dived into the role of societal trends on snack flavour preferences. For instance, looking back in time, there was a clear link between the perception of the USA as aspirational, the rise of informal eating in Europe and the eventual popularity of barbecue flavour – translating how shifts in society's attitudes impact flavour popularity was something we explored in depth with the client teams and project experts.

To understand how emerging trends today might play out in snacks tomorrow, we turned to Gen Z. Social media listening had revealed that Gen Z consumers were breaking snack flavour rules in unique ways: there were hundreds of thousands of posts about unusual flavour combinations and novel uses of snacks as ingredients. With our expert group, we analysed how this might be connected to wider societal shifts.

**Chaos Cooking & Hacked Snacks**

Gen Z influencers and followers were breaking the rules with snack flavours. Why was this happening?

Source: Instagram, @fannychupp

Source: TikTok, @thecaketoonist

Our semiotics and anthropology experts linked this behaviour to significant societal shocks since 2020: COVID-19, cost of living crisis, nationalism, climate change, AI, together with a "snackification" of social media. All were having a profound impact on the way teens and young adults perceived the world and their role within it. As the existing order was fragmenting around them, the sense of "the impossible is possible" translated into food.

Enter Chaos Cooking: transforming humble snacks into unique, quick but highly impactful experiences, that shock and delight in equal measure, express individuality and boost social status. My world, my rules.

This "hacked snacks" phenomenon helped us understand the link between macro-trends and emerging flavours, leading us to bring several societal features into the dimensions of success.

*Figure 2.*

---

[2] Client's sensory mapping of iconic savoury snack flavour profiles across multiple European markets.

[3] Articles and papers in the public domain, such as publications from the Food Institute (www.foodinstitute.com), blogs (e.g., www.generationzalpha.com), press articles (e.g., How do you reduce a national dish to a powder? The weird, secretive world of crisp flavours, | Snacks | The Guardian), trends (e.g. www.insighttrendsworld.com), retail/food service sources (e.g., https://newsroom.justeattakeaway.com), social media (e.g., www.disa.org).

[4] Mix of client proprietary insights, social media listening and secondary data reports.

For instance, this led us to uncover the phenomenon of "Chaos Cooking" a.k.a. Hacked Snacks, summarised in the example above.

As we reached the end of this stage, we had synthesised all insights into 10 core dimensions of success, comprising 56 individual event hypotheses. Our next task was to determine their relative importance and impact.

### *Stage 2: Validation and Prioritisation*

As we reached this stage, we needed to stress-test our framework and home in on what exactly needed to happen when, for a flavour to move up the curve. We set about identifying specific order and time effects, and assessing the relative impact of individual events on a flavour's chances of success.

For instance, was social media traction a predictor of long-term commercial success, or more of a longevity risk? Was a flavour's cultural resonance more or less impactful than brand popularity?

To answer this, we initiated an iterative research loop where historical insights, emerging trends and on-the-ground market realities constantly informed and validated one another.

At a macro-level, we worked with European academic and technical experts[5] to bring to light the unconscious drivers of flavour preference across human behaviour, food cultures, trends and sensory experience – uncovering how historical events from centuries ago still influence flavour preference today.

At a micro-level, store visits, product images, and pack semiotic analysis enabled us to define the flavour dimensions of several countries, directly contrasting why the same flavour can triumph in one market but fail in another. These localised differences were fertile hunting grounds in validating the impact of specific factors.

Throughout this stage, by continuously intersecting macro and micro insights with our expert panel and in ongoing client workshops, we bridged many seemingly unconnected or hidden factors that influence a flavour's success potential.

The next example illustrates the long-lasting impact of food heritage, as we set out to explore the cultural roots of today's flavour icons such Paprika and Salt & Vinegar.

Why was Paprika an iconic flavour in most of continental Europe while it failed to capture the British market?

Why was Salt & Vinegar flavour iconic only in the UK?

History provided the answer.

---

[5] We brought in three leading academics specialising in food cultures (a food historian, a cultural anthropologist, a culinary art historian), three trend experts (a semiotician, a futurologist, a food journalist) and six sensory experts (an Oxford professor, five flavour creation specialists), across the client's priority markets.

How could Paprika be an iconic flavour in so many diverse countries around Europe, and not in Britain? Because of empires! Introduced to Spain from South America in the early 1500s, paprika travelled East with Spanish merchants, crossing Ottoman and Habsburg trade routes.

Its ability to season simple, often potato-based dishes rapidly made it a culinary staple throughout the Habsburg empire (Netherlands, Spain, Austria), adjacent lands, and Ottoman-influenced nations (Balkans, Turkey).

Later on, Paprika was often the first flavoured crisp variant to be launched in these markets, as the familiar combination of paprika with potatoes helped American-style crisps enter food repertoires. It remains a highly popular flavour today.[6]

Why did Paprika flavour not succeed in Britain? Because it was Britain's geopolitical rivals - Spain, Ottomans, the Holy Roman Empire - who, century after century, drove paprika's expansion. The spice never entered British food traditions!

Why is Salt & Vinegar one of Britain's most popular savoury snack flavours, when it has limited appeal elsewhere? Because of Britain's unique industrial and seafaring heritage.

In the 18th century, Britain was already mass-producing malt vinegar as a byproduct of its brewing industry. A century later, railway expansion brought fresh fish to industrial cities. Paired with fried potatoes, a popular staple was born: fish and chips. Malt vinegar's sweet acidity cut through the greasiness and intense saltiness of this simple dish.

In contrast, most European markets have no tradition of using sweet vinegar with potatoes: they prefer softer acidic notes from dairy products like sour cream or mayonnaise.

*Figure 3.*

As we sharpened our understanding of how diverse individual elements impact flavour success, we continually fused human behaviours and category dynamics, history and food trends, culinary science and sociology.
Importantly, whilst the 10 overarching success factors were fixed, the attributes defining those factors allowed for flexibility in terms of brand fit and audience type. This reflected market realities: a potential flavour icon for a youth-focused brand would likely be different from that of a traditional premium player.

This generational perspective – along with other behavioural trends – suggested that in future, consumers across Europe might expect more intense snack flavours, with new trends spreading more quickly across markets. For brands, this means that the ability to pivot and identify early flavour winners in their innovation pipeline becomes even more critical in the near future.

The following example shows how we decoded the "spicy" phenomenon led by Gen Z, its role and its impact on the evolution of snack flavours in Europe.

[6] Nearly half of all potato crisps consumed in Germany today are Paprika-flavour. https://www.courrierinternational.com/article/snack-pourquoi-les-chips-au-paprika-font-craquer-les-allemands_233942

**Spicy: Dialling up the Heat**

Spicy flavours are growing in European snacking. Driven by Gen Z, heat is no longer just a taste preference: it is a pan-European cultural rite of passage and a valuable social currency.

On social media, young consumers flock to "spicy challenges", sharing their attempts to consume ultra-spicy snacks in the hope of gaining social kudos. As one expert noted: "*They don't dare bungee jump, but they dare eat chilli*."

Source: foobetrading.com

To signal this "fiery experience", youth-targeting brands like Takis, Doritos, and Pringles have dialled up heat cues on pack - as shown in the Pringles example below, summarising semiotic analysis.

This thrill-seeking behaviour is raising the entire region's spice baseline in snacks and beyond. Even in historically mild flavour markets like Germany, mainstream brands are cranking up the heat to keep pace with youth culture and address the growing acceptability of spice.

*Figure 4.*

The end-point of this intensive synthesis process was to transform these critical success factors and their diverse attributes into a single observation-based mathematical model that accurately captured the full dynamics of flavour success.

## *Stage 3: Consolidation into a Fully Working Model*

We had validated all the building blocks and pathways that lead to iconic flavour status. Now, to achieve accurate outcomes in a fully operational commercial model, we needed to quantify exactly how much predictive leverage to assign to each variable. Translating influences and attributes into precise figures required a sensitive, data-driven mathematical calibration.

For instance, what should be the numerical value of a flavour's alignment with popular cuisine? Was it worth more or less than its appeal to early adopters?

Using historical performance data of known flavour icons and past market failures, we pressure-tested the model to see if it could accurately and retroactively predict the correct outcomes. Thanks to Givaudan's database, we knew which existing flavours were actual icons, and which had failed: would our model predict them accurately based on independent inputs?

To supercharge this validation phase, Givaudan deployed a large-scale, multi-market quantitative survey, capturing inputs from over 5,000 consumers across hundreds of flavour ideas, from existing stars to emerging trends. This provided a robust data baseline to feed into the model, allowing us to plot known icons versus emerging trends, compare several markets and contrast Gen Z and early-adopter preferences against traditional cohorts.

As we piloted dozens of flavour ideas, it was encouraging to realise that only minor adjustments were required to finalise the mathematics of success and consolidate our original framework.

Our fully working model's maiden voyage was a co-creation workshop, which brought together the client's insights, marketing and flavour creation teams from across Europe to work with the tool and its supporting insights in order to create some new flavour ideas – a live test, as it were, in front of 30+ senior stakeholders.

Here is an example of how we used the model in the session.

**Using the model to identify the most promising Chicken flavour variant**

Chicken is a growing flavour trend across savoury snack food. But which chicken inspiration has the most potential for an upcoming savoury snack brand targeting young adults?

Bringing Givaudan's consumer data into the model, we were able to directly compare several chicken directions side by side across the 10 success dimensions, such as Korean Fried Chicken, Jamaican Jerk, Chicken Chasseur, Teriyaki, Barbecue, Chicken Nuggets etc.

In under 2 minutes, the model revealed which direction had the best fit for Gen Z and what was driving those high scores. Unfortunately, we cannot disclose the outcome of this simulation, but the model enabled the client to prioritise the innovation route with the most potential.

*Figure 5.*

As co-facilitators of this high-profile session, we were able to witness firsthand how easy the model was to understand and work with, and how impactful the narratives we had uncovered throughout the project were to the whole audience. Our predictive model had fulfilled Givaudan's ambition. To further consolidate the value of the model, Givaudan developed flavour ideas created in the European workshop into prototype snacks, transforming data into the taste of success!

Subsequently, the client team refined the tool's look and feel prior to making the model accessible to its teams around Europe, along with supporting materials.

Copyright © **ESOMAR** 2026
Congress 2026

## What We Delivered

Did we crack the codes of iconic flavours? Givaudan's opinion was a resounding YES!

Our pioneering approach worked: we solved the "iconic flavours" challenge by creating a rigorous mathematical scoring model informed by a unique blend of diverse disciplines and data. Anchored in real consumer behaviours, the model recognises that achieving iconic status requires a combination of factors and that there is more than one route to success. The tool's design enables Givaudan to assess the strengths and weaknesses of any snack flavour in detail, and to maximise the flavour fit based on country, audience and brand specificities.

An immediate impact is that early-stage failures have been eliminated, unlocking efficiencies: the model boosts the innovation process by prioritising likely winning ideas faster. Consumer inputs are integrated fully, with Gen Z and early adopters' flavour preferences futureproofing its predictive power and longevity.

Critically, our adventure didn't just uncover culturally rooted insights. It transformed them into an eminently practical and powerful predictive tool with commercially valuable results.

In the words of our client and co-author:

"*This builds a robust predictive model that puts all the evidence together in a simple way: we can take any flavour ideas, compare them, and see what we need to dial up or down to maximise their potential.*"

## The Wider Impact for the Client

As a direct result of this work, within months of launch, many flavour creations projects have been received across Europe and are under development by Givaudan for a variety of brands: from large multinationals to regional players. In terms of number of projects, the impact of this work at the time of writing exceeds that of previous Givaudan initiatives in snacks.

At a corporate level, this research gives Givaudan a genuine USP as a thought leader in flavour innovation for savoury products. Having delivered successful commercial outcomes for European snacks, this project is now generating significant interest in the business's other regions and product categories.

It has underlined the importance of insights, and enabled Givaudan's insights team in Europe to support business growth even more closely: thanks to this work, Givaudan sales teams are continuing to win new commercial discussions and innovation sessions with multinational snack brand owners.

The collaborative approach put in place for this project together with the tool's flexibility within and across markets, its accessibility and ease of use have led to it being adopted rapidly by diverse functions across the organisation in Europe. Inspired by the outcome, the business is also piloting a new dynamic to ensure future cross-team collaborations are equally productive.

The tool comes with a rich databank of flavour stories, consumer insights and expert evidence, demonstrating deep understanding of the snack category dynamics and bringing an inspiring narrative to the centre of the client's storytelling. Just as this explorative project blended science and emotions, so the deliverables combine data prediction with human resonance.

Flavour prototypes developed as a result of the tool provide a tangible illustration of what the model can help create. They are the ultimate "taste of success" and directly link insights to action.

This work is stimulating conversations far and wide: the company's own customers continue to be "wowed" by the ability to use predictive, culturally rich insights to deliver mathematically robust flavour recommendations.

## Overall Conclusions

This project proved that bold research briefs with an open approach to the "how" can lead to significant commercial step-changes.

Having the right blend of individuals and expertise was a fundamental success factor – but so was the team mindset. Two elements stand out: i) a culture of openness and mutual trust between client and agency teams, internal and external experts, core contributors and extended team members; combined with ii) inquisitive determination: going the extra mile was critical, even when not initially mapped out or part of the project plan.

To conclude, our approach demonstrated that connecting multiple disciplines to research insights can magnify their impact and uncover new solutions.

*With thanks to Givaudan for embracing this project's adventurous mindset and supporting this joint submission to ESOMAR.*

## About the Authors

Muriel Esposito, Director, Esposito Research & Strategy, France.
Laurence Minisini, Senior Consumer & Market Insights Manager, Givaudan Taste & Wellbeing, Switzerland.

esomar

# Decoding the dine-in micro-moments that build brand love for Nando’s BA and Nando's

*Anneri Venter*

*Simone Kakana*

Esomar

Office address:

Burgemeester Stramanweg 105

1101 AA, Amsterdam

Phone: +31 20 664 21 41

Email: info@Esomar.org

Website: www.Esomar.org

Publication Date: September 2026

Esomar Publication Series Volume S424 Congress 2026

ISBN 92-831-0337-8

Copyright © **Esomar** 2026

Copyright © **Esomar** 2026

Congress 2026

# Decoding the dine-in micro-moments that build brand love for Nando’s BA and Nando's

*Anneri Venter*

*Simone Kakana*

## Introduction

Nando's brief to us was simple to state and fascinating to answer: identify the emotional cues inside the restaurant itself that we can dial up across communications and channels, to build brand love through the food and the experience—not through advertising alone. Customer journey work already maps the touchpoints of a visit, so that part is not a mystery. What it rarely captures is how a guest feels at each of those points while it is happening. Brand trackers, satisfaction studies and exit surveys sit either side of the visit, and journey mapping usually reconstructs it afterwards, from memory. This brief was about the emotional signal inside the restaurant, in the moment. Thus, we went inside the visit. We focused on the dine-in occasion and captured a seven-second emotion signal from guests at five natural milestones: arrival, seated, order placed, drinks arriving and first bite. Those five moments are the spine of this paper.

Alongside them we measured “brand love”: a composite score of the guest-brand relationship, covering identity fit, emotional connection, passion-driven behaviours, long-term bond and separation distress. The two measures answer genuinely different questions. The moments describe how the visit felt while it was happening; brand love describes what the guest carries away from it. Holding both at once is the point of the study. A brand can know its brand love score and still have no idea which moment of a visit is building it or eroding it. Equally, it can map a customer journey in detail and never establish whether any of those moments reach the relationship at all. Our three questions followed directly:

1. What happens emotionally, moment by moment, inside a single visit?
2. Does any of it reach brand love?
3. What a change to any one moment might be worth before a business spends to find out?

There is a reasonable objection to measuring five moments. A well-established finding in psychology is that retrospective evaluations are dominated by the most intense and the final moments of an episode rather than by its average (Kahneman et al., 1993; Redelmeier & Kahneman, 1996). If memory works that way, why not simply ask afterwards? Our answer is that the peak-end effect describes how a memory is assembled, not which moments influenced the guest at the time. Both things show up in our data. First bite, the final moment in our sequence, is indeed the strongest single driver of brand love. However, the moments in between carry real weight, and asking afterwards would have made them considerably harder to see. Measuring live is what brings them into view.

## Methodology

We used a sequential, three-phase mixed-method design. Each phase answers a question the previous phase cannot, and each is built on the same guests rather than on separate samples. A note on the sequence of moments. Nando's dine-in format varies: In some markets guests are seated before they order, whereas in others they order first. We present the journey in the order that is most common across Nando's dine-in occasions, but guests logged each milestone as it occurred for them. A note on markets: fieldwork ran in three markets chosen for genuine cultural distance from one another, and for spanning very different stages of Nando's relationship with its guests. A home market where the brand is long established (Southern Africa), a mature market where it is familiar but competitive (Western Europe) and a market where it is still building presence (South Asia). To protect country-specific data and results, we refer throughout to regions rather than to individual countries.

Copyright © **Esomar** 2026


### *Phase 1. Quantitative: Where Emotion Moves*

Guests aged 16+ who confirmed they were planning a Nando's dine-in visit were recruited and screened beforehand. During the visit itself, guests self-logged a seven-second in-moment affect tap at each of the five milestones, followed by a two-to-three-minute post-visit survey containing the brand love statement battery and outcome measures. This in-the-moment capture is an applied form of experience sampling, developed precisely to avoid the reconstruction problems inherent in retrospective self-report (Csikszentmihalyi & Larson, 1987). The final targeted and achieved sample was n=120, evenly split at 40 per market. The sample was 27% aged 16 to 25, 23% aged 26 to 35, 28% aged 36 to 50 and 22% aged 51+, and 44% female versus 56% male. As a pilot, this base is sized to establish whether the approach detects real and consistent differences between moments.

### *Phase 2. Qualitative: Why it Moves*

12 45-minute online video in-depth interviews (four per market) were conducted with a subset of the same guests, within 48 hours of their visit. The timing is a design decision rather than a convenience: 48 hours is long enough for the guest to have left the environment and reflected, but close enough that sensory and situational detail is still retrievable. Using the same five-milestone framework, guests were asked to describe what they remembered rather than to re-rate what they had already scored. This phase does the work the numbers cannot. A score tells us a moment underperformed, but the interviews gave us the context behind each score, the guest's own language for it and the concrete behavioural detail from which realistic changes could later be written and tested. Because the same guests appear in both phases, we can also see where the two sources agree and where they diverge, which is itself a check on the quantitative reading.

### *Phase 3. AI: Sizing the Opportunity*

The “moment twin” is a small predictive model of the arrival-to-first-bite journey, trained exclusively on guest-level data from Phases 1 and 2. It is not a general-purpose AI model. It works as a collaboration rather than an automation. We seeded the model with suggestions based on the results and brainstorms with Nando’s, and the model then also generated further options of its own, including combinations we had not considered. Borderless Access and Nando's reviewed every option (ours and the model's) across several rounds, accepting, adjusting or rejecting each one for cultural tone, operational feasibility and brand fit. That review narrowed a long list down to eight final options per moment. Only those eight were then tested against the model to estimate the likely direction and scale of movement in brand love. Therefore, the AI contributed ideas, but what it didn’t do was decide. Every option that survived to be modelled had been read and approved by a human being on one side of the partnership or the other.

### *How the Three Phases Connect*

The three phases are sequential and dependent. Phase 1 establishes which moments show a consistent relationship with brand love, and therefore where the qualitative phase should concentrate. With 12 interviews available, that focus matters. Phase 2 explains the mechanism behind those moments in the guest's own language and supplies the behavioural detail from which realistic changes can be written. Phase 3 takes those reviewed changes and sizes them. In short: the quantitative identifies where, the qualitative explains why and the model estimates how much might be at stake.

## Findings

### *A Journey with No Weak Moment*

The first finding is that nothing is broken. Across all five milestones, mean in-moment feeling sits in a narrow band between 4.38 and 4.58 out of five. When guests were asked which moment felt least enjoyable, 43% said no moment stood out as bad at all.

Copyright © **Esomar** 2026


| Moment (mean /5) | Total | W. Europe | S. Africa | S. Asia |
|---|---|---|---|---|
| **Arrival** | 4.42 | 4.38 | 4.40 | 4.48 |
| **Seated** | 4.40 | 4.33 | 4.43 | 4.45 |
| **Order placed** | 4.38 | 4.30 | 4.53 | 4.33 |
| **Getting drinks** | 4.42 | 4.40 | 4.38 | 4.48 |
| **First bite** | 4.58 | 4.73 | 4.68 | 4.33 |

*Table 1*

This reframes the exercise mostly from remediation to elevation, and it raises the analytical bar: if nothing is failing, the question becomes which of several good moments is carrying the most weight and which is carrying none.

### *Deposits and Withdrawals: How We Read the Visit*

Before going into individual moments, we offer one piece of framing: throughout the qualitative work we read each moment as either a deposit or a withdrawal in an emotional account the guest holds with the brand. A deposit is any moment that makes a guest feel good, noticed, welcomed or reassured. It adds to their emotional balance with the brand. A withdrawal is any moment that creates friction or disappointment: being ignored, made to wait or left uncertain. It draws that balance down, even when everything else about the visit is fine. The useful part is the scale of the entries. Almost nothing guests described in this study was large. No disasters, no service failures, no bad meals. The account moves in units of eye contact, a single sentence, 30 seconds of attention. That is what makes the findings actionable; small entries are cheap to change.

### *The Moment Nobody Claims*

Asked which single moment felt best, guests chose first bite (39%), as expected. The revealing answer is at the other end. Getting drinks was named best by only 8%, the lowest of the five, while also being named least enjoyable by only 7.5%, among the lowest. Drinks are not a problem; it is an absence—i.e., the one moment that passes without registering in either direction. The interviews explain why, and it has very little to do with the drink. Consider the guest's position at that point in the visit. They have ordered. They have sat down. They are hungry—that is why they came. They are now waiting, with no information about how long the wait will be. The drink arrives, they take a sip and then nothing happens. There is no further contact until the food appears. That is the stretch in which hungry starts becoming impatient. Nothing has gone wrong; nobody has told them anything. Yet, with no information to occupy them, guests start doing arithmetic on their own wait: “the drink should have come earlier because there is nothing to cook in it, and it takes less time compared to food, but it took so long for my drink to come” (drinks moment). That is not a complaint about service speed. It is someone with nothing else to think about, reasoning about why a drink that requires no cooking took as long as it did. Left unfilled, the drinks moment gives guests time to build a case against the wait.

What makes this an opportunity rather than a problem is that the same moment demonstrably can carry emotion when something happens in it. Two things in particular register as deposits: generosity and staff contact.

1. “Bottomless drinks feel like they are free even though they are not. You feel gratified” (drinks moment);
2. “The staff were warm and engaging when they brought them” (drinks moment).

Both of those describe the moment doing emotional work, and neither is currently reliable or systematic. Also, structurally, drinks are the most available moment in the entire visit—the guest is seated, stationary, holding something, with a member of staff already at the table. It is the only natural pause in the first 30 minutes, and at present it is doing less than it could.

Copyright © **Esomar** 2026


### *The Wait is Where the Visit is Decided*

The seated moment is the one most operators treat as dead time and, in our data, it is the moment with the strongest and most consistent relationship to how guests feel about the brand after they leave, ahead of the food itself. Again, the qualitative evidence shows this is not about speed.

1. "You keep waiting, and we feel bad for asking questions with a massive queue" (seated moment);
2. "It shouldn't be that you enter a restaurant and get a seat, but then you have to wait for someone to come and give you a list of items that you can order from" (seated moment).

Neither guest is angry, and neither is describing a failure. They are uncertain and slightly uncomfortable about being uncertain—the first is reluctant to ask, the second is waiting for something to begin. Both are withdrawals made entirely of absence. The reverse case is equally instructive. The same wait, with acknowledgement, becomes a deposit: "they came and followed up with glasses, waters, as soon as we sat" (seated moment). This is an information problem, not a throughput problem. Consistent with that, the most commonly given reason for feeling good at this moment is feeling settled and looking forward to the food: an anticipated wait is pleasant, an unexplained one is not. It also explains why drinks and seated are best read together; the drinks moment is the structural opportunity and the uncertainty created while seated is the thing worth spending it on.

### *What Brand Love Is, and What Feeds It*

Brand love was measured in the post-visit instrument, as the mean of 14 attitude statements. The construct draws on the framework established by Batra, Ahuvia and Bagozzi (2012), whose higher-order model identifies various core elements of brand love, among them self-brand integration, passion-driven behaviours, positive emotional connection, long-term relationship and anticipated separation distress. We adapted rather than replicated their published scale for two practical reasons. First, the post-visit survey had to be completed in two to three minutes, in a restaurant, by a guest who had just eaten; a full academic battery was not viable in that setting. Second, we needed statements a brand could act on, so the battery deliberately mixes emotional statements with the concrete experiential attributes a restaurant actually controls. The 14 statements are below, listed in descending order of their strength as drivers of the overall brand love score.

| The 14 brand love statements, strongest driver first |
|---|
| The chicken served at Nando's is fresh. |
| Nando's offers a pleasant dining atmosphere. |
| The chicken served at Nando's is of a high quality. |
| Nando's makes eating out feel more enjoyable |
| The other food (non-chicken items) served at Nando's is of a high quality. |
| Nando's is a brand I would confidently recommend to others. |
| The other food (non-chicken items) served at Nando's is fresh. |
| I see myself coming back to Nando's many times in the future. |
| I feel like Nando's is a brand for people like me. |
| I feel a positive emotional connection to Nando's. |
| Nando's is one of my favourite restaurant brands. |
| Nando's is a brand I am truly passionate about. |

Copyright © **Esomar** 2026



| The 14 brand love statements, strongest driver first |
| --- |
| I would feel a real sense of loss if Nando's was no longer around. |
| Nando's is a brand I find myself thinking about or craving even when I'm not there. |

*Table 2*

Read as a set, these statements are the residue of the visit—i.e., the moments are where the experience is created and the statements are where it is stored. The composite brand love score is 4.24 out of five overall. It correlates 0.63 with revisit intent and 0.59 with overall experience, which is what makes it a commercial metric rather than a soft one.

## *Three Markets, Three Relationships*

| Metric | W. Europe | S. Africa | S. Asia | Total |
| --- | --- | --- | --- | --- |
| **Favourite fast-food brand = Nando's** | 40% | 48% | 20% | 36% |
| **Brand love score (/5)** | 4.19 | 4.58 | 3.95 | 4.24 |
| **First bite mean feeling (/5)** | 4.73 | 4.68 | 4.33 | 4.58 |
| **Overall experience, top-2-box** | 90% | 90% | 80% | 87% |
| **Revisit intent, top-2-box** | 70% | 95% | 68% | 78% |

*Table 3*

The three markets are not at different levels of success on one scale; they are at different stages of the same relationship. Southern Africa has effectively already won, with an attachment rooted in shared, social dining: “Nando's is literally in our blood here. It's the taste of home, the humour we all share and the place where you always feel like you belong the minute you walk through the door” (Southern Africa). Its spontaneous standout is staff friendliness, volunteered by 43% (more than double any other market), and its journey shape is distinctive—order placed rivals first bite as its best moment (40% versus 17% overall), the inverse of the overall pattern. Western Europe is a mature relationship with a ceiling: strong on fundamentals, more moderate on the deeper emotional statements and the only market where speed of service is spontaneously top-of-mind (35%): “I've literally grown up with Nando's; it's a taste that is deeply familiar to me. If Nando's disappeared, it would feel like losing a comfortable, reliable piece of my daily routine” (Western Europe).

South Asia is the most instructive case, and the one that most needs a hypothesis rather than a description. It reports the most positive in-moment feelings of the three markets: highest on comfortable (63%), relaxed (53%) and excited (45%), yet sits lowest on the attitudinal measures. In-moment enjoyment is clearly there, but durable attachment has not yet formed behind it. Our reading is that the two are measuring different things, and that the gap is explained by how the visit sits in the guest's life. The South Asian occasion is typically one stop inside a wider day out rather than a destination in its own right, and the relationship is younger, so the visit is enjoyable in the way an outing is enjoyable, without yet having become the repeated, social ritual that produces attachment in Southern Africa. Consistent with that, first bite is the one moment that does not stand out as the emotional peak in this market as it does in the other two, which points to an opportunity to make more of that moment rather than to any weakness in the visit itself.

Copyright © **Esomar** 2026



## From Insight to Action: What the Model Selected

Eight reviewed options per moment were tested against the "moment twin" for each market. The strongest two moments per market were then combined into a single recommended action.

| Market and action | Brand love |
|---|---|
| **Western Europe: fix the wait, then let the aroma lead** | 4.19 → 4.74 (+0.55) |
| **Southern Africa: proactive welcome into food ceremony** | 4.58 → 4.74 (+0.16) |
| **South Asia: unlock the wait, then frame the bite** | 3.95 → 4.45 (+0.50) |

*Table 4*

In Western Europe the selected combination is a table card carrying an honest wait estimate plus a staff check-in within two minutes, followed at food delivery by a line that primes the sensory peak: "You're going to smell this before you taste it…that's the peri-peri". Western Europe is the only market where aroma leads the first bite, so that cue is market-bound. In Southern Africa, where there is least headroom, the model selected a designated host per shift ("come sit down, we've got you") and food placed as ceremony ("there it is, proper peri-peri, made for you"). The smaller predicted movement reflects proximity to ceiling, not a weaker idea, so the objective there is protection rather than growth. In South Asia the combination addresses the wait and then gives the first bite the framing it currently lacks: a staff check-in within two minutes ("Your order's in, shouldn't be long") and, at delivery, an explicit freshness cue ("Our chicken is grilled fresh for every order, never reheated"), reinforced on the tray liner. Our reading is that where the brand is long established this credential is largely assumed, whereas in a younger relationship stating it out loud at the moment the food lands is what gives that moment its emotional weight.

The most significant result is what these options have in common. Across all three markets, and across every moment tested, the changes that moved brand love most were not product changes, not speed improvements and not giveaways. They were acts of information and acknowledgement: i.e., things staff say. This is a useful corrective to intuition. One tested drinks-moment option—bringing three small peri-peri sauce pots with the line "try these while you wait"—is generous, tactile and strongly on-brand. However, it was not selected in any market. A spoken cue that resets wait expectations outperformed it, at no cost, no stock and no operational change.

## Implications for Nando's

Four implications follow from the evidence:

1. Protect the welcome and the counter. These staff-led interactions set the emotional tone of the visit, and in Southern Africa the counter experience is already a competitive strength worth exporting.
2. Close the information gap while guests are seated. Uncertainty about wait time has a disproportionate effect and is the cheapest thing on this list to address.
3. Activate the drinks moment, which carries no emotional load today despite being the most structurally available point in the visit.
4. Frame the first bite explicitly in markets where the brand is still building familiarity, so that the food's credentials are stated rather than assumed.

Market strategy should follow relationship stage rather than a single global playbook: protect and export in Southern Africa, deepen emotional bonds in Western Europe and convert in-moment enjoyment into durable attachment in South Asia.

Copyright © **Esomar** 2026


## Implications Beyond Nando's

The transferable finding is not just about restaurants; it is that brands systematically invest in the moment they designed, while the relationship is decided somewhere they have not looked, typically a moment in which the customer is waiting, uninformed and not in control. The same structural pattern recurs across categories with sequential, physical or procedural journeys:

- Airlines invest in the cabin, and the trip is frequently remembered at baggage reclaim where the passenger has no information and no control;
- Private healthcare invests in the consultation, and the experience is often decided in the waiting room where the patient does not know how far behind the clinic is running or whether it is acceptable to ask;
- Automotive retail invests in the showroom and the test drive, and the relationship is frequently decided at the first service visit when something has already gone wrong;
- Financial services invest in the branch or the app interface, and onboarding and claims (the low-control waiting moments) are where trust is made or lost.

In each case, the deciding moment shares three properties: it is low in customer control, low in information and low in design attention. That combination is diagnosable in advance. The method transfers as three steps rather than as a fixed instrument:

1. Measure in-the-moment rather than just in recall, because recall compresses a sequence into a single impression;
2. Look for emotional transitions rather than just operational touchpoints, because a touchpoint that changes nothing emotionally is forgettable even when executed perfectly, whereas the moments that change a customer's state are the ones retained;
3. Model changes on a twin before testing them on real customers, which allows a brand to rank and discard options cheaply and reserve live testing for the few that survive.

The design constraints transfer, too. The approach needs a journey that decomposes into discrete, sequential milestones (a capture moment that does not itself damage the experience) and enough base to see which milestones behave consistently before qualitative resources are committed. Categories with long, asynchronous or largely digital journeys would need a different capture mechanic, though the underlying logic (locate the un-owned moment, measure it live, model the fix) still holds.

## Reliability And Limitations

This was a pilot, so its purpose was to establish whether in-moment emotional capture, combined with same-guest qualitative work and predictive modelling, produces findings a brand can act on. The following should be read with that intent in mind.

- Base size: n=120, at 40 per market, is a pilot base. It is sufficient to show consistent patterns across moments and to identify clear differences between markets, but small differences should be treated as indicative.
- Regional labels: To protect country-specific data and results, we have referred to regions rather than individual countries throughout.
- How the moment rankings should be read: When we say one moment matters more than another, that comes from a statistical model estimating how much each moment contributes to brand love. On a pilot base, that model is reliable for ranking the moments—telling us that seated matters more than arrival, for example, but the exact size of each contribution carries uncertainty.
- Correlations: The associations between brand love and revisit intent (0.63) and overall experience (0.59) are strong for attitudinal data, but they are associations, not causal proof. All three measures sit in the same post-visit instrument, so a degree of common-method variance cannot be excluded.
- Model predictions: The predicted brand love movements are counterfactual estimates of what the model expects if an option became standard practice. They have not been validated by an in-market experiment and, on a pilot base, they should be read as a way of

ranking opportunities and sensing their relative scale, not as forecasts. The model's reasoning is also bounded by its training data: it can only evaluate moments and levers observed in this study.

- Participation effects: Guests knew they were taking part in research during the visit, which may have raised their attention to moments they would ordinarily pass over. This is a known trade-off of in-moment capture.

A natural next step would be to run one of the selected options as a live in-market test and compare observed movement against the model's estimate, which would both validate the moment twin and extend the pilot into something scalable.

## Conclusion

The food earns the trip. The welcome, the wait and the words earn the attachment. Across three culturally distinct markets, Nando's guests describe the brand in the language of belonging. No moment in the visit is failing; 43% could not name a least-enjoyable moment at all. However, the moments are not equal and they are not interchangeable: the first bite is what guests love the brand for, the wait is where that love is put at risk and the drinks moment is the one nobody is using. The most actionable finding is also the least expensive. When a predictive model trained on this journey was asked what would move brand love most, in every market it selected a spoken cue; a few words from a member of staff. It comes down to information and acknowledgement, delivered at the moment a guest is least certain and least in control. For the wider industry, the implication is a question worth asking of any customer journey: which moment have we never looked at, because nobody owns it? On this evidence, that is where the relationship is being decided.

## References

Batra, R., Ahuvia, A. and Bagozzi, R.P. (2012). Brand Love. *Journal of Marketing, 76*(2), 1–16. doi:10.1509/jm.09.0339.

Borderless Access and Nando's. (2026). *Decoding the dine-in micro-moments that build Brand Love for Nando's. Full research report, n=120.* Unpublished internal PowerPoint report.

Csikszentmihalyi, M. and Larson, R. (1987). Validity and Reliability of the Experience-Sampling Method. *Journal of Nervous and Mental Disease, 175*(9), 526–536. doi:10.1097/00005053-198709000-00004.

Kahneman, D., Fredrickson, B.L., Schreiber, C.A. and Redelmeier, D.A. (1993). When More Pain Is Preferred to Less: Adding a Better End. *Psychological Science, 4*(6), 401–405. doi:10.1111/j.1467-9280.1993.tb00589.x.

Redelmeier, D.A. and Kahneman, D. (1996). Patients' memories of painful medical treatments: real-time and retrospective evaluations of two minimally invasive procedures. *Pain, 66*(1), 3–8. doi:10.1016/0304-3959(96)02994-6.

## About the Authors

Anneri Venter, Research Director UK and Europe, Borderless Access, South Africa.

Simone Kakana, Product and Insights Manager, Nando's, South Africa.

| Text | | | |
|---|---|---|---|
| **Text** | Text | | |
| | | | |
| | | | |
| | | | |

*Table captions*

# Deep Breath

## Uncovering the Missing Dimension in Consumer Emotions

*Graham Page*

Esomar

Office address:

Burgemeester Stramanweg 105

1101 AA, Amsterdam

Phone: +31 20 664 21 41

Email: info@Esomar.org

Website: www.Esomar.org

Publication Date: September 2026

Esomar Publication Series Volume S424 Congress 2026

ISBN 92-831-0337-8

Copyright © **Esomar** 2026

Copyright © **Esomar** 2026

Congress 2026

# Deep Breath

## Uncovering the Missing Dimension in Consumer Emotions

*Graham Page*

### Introduction

Innovation in market research in recent years has become almost synonymous with one thing: the more efficient generation of data and findings. The templating of research protocols has minimised the number of design decisions required; the rise of automated research platforms has cut increasing numbers of human steps out of the process of data collection and tabulation; and now the adoption of large language models (LLMs) and other forms of Generative AI technology has greatly accelerated the generation of research summaries and reporting, as well as brought into question the need for real respondents. However, as data has become faster and cheaper to generate, paradoxically the ability to generate insight, particularly about consumer motivation, has become more difficult. By cutting people out of the interview and analysis process, it is harder for research professionals to go beyond what is said, to see what is really meant. LLMs, by definition, take what people say at face value and can't see beyond the literal statements to get at the unspoken meaning. This is particularly problematic when it comes to what arguably is the core of the insights industry's job, which is to understand what really motivates people to buy, or watch, or vote. It is not enough to understand what people do and who does it, or even why they say they do it. We need to understand how people feel, or how they expect certain purchases or behaviours to make them feel, as emotion is a fundamental driver of behaviour that many consumers fail to realise is at work in their choices.

Researchers have long used observational methods that allow them to go beyond what people say, such as tone of voice, body language and facial analysis. However, this has become increasingly challenging as global research has shifted online, as the human steps have been progressively eliminated from the interview process. In this context, researchers (and now, LLMs) know what people said, but how do they know what they really meant? In response, many researchers in the last decade or so have turned to automated analysis of participant video, using algorithmic tools to assess facial expressions, body posture and verbal sentiment. Nonetheless, these tools only measure certain aspects of people's emotional experience, as there are only so many emotions that play across the face. In particular, it has been difficult to measure emotional arousal (i.e., tension, or excitement, or anxiety) from remote participant video or stated sentiment. The tension in the build up to a dramatic moment in an ad, or the frustration at a poor online experience, can be missed solely by looking at expressions, or after the fact recollections. This is a key issue, as arousal is one of the fundamental dimensions of emotion that is studied in academic investigation, and is a key component of many well-known models such as Russell's 1980 Circumplex model of emotion, or the Vector model of emotion (Bradley et al., 1992; Watson & Tellegen, 1985; see Figure 1).

Copyright © **Esomar** 2026

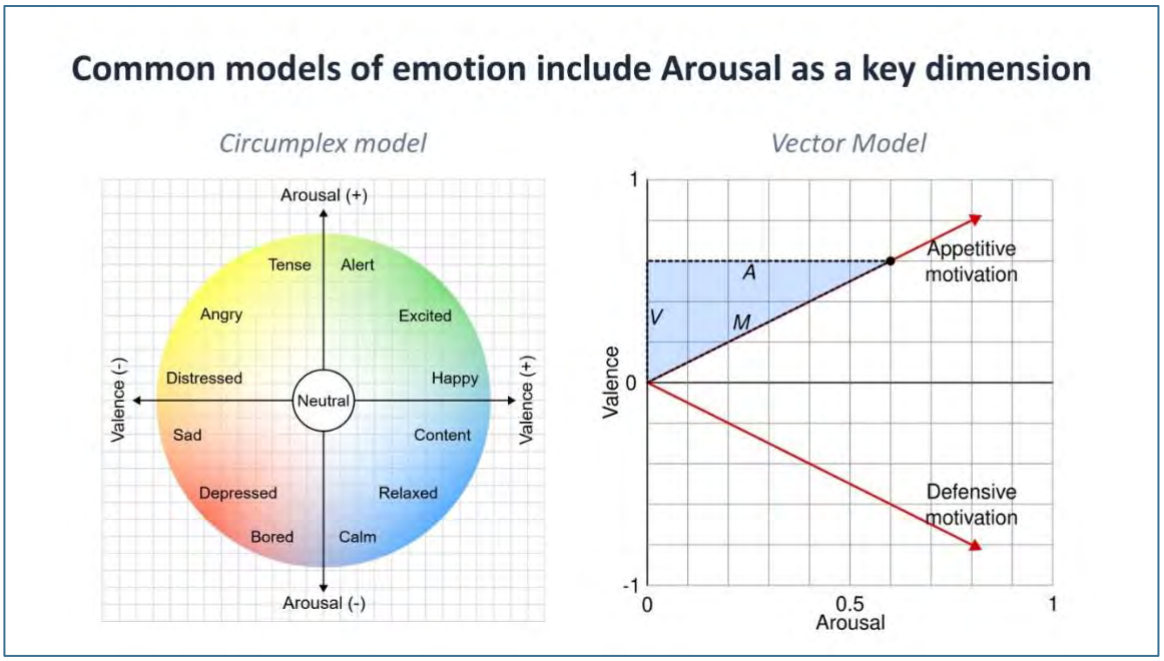




*Figure 1. Russell's Circumplex model of emotion (1980), and the Vector model (e.g., Bradley et al., 1992). Emotions are characterised along two fundamental dimensions: valence (positive or negative affect) and physiological arousal (intensity).*

In academic settings, arousal is measured via physiological signals which indicate the state of the nervous system such as Galvanic Skin Response (GSR, the electrical conductivity of the skin), heart rate or respiration rate. This presents obvious challenges when applied to large scale market research, as such methods require electrodes or devices to be applied to participants' bodies. While some small-scale research panels exist for this work (such as those offered by Mediaprobe) these still require participants to apply a sensor or hold a device during fieldwork, and as such these approaches are limited in how quickly they can be applied, or the scale and speed at which research can be conducted. However, there are additional signals available in remotely-collected participant video that can provide some of this missing insight. Our paper details how iMotions has developed a groundbreaking approach to measuring emotional arousal via the webcam-based measurement of respiration rate, which unlocks this additional dimension of emotional response for use at much larger scale by the industry.

## Video–Based Measurement of Respiration

iMotions has built software over many years, which allows the signals from different types of bio-sensors to be brought together and analysed in one platform. As such, we have lots of experience in both direct bio-signal measurement and measurement of signals from respondent video (historically for facial expression detection). iMotions' Affectiva facial expression technology has been used for almost 15 years by advertisers worldwide to assess audience's emotional responses to ads, entertainment content and experiences. In this sense, this represents a huge corpus of data from which to assess and develop additional signals. Reviewing such data, it is apparent that there are subtle signals in videos of research participants that can signal more than just facial expression. Changes in skin tone can indicate heart rate, and body movements can provide small but meaningful indications of respiration rate. In particular, it was noted that subtle changes in the shoulders and upper torso have the potential to indicate when viewers breathe in and out. Based on this insight, iMotions developed a machine-learnt algorithm, which isolates such body movements and, from this, established a continuous measurement of breathing rate and its variability (see Figure 2). Given the higher potential for noise in a participant video compared to a direct bio-sensor, quality filters were also developed to isolate the signals related specifically to breathing, and to identify cases where the signal does not conform to a normal breathing pattern.

Copyright © **Esomar** 2026

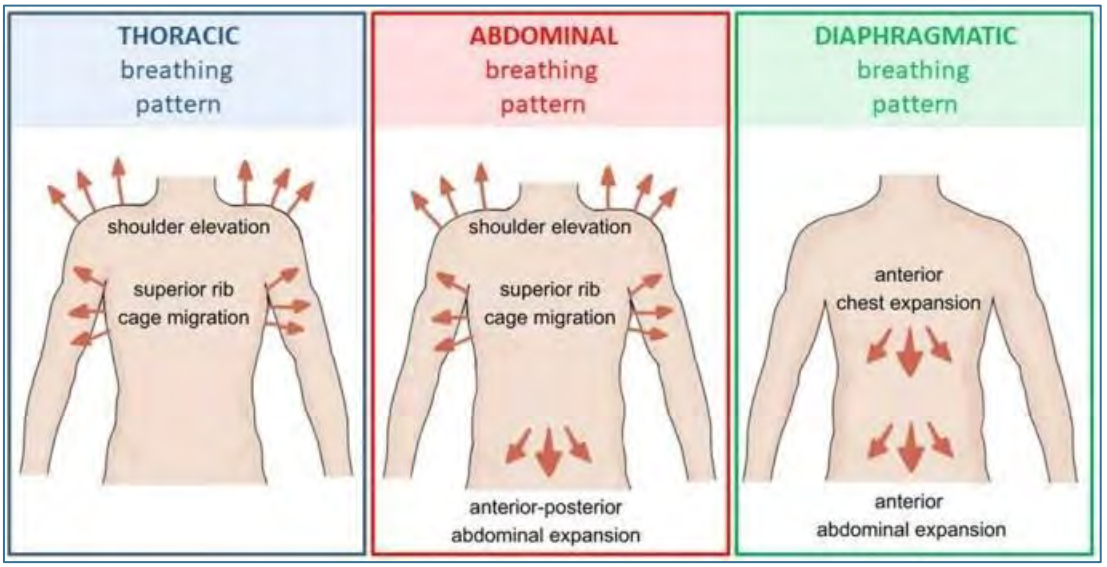




*Figure 2. Body signals that indicate breathing actions (Sikora et al., 2024).*

## Validation

As a vendor of biometric software and hardware, iMotions has unique access to lab-standard technology with which to validate the algorithm. Therefore, two validation studies were conducted. Firstly, to compare the signal from participant video to that derived from gold standard technology, in this case the Biopac respiration belt, which directly measures the movement of the ribcage via a physical sensor; secondly to establish whether good data could be collected from research panellists in the wild (Gregoret, Seernani & Wilson, 2024). In the first experiment, 10 participants were recruited and asked to watch a number of videos with varying emotional content—one baseline of neutral stimulus and then a series of funny or emotional ads, as well a positive or disgusting images. In each case, responses were recorded via webcam and respiration rate and number of cycles was also measured via a Biopac belt. When comparing the results, despite the small number of participants, highly significant correlations ($r>0.8$) were found between the respiration metrics (respiration rate and number of breathing cycles) from the webcam method and the Biopac belt (for example, see Figure 3), strongly suggesting that the webcam-based method can be used as an accuracy proxy for direct measurement of respiration.

Copyright © **Esomar** 2026



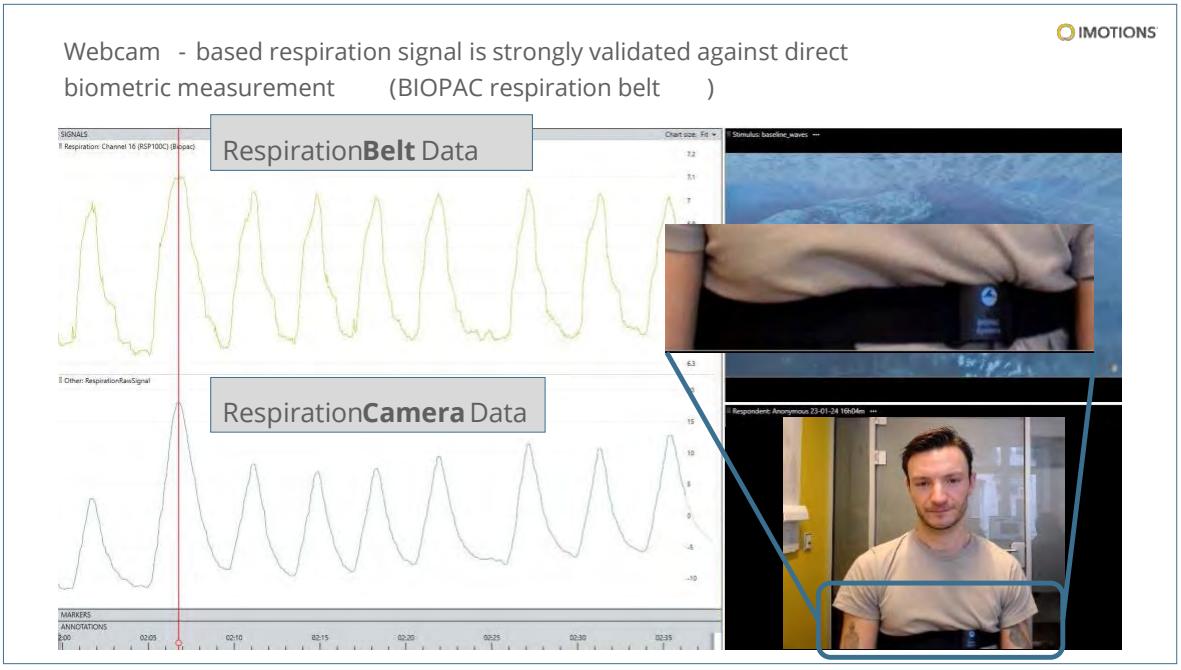



*Figure 3. Comparison of respiration rate data from one participant, collected via webcam and Biopac sensor.*

The second experiment focused on the scalability of the method: can data be collected effectively from typical research panellists? To investigate, a larger scale study was completed using research participants recruited from the Prolific panel. Again, participants were shown a variety of videos, and their responses were recorded (with consent) via webcam with respiration metrics calculated based on those recordings. Of 329 qualified participants who reached the pint in the interview where they were asked for consent to be filmed, 77% of them consented to be recorded and continued through the positioning check to be exposed to the video and 205 completed the study in total. This opt-in and completion rate compares well with other comparable statistics obtained, for instance when asking participants for consent to be recorded for facial expression analysis. Equally importantly, the webcam respiration also successfully distinguished between the different types of videos shown to participants. The stimuli consisted of a neutral baseline video, two humorous ads and videos of positive and disgusting images taken from the *International Affective Picture System*. Significant differences in respiration rate were observed, both between the neutral baseline video and the positive and humorous content (higher breathing rate observed) and between the positive stimuli and disgusting images (lower breathing rates observed for negative stimuli). This suggested that not only can the webcam algorithm successfully measure breathing rate, but this can also be achieved at scale with research panellists and there is meaning in the resulting measurement of breathing rate that can successfully differentiate emotional responses to visual stimuli, such as ads or images.

As a qualitative example, Figure 4 shows the average respiration rate over time for a sample of participants exposed to four videos. The first and third video are neutral baselines, showing calming undersea imagery (video 1) and nature scenes (video 3). Video 2 is first-person user-generated content of parkour (specifically, a woman walking along the edge of a high-rise building), and video 4 is footage of a baby laughing. Both video 2 and 4 would be expected to be more arousing, and we do see clear evidence of higher respiration rate to those videos. The response to video 4 is also accompanied by clear positive facial expressions (smiles) but, crucially, the response to video 2 is not. This is a more anxiety-inducing video, and while there are small bumps in facial expressiveness there is a much clearer response in breathing rate, suggesting that, as expected, this approach can pick up emotional responses that do not show on the face.

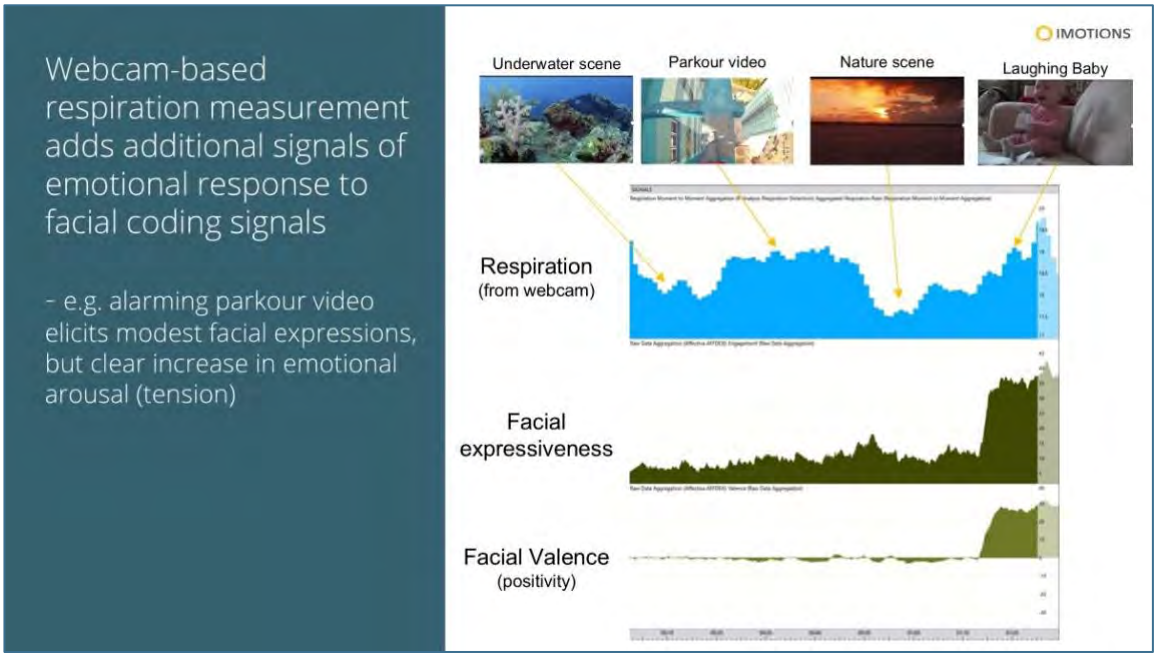



*Figure 4. Webcam–based respiration and facial measurements in response to four videos. Anxiety-inducing parkour video generates clear respiration (arousal) change, but limited facial expression change.*

## Implementation and Data Quality Safeguards

Given the successful validation of the method, the webcam respiration algorithm has been implemented within the iMotions suite of tools, specifically within:

- iMotions online research: a DIY research platform that allows commercial researchers to set up and run online survey projects, which also include webcam-based eye-tracking, facial expression analysis and now respiration measurement;
- The Affectiva media analytics platform: an enterprise-level research platform that allows the integration of webcam-based audience response measurement into ad testing and entertainment testing research tools (extensively used by the ad industry, with over 100k ads tested to date in the database in over 90 countries).

This has opened up the use of the approach for a variety of applications, both in video-testing but also in wider concept, UX and NPD work. Given that most applications in online research are within self-serve or templated products, where speed is of the essence, human intervention is kept to a minimum and where individual respondent videos are unlikely to be reviewed manually, the algorithms also have additional quality safeguards implemented as part of the process. At the data collection stage, participants are given clear guidance in how to sit, and are given a framing oval to confirm that they are visible to the camera as well as, at the analysis stage, additional data cleaning is carried out. For example, data which provides respiration rates outside the typical human boundary of seven to 30 breaths per minute is discarded, as is data which does not confirm to the expected waveform for respiration (this is a threshold which can be adjusted by the user in iMotions Online). Likewise, data which exhibits large swings in breathing rate or minimal movements is also excluded, as this is typically a signal of an artefact produced by participants moving around or moving their shoulders out of shot. This ensures that even when data is collected on mobile devices, only valid data is included in the analysis and motion artefacts are kept to a minimum.

## Insights Created

While technically impressive, the value of the respiration metric is only apparent in the additional insights that it can bring. Since implementation, the tool has been applied widely across projects run using the iMotions online platforms, and this has revealed many cases where new insights

have been uncovered. While client confidentiality limits our ability to share certain examples, the following cases give a flavour of the insights this approach enables.

### *Uncovering Excitement*

Understanding emotional responses is a critical element of movie trailer research, and iMotions has researched several thousand trailers and episodes. They often have highly complex and emotional storylines, and so widening the palette of emotional measurement is of clear benefit. Some years ago, we researched the trailer for the last James Bond movie *No Time to Die*, which includes the usual mix of action, drama, humour and stunts. Viewers love the trailer, and show very clear facial reactions, but there were always a couple of moments which produced surprisingly little facial expression. Testing the trailer again and using the respiration algorithm added clear colour to these moments. Figures 5 and 6 show the respiration rate of viewers and also some key facial signals, e.g., "Valence" or overall positivity and "brow furrow", often indicating tension in the context of trailers. During a key action and fight scene, facial expressions were slightly negative, but there was a clear spike in respiration rate, suggesting the action was indeed engaging viewer as we might expect. Similarly, later in the trailer there is a further action sequence including a dramatic motorcycle stunt, and while there was relatively little facial reaction to this, there was a clear increase in respiration rate. Therefore, as we might hope, measuring respiration can give a clear signal about emotional engagement in the absence of facial responses.

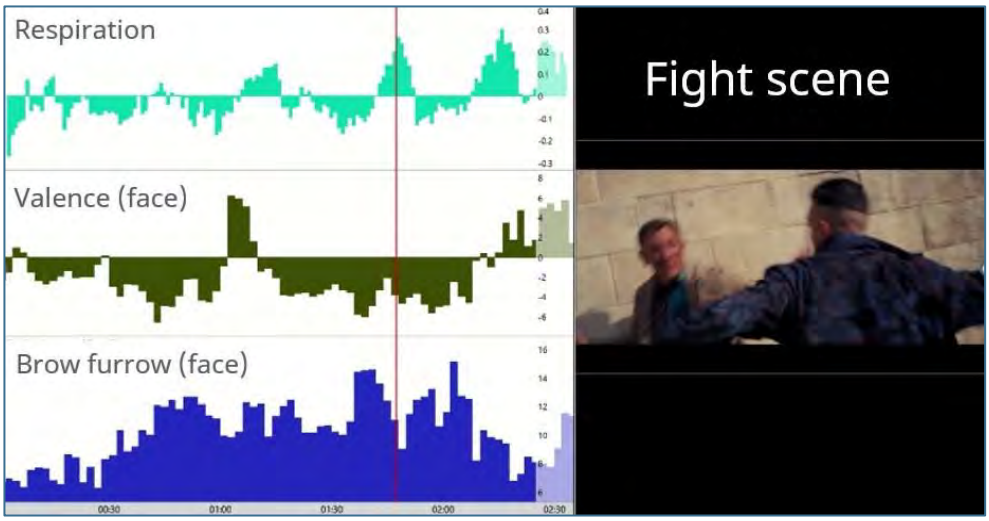



*Figure 5. Respiration and facial responses or fight sequence in James Bond trailer (key moment in red on the timeline).*

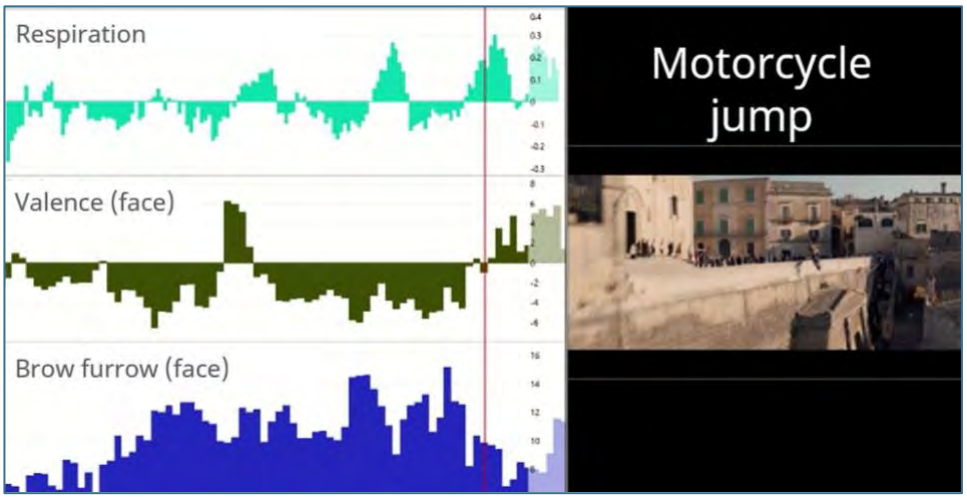



*Figure 6. Respiration and facial responses to stunt and action sequence in James Bond trailer (key moment in red on the timeline).*

## *Gasp Moments*

Respiration does not have a linear and positive relationship with engagement. Generally speaking, faster breathing is associated with excitement or tension, but there are other characteristics associated with additional responses. For instance, holding one's breath is part of the biological startle response, and can be an indicator of shock. An ad we recently tested for Columbia outdoor clothing illustrates this perfectly. It is a very intense commercial, with lots of fast cuts, threatening and occasionally violent and graphic imagery designed to convey how harsh the outdoors can be (and of course, why you therefore need Columbia clothing). Early parts of the ad, where the threat and intensity build, are associated with a clear increase in breathing rate. However, then at a key moment (see Figure 7), we see a sharp drop in respiration accompanied by a mouth open response—an almost perfect example of participants gasping slightly. The reason? At this moment, the ad recreated the moment where Aaron Ralston (of *127 Hours* and "I cut my own arm off to escape being trapped" fame) is trapped by a rock (again). Respiration measurement gives real clarity to the emotional power of this moment—inevitably the most memorable scene in a very intense spot.

*Figure 7. Gasp moment (marked in red) at shocking scene in Columbia ad.*

## *Converging Signals: Effort and Difficulty*

While communication and content testing are always fertile ground for bio-sensing technology, there are clearly other areas where measurement of arousal is important. In a recent study, we were able to demonstrate a clear interaction between advertising exposure and subsequent effort in a brand-location task at a shelf front. All participants were shown a virtual shelf and asked to locate a specific brand (CeraVe), but half the participants were shown an ad for the brand first, whereas the other half saw the ad after the shelf task. Participants were recorded using their webcam, and eye-tracking facial analysis and respiration measurement were conducted on the recordings. The hypothesis was that exposure to the ad would increase the saliency of the brand and packaging and make the search task easier, which is indeed what we found. Participants who saw the ad first showed greater visual attention on the target brand during the task than those who did not. However, there was also evidence that they found the task less onerous or stressful, as respiration rate was lower during the task amongst those exposed to the ad first and they also showed lower levels of brow furrow, which is often associated with effort in this scenario (see Figure 8). In this case, both the expression and respiration data converged to indicate lower effort during the task, and the improved attention to the brand showed a positive outcome of this lower effort. This both illustrates the potentially subtle but important role of point-of-sale advertising, and shows how respiration and other video signals can tell us something about how much effort we are requiring of consumers at retail.

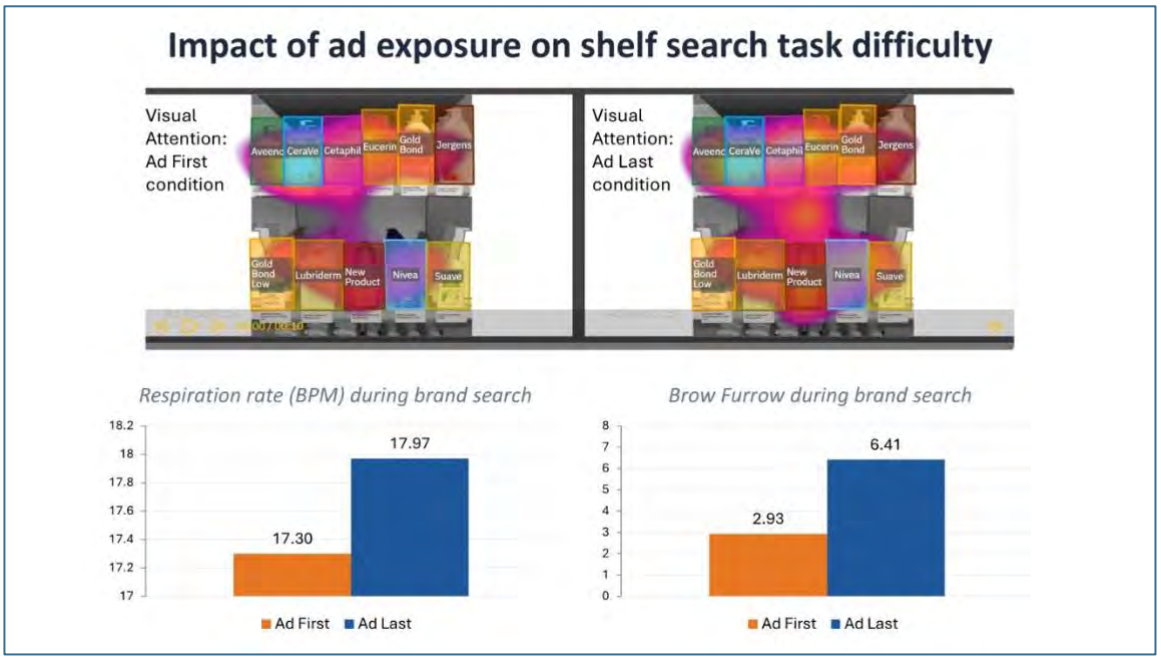



*Figure 8. Eye-tracking, facial expression data (brow furrow) and respiration rate data during brand search shelf task (target brand: CeraVe). Participants split into two groups, Ad First (shown brand ad before task) and Ad Second (saw brand ad after the task).*

## Applications

This novel approach to bio-signal measurement has many applications, as already suggested in the case studies above, and has already begun to be used at scale by many iMotions clients. This respiration signal can be applied whenever there is a need to go beyond what is said in an interview or survey, to dig deeper into consumer reactions that participants may be unaware of, unable to remember or feel reluctant to discuss. It has real potential to add a further dimension to the measurement of consumer reactions "in the moment", without requiring them to self-monitor and interrupt or distort their experience. By simply requiring a webcam and viewer consent, this can be applied alongside survey instruments as well as alongside existing currencies of direct measurement, such as facial expression analysis. Thus far, we have seen it to be particularly useful in:

- Advertising and brand content research: adding understanding of the emotional flow of video narrative, whether viewers are truly engaged in key moments and how this leads to stronger or weaker ad performance.
- Creator content research: this is particularly relevant for content which may be considered more edgy or contentious, and over which brand owners may need more objective measures of responses than available through online metrics of likes, clicks and shares.
- Entertainment content testing: the addition of direct arousal measurement is a significant step forward given the need to keep viewers engaged over long periods, and the wide emotional palette that different genres draw from. Tension and immersion are key elements that are hard to measure from dial-testing or recall, which take viewers out of the experience. This is a potentially powerful alternative to GSR measurement, which is difficult to scale beyond small research panels.
- Qualitative research: Alongside facial and voice signals, respiration measurement offers another signal to assess participants' non-verbal signals, which is increasingly difficult to acquire in online fieldwork. It also offers a further potential signal to tune the responses of Generative-AI moderators, and may allow them to go beyond the face value of participant speech and identify alternative meanings, or when to dig deeper with further prompting.
- User experience: Anticipating when users become frustrated with functionality is a critical element of UX research, and webcam respiration offers a potential means to spot rising tension and frustration. This adds fidelity to user recordings and may help pinpoint triggers of frustration that occurs before users seem "blow up" or abandon tasks. It may also more clearly identify moments of surprise (and combined with facial expressions when these are good or bad!).

- Retail: As illustrated above, identifying effort and pain points is critical in digital retail research, and respiration measurements is a means to identify when decision-making flows or where it sticks and purchase are potentially lost.

## Summary

*This is not the AI you were looking for…*

… But it may be the AI that you need. Not all technological innovations in research involve Generative AI. While LLMs are gaining the most attention, this may mean researchers are missing other novel approaches which add depth to insights, rather than just efficiency. In an era where the human role in the research process is being compressed, video-based respiration measurement can add back some of the signals that automated tools, which focus only on words, may miss. As ever, this is not a replacement for other tools, but it adds an additional modality, which taken with other signals such as facial expression or tone of voice and alongside stated answers and surveys, can provide a more rounded view of human reactions. It is an easily implemented, bio-based signal that can help humans and machines go beyond what is said, to what is felt and get to the truth of consumer experience.

## References

Bradley, M. M., Greenwald, M. K., Petry, M.C., and Lang, P. J. (1992). Remembering pictures: Pleasure and arousal in memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18*(2): 379–390.

Gregoret, L., Seernani, D., and Wilson, J.M. (2024). *In-Lab and Remote webcam-based Respiration: A promising candidate for neuromarketing?* NeuroPsychoEconomics Conference, May 2024.

Russell, J. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology, 39* (6): 1161–1178.

Sikora, M., Mikołajczyk, R., Łakomy, O., Karpiński, J., Zebrowska, A., Korstorz-Nosal, S., and Jastrzębski, D. (2024). Influence of the breathing pattern on the pulmonary function of endurance-trained athletes. *Nature: Scientific Reports, 14*(1113). Available: https://www.nature.com/articles/s41598-024-51758-5

Watson, D., and Tellegen, A. (1985). Toward a consensual structure of mood. *Psychological Bulletin, 98*(2): 219–235.

## About the Author

Graham Page, EVP, Commercial Research & Innovation, iMotions, Oxford, UK.

esomar

# Marketing's Most Expensive Oversight

## Teaching Researchers and Marketeers How to Engage with the Neurodiverse Community

*Alexander Holmes*

*Jessica Garlick*

*Heidi Kim*

*Sebastian Giugovaz*

Esomar

Office address:

Burgemeester Stramanweg 105

1101 AA, Amsterdam

Phone: +31 20 664 21 41

Email: info@Esomar.org

Website: www.Esomar.org

Publication Date: September 2026

Esomar Publication Series Volume S424 Congress 2026

ISBN 92-831-0337-8

Copyright © **Esomar** 2026

Copyright © **Esomar** 2026

Congress 2026

# Marketing's Most Expensive Oversight

## Teaching Researchers and Marketeers How to Engage with the Neurodiverse Community

*Alexander Holmes*

*Jessica Garlick*

*Heidi Kim*

*Sebastian Giugovaz*

### Setting the Scene

#### *Neurodiversity: A Growing Audience, Both in Size, Public Awareness and Monetary Value*

Neurodiversity (ND). It's a topic that's had an explosion of interest over the last decade. In part this is due to diagnoses steadily increasing over the last 10 years. Since 2017, one to five people in the US has some form of ND (NCLD, 2017). Based on ascertainment modelling (or diagnostic capture) this figure is set to rise by a minimum of ~1 million new identifications per year, as we move toward the "plateau" where identification finally matches the true biological prevalence. At the same time, our TV screens and social feeds are much more textured with people who wear their ND on their sleeves: think celebrities like Greta Thunberg, Billie Eilish, Richard Branso or characters like Sheldon Cooper or Sherlock. This shift is also changing behaviour: online search interest has surged, with people now searching for neurodiversity-related terms more than five times as often (based on Google search data for "neurodiversity", 2004-2026).

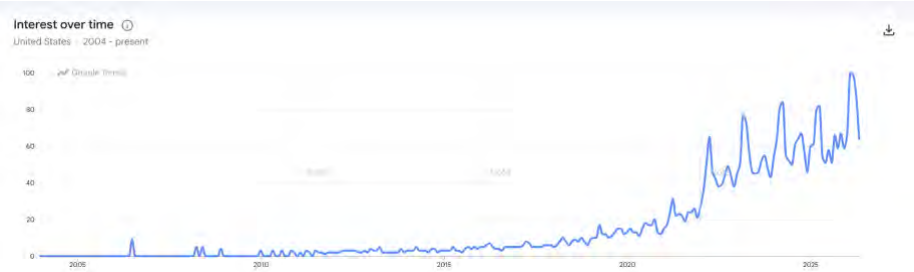



*Figure 1. Google Trends analysis.*

Yet, despite this rapid growth and cultural visibility, the commercial opportunity remains largely overlooked. Our estimates place the audience value at over $6 trillion USD (estimated annual value: 132 million people × ~$47.5K ≈ $6.3T). You'd expect brands to be clamouring to understand who they're engaging, and how to capture more of their spending. You'd be wrong.

#### *Neurodiversity: A Forgotten and Overlooked Audience*

Despite the value and size of this audience, when it comes to marketing research, interest seems to remain at near zero. Conducting a quick scan of relevant research sources about this community, you'll find next to nothing (see Table 1).

| Market research bank | Hits for neurodiversity |
|---|---|
| Archive of Market and Research (AMSR) | 5 |

| | |
|---|---|
| **Esomar** | 1 |
| **Mintel** | 2 |
| **Quirks Media** | 6 |

*Table 1. Results for marketing research "neurodiversity" searches.*

Adding to the irony, smaller consumer segments often benefit from disproportionately greater levels of research. Cross-referencing the same media outlets and/or sources with small cohorts or groups of people, Gen Z (20% of the population), luxury shoppers (8%), daily X users (7%) and plant-based eaters (5%)—the disparity is dramatic (see Table 2).

| **% of population *(decreasing)*** | **30-40%** | **20%** | **8%** | **7%** | **5%** |
|---|---|---|---|---|---|
| **Market research bank coverage** | Hits for neurodiverse people | Gen Z | Hits for luxury shoppers | Hits for plant-based eaters | Hits for X users |
| **Archive of Market and Research (AMSR)** | 5 | 61 | 830 | 145 | 114 |
| **Esomar** | 1 | 98 | 1,079 | 24 | 589 |
| **Mintel** | 2 | 264 | 336 | 444 | 300 |
| **Quirks Media** | 6 | 254 | 217 | 1,018 | 287 |

*Table 2. Results for marketing research "neurodiversity" searches, versus other consumer segment searches.*

The pattern is the same when you look at advertising focused publications.

| **Advertising bank coverage** | **Hits for neurodiversity** | | | | |
|---|---|---|---|---|---|
| **Campaign US** | 28 | 1,316 | 623 | 5,702 | 1,804 |
| **Little Black Book (focused on work)** | 19 | 426 | 383 | 112 | 202 |

*Table 3. Results for advertising research "neurodiversity" searches.*

The disparity is hard to ignore. Smaller, more narrowly defined audiences are studied hundreds, sometimes thousands, of times more frequently.

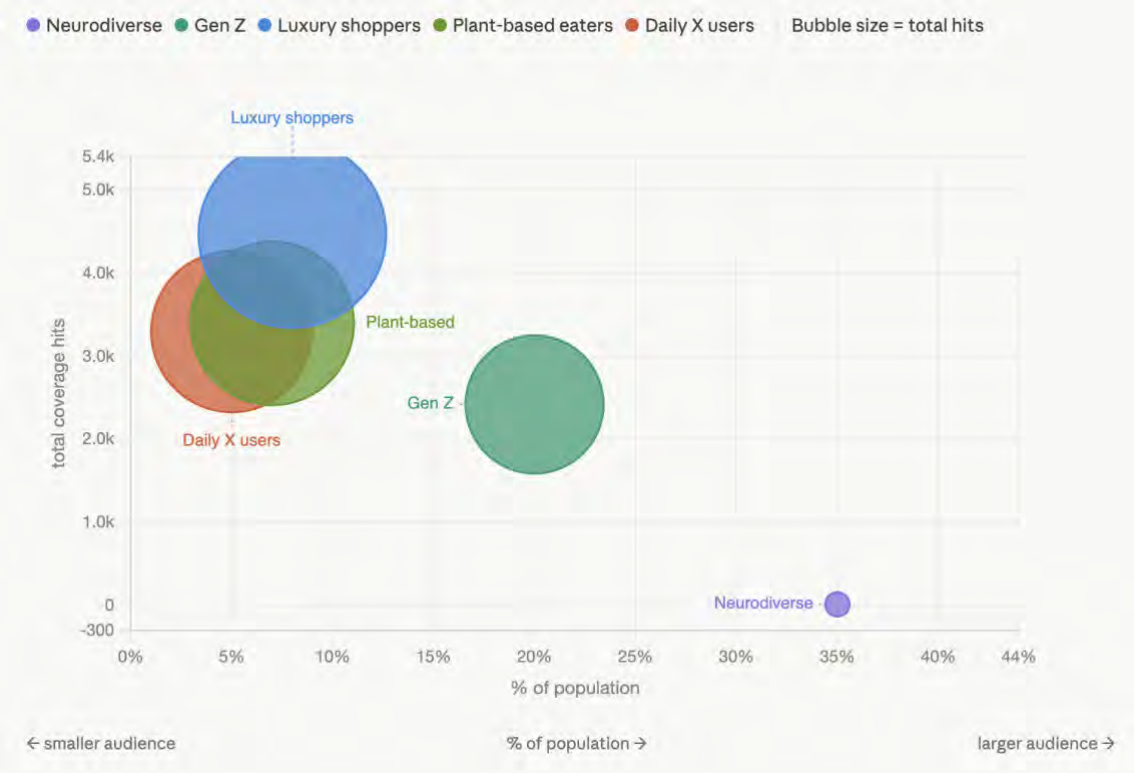



*Figure 2. Data visualisation built using Claude AI.*

This points to a clear and significant research oversight, and it's a surprising one. Because if there is one factor that fundamentally shapes how people experience the world, process information and make decisions, it's their neurology. Which raises an obvious question: if this audience represents trillions in spending power, and if neurology directly influences behaviour, why are brands investing so little in understanding it?

## *Neurodiversity: Often Faced with Negative Misconceptions*

Filling the vacuum of knowledge and inaccurate information, naturally, misconceptions have come to the fore as the wider public in the US (and in some cases, friends and family members) have little to no understanding of ND communities. Below are just some of the daily struggles and barriers those with ND tell us about:

- There is something wrong with us: "I have heard both 'wilful' and 'behaviourally obstinate' used to refer to neurodiverse people, which tells me people without vocab or lived experience glob onto the poles, the extremes. Seems some serious offenders are rather close to my family tree", woman, ADHD, ND Community Hub.
- We are less capable and less intelligent: "The biggest misconception I've seen is that neurodivergence equals incapability or, on the harsher spectrum of judgement, a lack of intellect or that one with neurodiversity should be approached and treated like a child in some aspects. I can see and feel the difference in the treatment I get from people, particularly in the workplace, when I've been transparent about my neurodiversity, which then just reinforces masking in the workplace and eventual burnout from it", woman, ADHD, ND Community Hub.

## *The Role of Silent Spenders 2.0*

Summarising where we're at:

- A large, growing and valuable audience;
- Increasing consumer interest;
- A sorely under researched audience;

- All set against the backdrop of misrepresentation.

In 1982, advertising agency Bartle Bogle Hegarty (BBH) coined the phrase "When the world zigs, zag" as the tagline of a Levi's advert for the first ever pair of black jeans on the market. This idea of embracing uniqueness and challenging the status quo has become their brand ethos, impacting everything from the ads they produce to their approach to insight. This "zag" philosophy inspired BBH USA's research project, *The Silent Spenders*. Its core mission is to illuminate large, valuable and dramatically overlooked audiences that represent massive opportunities for brands. Because ND communities tick every single box, they became the definitive focus of this second edition. To add to this, many of the project team were ND themselves. This didn't mean we leaned into assumptions from our own lived experiences; instead, we used this project as a unique opportunity to better understand ourselves and the way we process the world.

## Research Objectives and Our Approach

### *Our North Star: Our Objectives*

With *Silent Spenders 2.0*, BBH USA wanted to create a meaningful narrative around ND communities as consumers in the US to:

- Help brands better understand and support those with ND, recognising those with ND as just as powerful consumers as neurotypical ones;
- Disprove the many misconceptions about this audience;
- Further empower the agencies and/or third sector who support them.

Whilst BBH USA was the project champion and chiefly responsible for insight generation, the resulting key insights and final marketing guidelines, Shape Insight were their go-to consultant, ensuring the methodology and the varying data capture methods used were correctly tailored for this group, as well as guiding the analysis process.

### *Evolving as Researchers: From Neurotypical Research to Neurodiverse Research*

We were conscious that our tried and tested approaches, which work with neurotypical people, wouldn't work with ND communities. Irrespective of whether we did qualitative and/or quantitative research, we had to approach the topic and those we spoke to in a different way. Therefore, we changed our approach to research from:

1. Direct questioning to reflective and projective: For those who identify with ND, direct questioning feels confrontational, overwhelming and assumes that there is a right answer. For many, this triggers stress or shutting down. This is why we designed our discussion guides and survey questions to either be reflective or projective. Either method removes the pressure and allows them to engage at their own pace.

2. Pre-defined question flows to organic conversations: We moved from pre-defined question flows and probing to completely organic conversations. This shift fundamentally changed the power dynamic in the conversations, enabling participants to feel at ease. This is because those with ND communication styles and the way they process information (ergo, questions) often isn't linear.

3. High cognitive questioning to low, grounded in simple language: It's too easy to fall into the trap of constructing (and asking in the moment) complex, abstract and cognitively demanding questions. This is hard for neurotypical people to grapple with, even more so for those with ND. Our discussions, guides and surveys were designed to focus on low, simple language. This shift allowed us to create an inclusive language, which then yields better insight from our participants.

### *Evolving as Strategists: From Trusting Our Gut to Checking Our Bias*

As strategists, we were also conscious that we simply couldn't approach this project in the same way we normally would. That means we:

1. Rejecting our muscle memory to focus and striving to be more open: Some of our team were ad agency strategists, who are used to researching with a very focused objective in mind, e.g., understanding people's motivations and behaviours relating to a certain product or brand. We had to reject our intuitive style of drilling deep and instead be entirely open to being led by the community. Part of this was being less wedded to our discussion guide, and more guided by what we heard and observed from the community.

2. Continuously checking ourselves on our conceptions about neurodiversity: We had to make a conscious effort to park our conceptions about neurodiversity at the door, and recognising that as that would never be entirely possible due to our unconscious bias, we would need to plan a rigorous review process for our own insight versus "trusting our gut".

3. Welcoming pivots (and pivots, and pivots!): The majority of marketing theory was created with a neurotypical lens, so we knew some of the marketing ideas we think of as "gospel" may turn out to be misleading. We needed to be comfortable being wrong, wrong and wrong again to get to a better understanding of this community.

### *Building Our Methodological Feedback Loop*

Not only did our approach to how we did research evolve to better reflect our audiences, so did the entire methodology! We felt that to truly understand ND communities, our approach would need not to be cookie-cutter and traditional. Therefore, we designed an approach that acted as a "feedback loop" (see Figure 3).

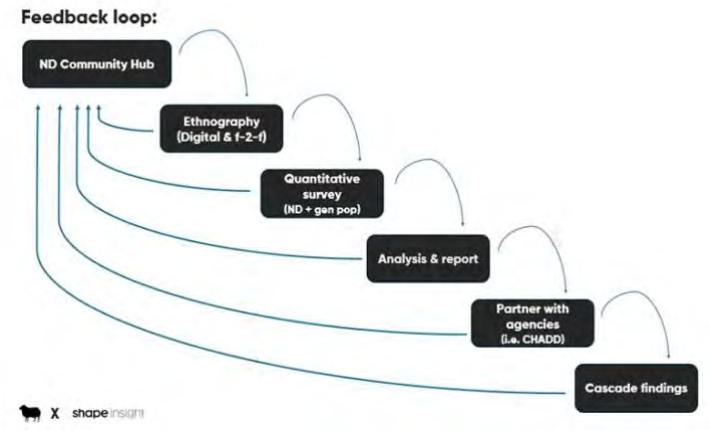



*Figure 3. The methodological "feedback loop".*

The ND community hub played a key role as a space where the core research team would continually share their thoughts, feelings and findings with the wider community as we moved through each step of the process, and vice versa. This "feedback loop" process was fundamental in not only shaping how we approached the next stage of the methodology, but also in refining our thinking and lines of questioning.

## The Zag and Our Findings

### *The "Zag": ND --> ADHD*

As we organically moved through the research, the challenge we had to overcome seemed to grow. This was driven by the diversity of the people we were speaking to. For instance, there are 17 broader spectrum conditions under the umbrella of ND (Cleveland Clinic, 2022). Based on the community we had built, we had representation of up to seven types of ND. It became clear to us that what might be a pain-point for a dyslexic person could be a superpower for those who had ADHD, as has been well-documented by Willcutt and colleagues as early as 2005. Therefore, BBH USA proposed a "zag"—a pivot in the name of stronger output—we would double down our efforts on those who identify as being ADHD because it's the fast-growing category of ND, with 13.9% of US adults (18–64 years of age) self-reporting receiving a formal diagnosis of ADHD from a doctor or healthcare professional (London, Monnat & Gutin, 2025). Meanwhile, among younger adults self-reported diagnosis rate reaches 20.9% (*ibid.*).

Copyright © **Esomar** 2026


### *The Big Overarching "Ah Ha!" Design for ADHD to win With Everybody*

We found that people with ADHD aren't just a large and growing consumer group—they're the blueprint for better marketing for everyone. This is arguably because they are, in many ways, the most discerning consumers: quick to spot friction, unwilling to tolerate unnecessary complexity and highly responsive to experiences that genuinely work. This makes people with ADHD a powerful barometer for quality. For example:

- They are 50% more likely than neurotypical people to abandon their cart while shopping;
- Their #1 reason for leaving a brand's website is that it is hard to navigate;
- As one study participant put it: "If a website triggers me, I'll just leave. I want to quickly navigate from A to B… this stuff affects me more than it does most people".

People with ADHD favour clearer pathways, fewer steps and intuitive navigation, making them the ultimate stress test for modern experiences. Also, moreover, optimising for ADHD isn't creating separate experiences for a niche audience—it's optimising for everyone.

### *Myth Busting and the Implications for Brands*

The aim of *Silent Spenders 2.0* was to smash the many myths and misconceptions that dog those with ND, but specifically those with ADHD. This north star helped frame not only the way we conducted our analysis, but also how we present our findings and resulting implications for brands (see Table 4).

| Myth | Reality | Explanation | Key quant evidence | Key qual evidence | Implication for brands |
|---|---|---|---|---|---|
| Unpredictable spenders | Ultimate unexpected loyalists | Once they love a brand, they'll likely stick with it. The mental work that comes with switching (heavy research, disruption to routine) turns them off, whereas familiar brands offer comfort and reassurance. | 90% stay loyal to brands they like, higher than non-ADHD. | "I like trying new things but only when they don't completely disrupt my routine. I think a lot about whether trying something new is worth the mental effort". | Deliver against brand promises<br><br>Focus on loved products as well as NPD<br><br>Make it easy to stay e.g., option to automatically refill |
| Impulsive spenders | Over research to feel confident | Knowing they can tend to impulsivity, they are overly cautious when shopping. They research hard, are | 4x more likely to research heavily than going with their gut | "I'm a textbook review doomscroller. I have to look at every option on the product | Show/prove how you're the best option e.g., highlight quality creds, reviews, comparisons |

| | | | | | |
|---|---|---|---|---|---|
| | | more likely than non-ADHD to abandon cart if it's not perfect and are more cautious on risk-heavy activities like investing. | when shopping. | page before buying…I don't want to miss out on the best". | Offer flexible return policies |
| Behavioural tactics work universally | Urgency tactics alienate and often kill the same | They're our go-to tools, but using wrong behavioural tactics can backfire. Urgency, scarcity or other marketing tricks can trigger intense anxiety that kills the sale entirely. | Sales and limited time offers are 1/3 less likely to influence purchase than for non-ADHD. | "They design things that prey on how you think to catch you out…like free trials that charge you $100 unexpectedly. That's not cool". | Let them shop on their own terms and in their own time<br><br>Focus on the core idea of delivering consumer trust to reduce reliance on this kind of tactic |
| Easily overstimulated, so safer to design plain and simple | Colour and pattern are key communication tools, used correctly | Beige brands are inaccessible ones. Color and pattern are essential communication tools for people with ADHD to navigate the world. "Under-stimulating" is as much of a problem as "overstimulating". | Brands with bold, vibrant colors are 2x as likely to make a lasting impression than non-ADHD. | "Minimalism is my worst nightmare. My brain sees colour and design first. If there are no colour-coded weights in the gym it can take me 10 mins to find the right ones!" | Train design teams on truly neuro-affirmative design—the right amount of colour and vibrancy.<br><br>Colour code products or allow them to be customised. |
| More drawn to action than relaxation | Relaxation need is there, just looks completely different | Relax through stimulation, not silence. Think less spas, LED masks and meditation and more stimulation, multitasking and passionate rabbit holes. | 2x as likely to find socially gaming with music on relaxing than non-ADHD. | "Zen looks completely different to me. Silence freaks me out. When I have music blaring, my brain gets to sit back while my body takeover". | Consider that your brand may be part of the comfort economy, even if it doesn't feel like it to you. |

| | | | | | |
|---|---|---|---|---|---|
| AI lifts some of the cognitive load | Ultimate AI adopters, but more stressed out by it. | Have a love-hate relationship with AI. Seeking ways to automate and lift cognitive load in almost every sector, but "set it and forget it" can actually leave them feeling less in control. | 50% more likely to use AI tools than non-ADHD, but 2x as likely to say they make them feel more anxious. | "Our AI assistant has been life changing. It reminds me to lock the doors at night, to get the kids to brush their teeth, to drink water and whatever else we need it for". | Reframe as "set it and don't forget it".<br><br>Consider that ADHD folks may seek AI help in more categories than most— e.g., not just time keeping but finance, personal relationship management, etc. |

*Table 4. Summary of key findings and recommendations.*

## *The Impact of AI: A Warning*

AI has come to dominate the public conversation. Naturally, the research and marketing community has begun to explore its impact on neurotypical audiences in depth. However, once again, ND communities, particularly those with ADHD, have been overlooked. To address this gap, we explored the impact of AI on this specific audience. As noted earlier, our survey found that this group is 50% more likely to use AI tools than non-ADHD. Their usage is defined by a love–hate relationship that creates a unique tension: AI empowers them, while at the same time heightening anxiety. In fact, they are twice as likely to say that AI makes them feel anxious. This finding, combined with the rapid expansion of AI across both public and private touchpoints, is concerning. We hope this work serves as a rallying cry for further exploration into how AI can continue to empower those with ADHD while reducing its negative effects.

## *The Research Format: We Told the World a Way that Was Designed For ADHD...Literally*

The research process taught us just how important thoughtful design is to ensure experiences are ADHD friendly. One study participant told us "Minimalism is my worst nightmare…my brain sees colour and design first". Our quant study found that people with ADHD's number one reason for leaving a brand's website was that it was hard to navigate, and that they are twice as likely as neurotypical people to leave due to bland design. Thus, we knew our research output had to be designed with best practices for ADHD in mind. We worked with neurodiversity expert Understood to brief our design team with the following principles:

1. Colours: Must be bold and appealing while avoiding harsh contrast or anything overly distracting;
2. Typography: Strong with easily readable fonts must be chosen that are comfortable to engage with;
3. Copy: Keep messages short, clear and meaningful. Convey the information in the most succinct way possible and "chunk out" different points;
4. Graphics: Visual elements should support the message and gently guide the eye, never overpowering the copy;
5. Sense check: Sense check the final product with true neurodiversity design experts at Understood.

Copyright © **Esomar** 2026

*Figure 4. The ADHD-friendly design identity.*

## The Impact (Now and Tomorrow)

We released the findings in April, during neurodiversity celebration month, and successfully began to create a narrative around the neurodiverse consumer. Our learnings achieved 12.7k impressions on LinkedIn, with a strong 55% engagement rate (over 10 times the benchmark for businesses on LinkedIn at 5% according to Rival IQ [Feehan, 2024]).

| Selection of LinkedIn Testimonials |
| --- |
| *"I have ADHD and six pairs of the same pants in different colours from Free People. This report finally explains my entire purchasing history. But beyond the personal recognition, the insight here is genuinely important for our industry. We're not the impulsive consumer. We're the most demanding one. Earn our trust and you've earned something rare. And it takes real rigour to centre research this thoughtfully on a community that marketing so often gets wrong. Well done, BBH".* |
| *"As someone with ADHD I was afraid to read yet another pathologising report, but this one was actually on point. Many people still struggle with impulse buying but this quote stood out: 'Far from reckless, adults with ADHD tend to be relentless about getting it right'. 💜 Also agree that the pressure and urgency tactics don't work well—we have a high B.S. radar".* |
| *"Great topic! ADHDers have been living in the future for a while. The challenges that are becoming more common for everyone (constant distractions, information overload, too many decisions, too many tabs open) are challenges ADHDers have been navigating our whole lives. I love the idea of ADHD as a stress test. The things that make life easier for ADHD brains often make life easier for all human brains ✨".* |
| *"Mum of an ADHD child who's grown into an amazing adult. This is great—thank you for making this research a priority".* |
| *"Feeling: Seen 👁".* |

*Table 5. Testimonials via LinkedIn.*

Six industry and wider business press houses wrote about the research, including Fast Company, Little black Book, Shoot, Design Rush, Quirks and the Breaking and Entering daily show—with more content to be published via WARC and Research Live in the near future. Combined, these publications delivered an estimated potential earned audience of more than 10 million monthly industry and business readers, viewers and listeners, based on publicly available audience figures and publisher-reported reach.

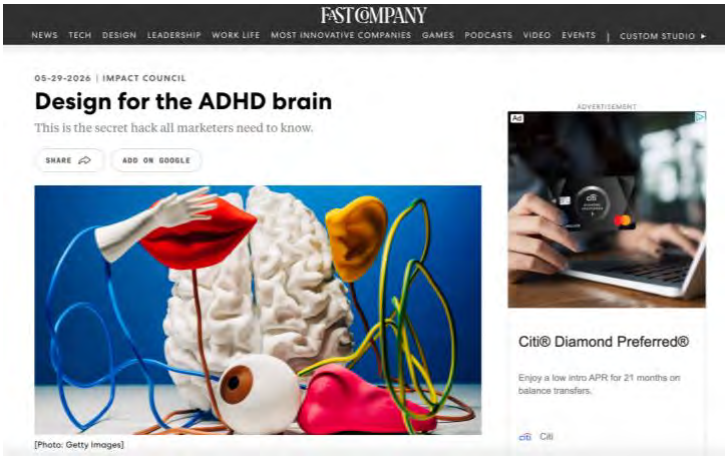



*Figure 5. Screengrab of work featured on Fast Company.*

We also created a narrative in our own companies: BBH USA is now in contract with Understood to better design its office space and hiring practices to be more neuro-affirmative.

> *"The ADHD edition of silent spenders is a true research 'zag', from our focus on ADHD consumers to our methodology. We rejected the typical 'focus group' default in favour of a process truly tailored to this audience, designed to mitigate against our biases. What we found is that when you design for neurodivergence, it's better for everyone, and that's something we're taking forward into so much of our business: how we conduct our research, how we operate in our teams, the work we create, even in how we think about our office space",* (Samantha Deevy, BBH CSO).

That's just an overview of the impact since we published our findings. This piece of work continues to gain traction and is fundamentally changing how our clients see those with ADHD. The inspiration for this project was the radio silence around neurodiversity in the marketplace. We are proud to report that we have begun to address that problem.

And, finally, both BBH USA and Shape Insight undertook this study pro bono, and participants who took part in the "ND Community hub" similarly contributed their time, perspectives and personal stories over an extended period—freely and with great honesty. We want to use this as an opportunity to thank them for their instrumental input! Without them, this project wouldn't have been the success it is today.

## References

Cleveland Clinic. (2022). *Neurodivergent*. Available: https://my.clevelandclinic.org/health/symptoms/23154-neurodivergent

Google Trends Analysis. (n.d.) *Interest over time: neurodiversity.* Available: https://trends.google.com/explore?q=neurodiversity&date=all&geo=US

London, A. S., Monnat, S. M., and Gutin, I. (2025). Self-reported ADHD Diagnosis Status Among Working-Age Adults in the United States: Evidence from the 2023 National Wellbeing Survey. *National Center for Biotechnology Information.* Available: https://pubmed.ncbi.nlm.nih.gov/39963833/

Copyright © **Esomar** 2026


NCLD. (2017). The State of Learning Disabilities: Understanding the 1 in 5: Executive Summary. *National Center for Learning Disabilities.* Available: https://assets.ctfassets.net/p0qf7j048i0q/2Q2TsAzUSM9TXq22LPxQwY/658d54ed0529acddf70f89d367db7915/2017_State_of_LD_-_Executive_Summary_Final_Accessible.pdf

Feehan, B. (2023). 2024 LinkedIn Benchmark Report. *RivalIQ*. Available: https://www.rivaliq.com/blog/linkedin-benchmark-report/?utm_source=chatgpt.com

Willcutt, E. G., Pennington, B. F., Olson, R. K., Chhabildas, N., and Hulslander, J. (2005). Neuropsychological Analyses of Comorbidity Between Reading Disability and Attention Deficit Hyperactivity Disorder: In Search of the Common Deficit. *Developmental Neuropsychology, 27*(1), 35-78.

## About the Authors

Alexander Holmes, Director, Shape Insight, London, UK.

Jessica Garlick, Strategy Director, BBH, US.

Heidi Kim, Senior Strategist, BBH, US.

Sebastian Giugovaz, Junior Strategist, BBH, US.

# The Ripple Effect

## Turning a 60+ Market Brand Tracker into a Strategic Decision System

*Alexander Shmelkov*

*Sylvia van der Waal*

*Muge Vardar*

## Esomar

Office address:

Burgemeester Stramanweg 105

1101 AA, Amsterdam

Phone: +31 20 664 21 41

Email: info@Esomar.org

Website: www.Esomar.org

Publication Date: September 2026

ESOMAR Publication Series Volume S424 Congress 2026

ISBN 92-831-0337-8

Congress 2026
Copyright © **ESOMAR** 2026

Copyright © **ESOMAR** 2026
Congress 2026

# The Ripple Effect

## Turning a 60+ Market Brand Tracker into a Strategic Decision System

*Alexander Shmelkov*

*Sylvia van der Waal*

*Muge Vardar*

### 1. The Problem with Brand Tracking

Most organisations have never had more data than they do today. The sources are diverse: always-on digital signals, syndicated panel data, proprietary survey infrastructure and a growing ecosystem of third-party suppliers. And yet, the demand for more data-driven decision-making persists across almost every business. That persistent gap is the first thing worth examining, because it points to something the industry has been slow to accept.

Brand tracking, in its most common form, has settled into one of two failure modes. The first is the KPI machine: tracking degraded to a transactional reporting tool that produces a clean quarterly deck, gets presented at the brand review and changes nothing. The second is the sprawling omnibus: a survey that tries to answer every open question in the organisation, runs to 40 minutes or more, delivers rich but unactionable output and struggles to justify itself in a budget cycle. Neither failure mode is a methodology problem. Both are a design problem – a mismatch between what the tracker is built to produce and what the organisation actually needs to decide.

IKEA's previous tracker, Brand Capital, had operated for over twenty years. It was well embedded in the business and provided a genuine long-term view of the brand across global markets. But over time, it had become a bureaucratic milestone rather than a strategic input. Teams knew when it was coming. They built presentations around it. They did not build decisions around it.

The question we set out to answer was not how to build a better tracker. It was how to build a system that changed the way decisions were made. That distinction matters. A better tracker produces better reports. A decision system changes what happens after the report.

### 2. Organisation Setup has a Significant Impact on Design Decisions

Replacing a 20-year-old tracking system in IKEA was not primarily a methodology problem. It was an organisational one. Understanding why requires understanding how IKEA is structured.

IKEA operates as a franchise system. Multiple organisations in the system contribute to building and managing the IKEA brand globally. At the time of the Brand Pulse transition, this meant navigating alignment across a complex organisational setup. In several cases there were other brand tracking studies that co-existed alongside Brand Capital.

Brand Capital was launched in 2005 and went through a redesign in 2015. There were several attempts to revitalise the brand tracking that did not land anywhere. That history shaped the conditions under which Brand Pulse was developed. Any new solution had to earn organisational trust before it could expect organisational adoption.

Building one trusted source of truth for brand monitoring across the IKEA franchise system was therefore the primary objective. The most important motivation for transition was a need to bring alignment both to measurement and to the business frameworks where brand performance data plays a role.

Brand tracking does not exist in isolation; it is part of a broader brand management process. To ensure the acceptance and adoption of Brand Pulse, it had to be fully in sync with the key strategies and business frameworks, cadence of decision-making and performance language that different parts of the franchise system already used. This meant investing significant time before any survey design decisions were made in understanding how brand performance was discussed, evaluated and acted upon across global and local functions.

The second was brand performance measurement alignment. Different organisations within the IKEA franchise had developed different approaches to reporting brand health over time. Establishing a shared measurement model grounded in the empirical laws of brand growth required making explicit choices about what would and would not be measured and defending those choices against the natural pressure to accommodate every existing practice.

Together, these two workstreams ensured that when Brand Pulse launched, it reflected the actual structure of brand management decisions across the franchise. It was built to serve the whole value chain of IKEA.

## 3. The Brief as Design Philosophy

There is no universal template for designing a brand tracker. The right design depends entirely on the organisation's structure, its decision-making processes and its vision of how a brand grows.

The vision for Brand Pulse was inspired by the evidence-based marketing principles devised by the Ehrenberg-Bass Institute. IKEA is sponsoring their research and this collaboration provided us with the opportunity to bring grounded theoretical principles into the business environment. The foundational elements of Laws of Growth became the backbone of Brand Pulse: Mental and Purchase Availability, Category Entry Points, Distinctive Brand Assets (for more details see the key publications of the institute; Sharp 2010, Romaniuk 2023, Romaniuk 2018).

The brief for Brand Pulse was the final output of an extensive R&D process. Before a single supplier conversation took place, we invested in understanding the failure modes of Brand Capital and the conditions under which a successor system would need to operate. That R&D program included:

A series of 45 stakeholder interviews to identify the key functional pain points and decision-making gaps across global and local teams.

A consumer centricity audit of the Brand Capital survey, assessing the degree to which its measurement architecture reflected how consumers actually engage with the IKEA brand.

A review of internal data delivery processes, mapping the gap between when data was produced and when decisions were being made.

A series of qualitative and quantitative projects to identify and validate the Category Entry Points relevant to IKEA's category. Four dedicated CEP studies were completed before the tracker design was finalised.

A series of proof-of-concept projects testing different survey approaches, modular architectures and design principles under real fieldwork conditions.

A design thinking workstream developed with the key user group to adjust the study design to specific user patterns.

This upstream investment produced a brief that gave potential suppliers full transparency on the typical topics reflected in the brief (e.g., business situation, key strategies and frameworks, challenges, study objectives) as well as specific research conventions and positions. Such clarity of objectives upstream makes every downstream decision faster and more defensible. It also allows the research partner to build a solution in genuine collaboration and on a solid knowledge foundation, rather than a blank specification sheet.

NielsenIQ eventually won the pitch. Reflecting back on this, they commented what made Brand Pulse different to design was the quality of the starting point:

*"Most clients come to us with a list of questions they want answered. IKEA came with a clear theory of how the brand grows, a specific set of entry points they needed to track, and a budget commitment that reflected genuine strategic intent. That combination [of] clarity of purpose, methodological flexibility, and commercial seriousness is [what] allowed us to make design decisions based on what the data needed to do, not what was easiest to produce."*

## *The Modular Architecture*

Brand Pulse is a modular brand tracking system. It is designed to provide ongoing measurement of core performance metrics as well as the ability to dive deeper in specific areas. This approach allowed us to keep the length of the interview under control at 15 minutes while covering a wide array of topics.

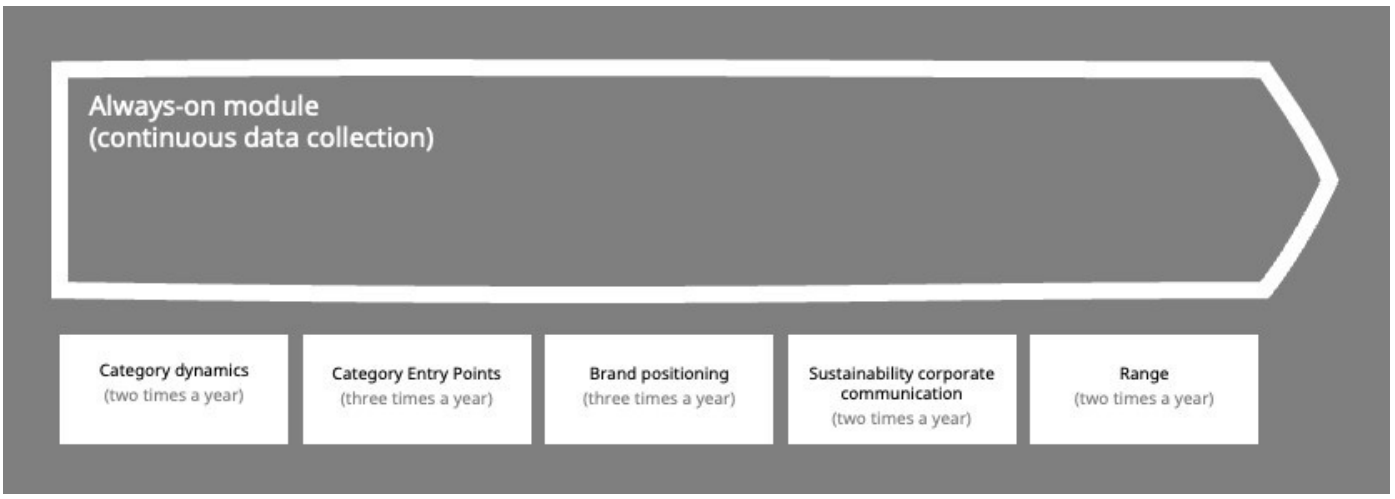



*Figure 1.*

From the change management perspective, the study modularisation created the biggest change in the brand performance measurement in IKEA. It challenged the role and the definition of the brand tracking – what it is and what it

is not – and led to significant reconsideration of the scope of the study. It also allowed us to establish a discipline for treating the business request to include a topic or a metric in the study.

This discipline is directly connected to the operational outcomes. The lean, focused architecture is what enabled the program to reduce cost-per-interview by 45% and at the same time triple the global sample in comparison to Brand capital. Across 64 markets, Brand Pulse now captures over 440,000 interviews annually. The consolidation of duplicate tracking infrastructure delivered another substantial annual cost savings to the IKEA franchise system.

## 4. Change Management as a Key Differentiator

A tracker that produces better data is not the same as a tracker that produces better decisions. The gap between the two is where most insights functions lose the value they have created. Better data requires a capable organisation to receive it, interpret it and act on it. Building that organisational capability is change management work.

Brand Pulse moved into operational delivery in autumn 2024 and was made fully available to IKEA in June 2025. By that point, the study was live across all 63 active IKEA markets, ETL pipelines were ingesting 12 million records per month into the data lake and market-level users had access to results via Power BI dashboards. Having access to data, however, is not the same as using data. The launch of the product was the beginning of the change management programme.
The programme was structured around five sequential stages of product adoption.

### *Stage 1. Creating Awareness*

It is crucial to create awareness and set expectations for the transition in advance and among the broadest possible target group. We approached this stage as an internal communication project – from developing a distinctive visual design code of Brand Pulse deliverables, through pre-launch internal communication, to a launch week with events and trainings on site and online for over 700 IKEA co-workers across the globe.

### *Stage 2. Stimulating Desire*

This stage relied primarily on targeted digital meetings with key stakeholders, building personal familiarity with the tool and surfacing the specific decision contexts where Brand Pulse data was most directly applicable.

### *Stage 3. Building Foundational Knowledge*

Targeted training programmes were designed for specific stakeholder groups from both local and global functions ranging from focused three-hour onboarding sessions to full two-day immersion programmes depending on the depth of engagement required.

### *Stage 4. Increasing the Ability to Use*

A bi-weekly drop-in session format was introduced: open calls for the local IKEA insights community with no fixed agenda. Country insight leaders and specialists could join and ask the team any question about Brand Pulse. This format proved foundational in building the transparency and organisational trust that sustained adoption beyond the initial launch period.

### *Stage 5. Learnings Reinforcement*

The approach here was to identify and cultivate internal champions who could drive adoption by example and provide early warnings of emerging challenges or misinterpretations before they became structural problems.

## The Trend Break Problem

One specific challenge deserves naming directly, because it is a problem any organisation replacing a long-running tracker will encounter. The design choices made for Brand Pulse made direct comparison with the legacy trend impossible. For markets where leadership teams had been managing against Brand Capital trends for years, this created challenges, especially when it came to follow-ups and goal setting.

To offset the challenge of the trend break, we took a deliberate decision to run Brand Pulse in parallel with Brand Capital during the launch year. This allowed us to connect the two trend lines and observe how the new and legacy systems related to each other in practice. From the NielsenIQ perspective, the calibration of the legacy trends represented a methodological challenge, as the designs of the studies were substantially different:

*"Together with IKEA we identified limitation[s] and constrain[t]s in granularity and scope for the calibration that allowed [us] to find a methodologically sound solution. And from the change management perspective, IKEA has set specific requirements to where, when and what for the calibrated trends shall be used."*

Although the calibration allows us to partly offset the challenge of the trend break, it is not a silver bullet. The trend break is however also a learning opportunity for the organisation to realise the probabilistic nature of survey research and develop better understanding of the relativity of brand performance in the market context. We demonstrated that despite changes in absolute metric values, the structural view of market dynamics remained consistent. The relative positioning of markets, as well as the directional signals of brand health, was consistent across both systems. That structural stability, rather than statistical calibration alone, was the more durable answer to the anxiety about baseline loss.

## 5. Early Results and What We Are Still Learning

It is important to be honest about the difference between what we know and what we believe. The operational results are clear and measurable. Brand Pulse now captures over 440,000 interviews annually across 64 markets. Cost-per-interview is half what it was under Brand Capital, while the global sample has tripled. The consolidation of duplicate tracking infrastructure delivered significant cost savings to the IKEA franchise system. Data that previously was updated once a year is now available monthly via Power BI dashboards and supported by reports and insight deliverables that are defining the consistent brand performance narrative.

The strategic impact of the transition will only be fully known in the coming years. However, at this stage we can talk about the evidence of the positive impact that Brand Pulse has had on the organisation.

Consistency. Brand Pulse data is now used as a single source of performance data across the whole IKEA value chain. This means using Brand Pulse as a source of global and local goal setting and ambition-level identification.

Change of the marketing model. Brand Pulse and the R&D work conducted for it contributed to the evolution of IKEA's marketing model.

Contribution to creative excellence. Category Entry Points was by far the most anticipated and successful innovation of Brand Pulse.

Performance reporting. Brand Pulse directly contributes to the enterprise-level strategic performance reporting.

Strategic exploration. Brand Pulse is the only available study that represents the general population of 18–75-year-olds with a consistent methodology and cadence for all 63 IKEA countries, allowing for analysis on a global, country and store level.

The unique opportunities to crack the long-standing dilemmas and new challenges are now to be put to the test with several strategic exploratory projects; the full benefits of the program are yet to be unlocked.

## 6. Three Principles for Strategic Insights Design

Brand Pulse is a case study in organisational transformation, not just a template for building a tracking system. The specific design choices emerged from IKEA's specific context. A different organisation, with different structural conditions and a different theory of brand growth, would build something different.
What can be transferred are the principles that shaped the process.

### *Principle 1: Ruthless Prioritisation and Scoping Are Ways to Improve the Value of the Brand Tracking*

The discipline required to keep a global tracker lean is counterintuitive inside large organisations. The pressure to add questions, expand scope and accommodate every stakeholder's knowledge gap feels reasonable at the local level. The explanatory power of Brand Pulse is a direct product of cutting everything that did not serve the core measurement objective.

### *Principle 2: Brand Tracking as a Strategic Decision-Making Tool*

A well-designed brand tracking study that properly represents your footprint will come at a cost. The opportunity is that it can deliver so much more than just performance KPIs. A global brand tracker has a unique advantage: it uses a standardised methodology across multiple countries and cultures. It allows the company a unique opportunity to uncover cultural nuances and similarities and make strategic choices. This requires balancing the global comparability and local relevance during the study design. But our experience shows that it is possible to find a sweet spot between them.

### *Principle 3: Prioritise Scope Over Granularity*

The common challenge of global tracking is excessive detail, such as representing too-narrow audiences or going to a specific level of granularity. Good tracking is, in the words of Jenni Romaniuk, "designed for the category, reported for the brand" (2023). Our experience shows that avoiding the prioritising of customers, category shoppers and decision-makers and focusing on the general population provides much more opportunity to see the real market potential of the brand and identify the ways to harness that potential.

Copyright © **ESOMAR** 2026
Congress 2026

The work is not finished. Brand Pulse is now a functioning strategic decision system. Making it an indispensable one – embedded in every significant marketing and investment decision across the IKEA franchise network – is the next phase.

## References

Romaniuk, J. (2023). *Better Brand Health: Measures and Metrics for a How Brands Grow World.* Oxford University Press.

Romaniuk J, (2018). *Building Distinctive Brand Assets.* Oxford University Press.

Sharp, B. (2010). *How Brands Grow: What Marketers Don't Know.* Oxford University Press.

## About the Authors

Alexander Shmelkov, Director, Global Brand Tracking, IKEA, Delft, Netherlands.

Sylvia van der Waal, Global Brand and Marketing Insights Leader, IKEA, Delft, Netherlands.

Muge Vardar, Global Client Lead, NielsenIQ, Madrid, Spain.

# Citizen Intelligence in Action

## Turning Mobility Insights into Public Transport Change

*Raquel Marzo*

*Josep Chiner*

*Amparo López-Vicente*

*Carol Soriano*

*Juan F. Giménez*

*Elisa Signes*

*José Solaz*

*Raquel Alario*

Esomar

Office address:

Burgemeester Stramanweg 105

1101 AA, Amsterdam

Phone: +31 20 664 21 41

Email: info@Esomar.org

Website: www.Esomar.org

Publication Date: September 2026

Esomar Publication Series Volume S424 Congress 2026

ISBN 92-831-0337-8

Copyright © **Esomar** 2026

Copyright © **Esomar** 2026

Congress 2026

# Citizen Intelligence in Action

## Turning Mobility Insights into Public Transport Change

*Raquel Marzo*

*Josep Chiner*

*Amparo López-Vicente*

*Carol Soriano*

*Juan F. Giménez*

*Elisa Signes*

*José Solaz*

*Raquel Alario*

### Introduction

The European Green Deal positions sustainable mobility as a cornerstone of the European Union's climate neutrality objectives, calling for cleaner, more accessible and more affordable public transport as part of the wider ambition to reduce greenhouse gas emissions by 55% by 2030, and achieve climate neutrality by 2050 (European Commission, 2019). Achieving this transition requires more than vehicle decarbonisation; it also depends on a modal shift towards collective and active modes, with public transport acting as the backbone of urban mobility. However, in many European cities, the full potential of public transport remains underused. Private car use continues to be sustained by perceptions of flexibility, reliability and comfort, while collective public transport is often chosen when it is affordable, well-connected and easy to integrate with other modes (UPPER Consortium, 2024a). This suggests that increasing public transport uptake is not only a question of expanding infrastructure or service supply. Citizens must also perceive public transport as a credible, accessible and reliable alternative to private vehicles, which means improving the everyday experience of using it.

Understanding everyday public transport experience requires moving beyond traditional operational metrics such as ridership, punctuality, infrastructure capacity or service frequency. While these metrics are essential for managing transport systems, they fail to reflect the subjective, lived factors that ultimately determine whether individuals adopt, maintain or reject public transport. These factors are especially relevant for vulnerable groups, including older adults, people with functional diversity, as well as users who experience uncertainty when navigating the transport system (Sree Yarlagadda & Gupta, 2025). This paper argues that improving public transport uptake requires insight systems capable of capturing lived mobility experiences and translating them into action. In this context, citizen science offers a valuable framework. It refers to the active engagement of citizens in knowledge generation, allowing them to contribute not only as respondents, but also as collaborators in identifying problems, generating evidence, shaping solutions and validating outcomes (Sauermann et al., 2020; Vohland et al., 2021). The literature highlights the importance of inclusive, flexible and adaptive research designs, where citizens can participate across different stages of the process (Robinson et al., 2018). This is particularly relevant for public transport, where the users most affected by accessibility or information barriers are often those whose needs remain least visible in conventional planning processes.

Although citizen science has been increasingly applied to mobility challenges, much of this work remains limited to specific stages of participation, such as data collection, reporting or consultation. Some projects also fail to involve sufficiently diverse and representative populations, limiting the potential of citizen science to capture lived experience and inform change (Sauermann et al., 2020; Sardo et al., 2022). In the mobility field, existing initiatives show the value of connecting citizens with local authorities and policymakers, but also highlight the need for stronger institutional commitment if citizen evidence is to influence decisions (Bike Barometer, 2022). Similarly, studies of public participation

in bus transit policymaking show that citizen engagement can inform policy and implementation, but do not necessarily provide an integrated methodology for turning citizen-generated evidence into actionable insights, co-created solutions and validated public transport interventions (Nalle et al., 2023; Sardo et al., 2024). This points to a key gap: the need for a structured citizen science-informed insight system that connects evidence generation, insight translation, co-creation, implementation and validation.

The Horizon Europe project (UPPER) responds to this gap by developing a citizen science-informed insight methodology for public transport innovation. Rather than treating participation as a stand-alone consultation activity, the project connects citizen-generated evidence with participatory ideation, implementation and validation. In doing so, it creates a structured process through which mobility experiences can be translated into actionable insights and used to inform public transport decisions. Grounded in the principle of *Mobility as a Right*, the paper frames mobility not simply as a consumer service or market-based utility, but as a condition for autonomy, equitable access, social participation and urban life (Mitropoulos, 2025).

## Methodology: A Citizen Science Methodology for Insight-Led Public Transport Change

The citizen science-informed insight methodology was designed as a learn–ideate–validate process:

- Learning from citizens' lived mobility experiences;
- Ideating with vulnerable users and mobility stakeholders;
- Validating implemented interventions through citizen feedback.

As shown in Figure 1, the methodological design combined scale, depth and actionability by connecting three consecutive phases: evidence generation, participatory ideation and real-world validation.

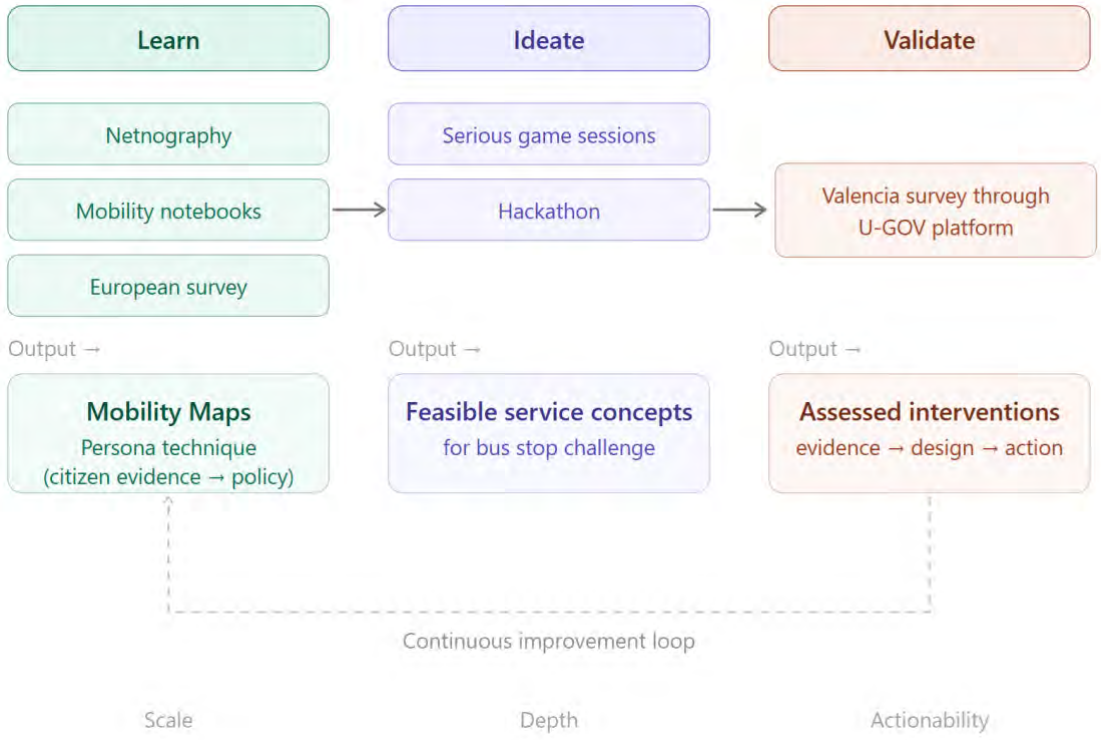



*Figure 1. Citizen science methodology.*

The "learn" phase was based on methodological triangulation, integrating netnography (Kozinets, 1998), experiential mobility notebooks (Fortin, 2014) and a large-scale European survey to capture citizen-generated evidence at scale, in context and across user groups. These insights were

synthesised into mobility maps using the "persona" technique (Pruitt, 2010), translating complex evidence into accessible narratives for mobility stakeholders. The "ideate" phase used serious game sessions (DiStaso, 2022) and a hackathon (Hevenstone, 2026) to convert identified needs into feasible concepts for the Valencia bus stop challenge. Finally, the "validate" phase assessed the implemented measures through user feedback, using U-GOV as the participatory platform that integrates the citizen science methodology developed in UPPER.

## Learn: Triangulating Citizen-Generated Evidence

### *Netnographic Analysis: Capturing Citizen-Generated Mobility Evidence at Scale*

The first source of evidence in the "learn" phase was a netnographic analysis of spontaneous citizen feedback on urban transport experiences. Netnography was used to analyse public reviews related to six transport modes (shared bike, bus, tram/subway, taxi, shared light electric vehicles and shared car) across five UPPER living labs: València, Île-de-France, Rome, Oslo and Mannheim (Soriano et al., 2024). This approach allowed the project to capture citizen-generated evidence at scale, based on what users voluntarily express when public transport and shared mobility services succeed, fail or create friction in everyday life. Across the full UPPER user research programme, the netnographic intervention analysed 23,739 reviews, including 3,765 reviews in Valencia. Textual comments were processed using natural language processing (NLP) to extract sentiment polarity (positive, negative or neutral), as well as emotional tone, including anger, joy and sadness, alongside indicators of hate or aggressive expression. As shown in Figure 2, this analysis helped compare user perceptions across transport modes and cities, showing that the bus was one of the most critical modes from the perspective of user experience. However, Valencia's bus service performed more favorably than the bus services analysed in the other UPPER cities.

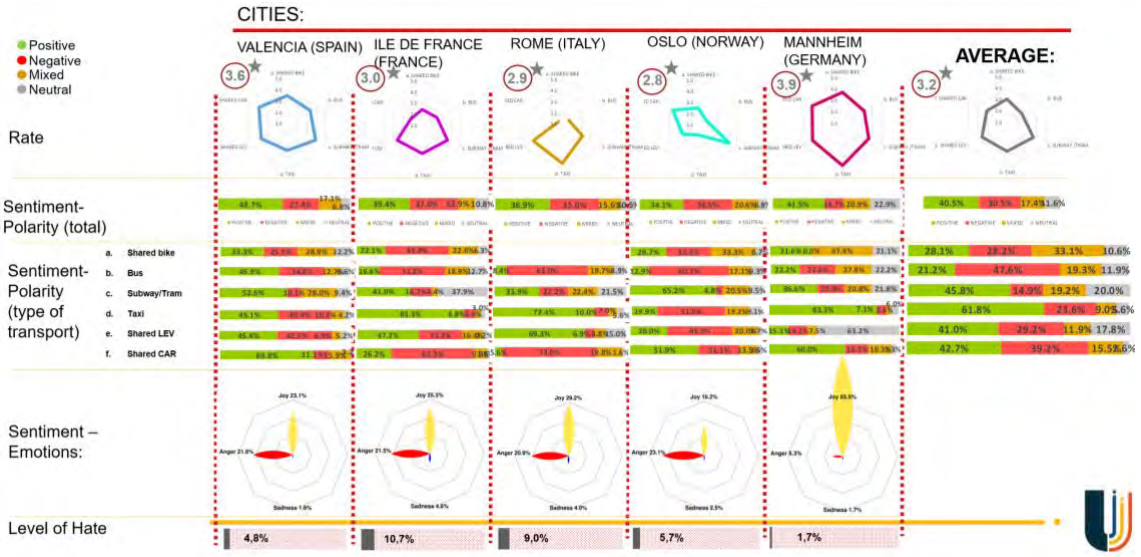



*Figure 2. Netnographic evidence on bus service perception across UPPER living labs.*

In addition to sentiment and emotion analysis, a semantic analysis of the collected reviews was conducted to identify the topics most strongly associated with positive and negative perceptions. In the case of bus services, negative comments were particularly linked to waiting times, lateness and delays, identifying predictability and reliability as central drivers of perceived public transport quality. These findings are especially relevant, because bus services operate in direct interaction with urban traffic and are therefore more exposed to everyday operational uncertainty.

### *Experience Notebooks: Understanding Mobility Decisions in Everyday Context*

The second source of evidence in the "learn" phase was the experience notebook, a qualitative technique designed to capture citizens' everyday mobility practices, decisions and barriers from their own perspective. Unlike netnography, which relied on spontaneous public reviews, the

notebooks provided contextual depth by asking participants to describe routine and leisure journeys, the reasons behind their transport choices and the difficulties encountered. A total of 72 participants from eight countries completed the notebooks, including 24 participants in Valencia. Participants included user groups whose needs are often under-captured in conventional mobility planning, such as older adults. The collected narratives showed how accessibility, perceived safety, physical effort, service information, journey complexity and confidence in the system influence the use of public transport and other mobility options. The material was analysed through semantic and thematic coding, and synthesised into visual summaries for each user profile. As shown in Figure 3, these outputs translated individual mobility experiences into structured insight artefacts for subsequent co-creation activities.

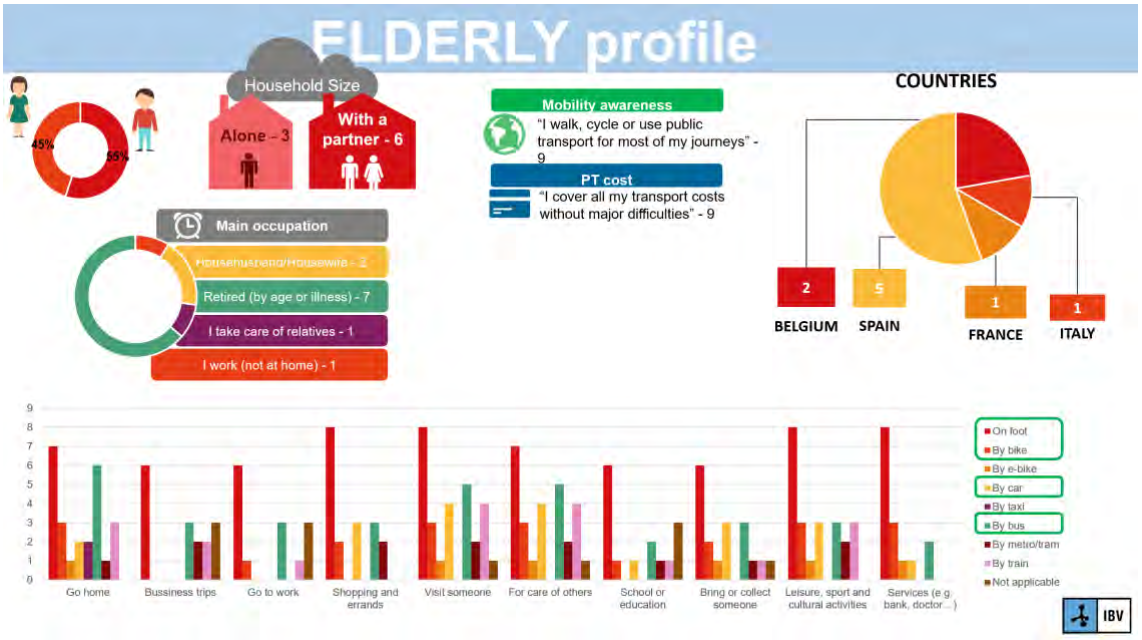



*Figure 3. Older adult mobility profile generated from experience notebook evidence.*

Figure 3 provides an illustrative example through the older adult profile. The notebooks showed that the bus is one of the most frequently used public transport modes among older adults, but also that accessibility is one of their main concerns. For this group, being able to access, understand and use the service safely and autonomously is often more decisive than travel time, reinforcing the relevance of clearer and more accessible information at bus stops.

## *Quantitative Survey: Validating Mobility Priorities*

The third source of evidence in the "learn" phase was a large-scale European survey designed to validate and quantify the insights emerging from the qualitative research. While netnography provided spontaneous citizen-generated evidence and the experience notebooks offered contextual depth, the survey enabled the project to assess the relative importance of mobility barriers, motivations and expectations across a broader population. The survey involved 2,676 participants across nine countries linked to the UPPER demonstration sites: Spain, France, Italy, Norway, Germany, Portugal, Belgium, Hungary and Greece. In Spain, the sample included 303 participants, with particular attention to vulnerable road users, such as elderly citizens, women and low-income travellers. The questionnaire explored mobility habits, frequency of use, perceived importance and satisfaction with different transport modes, reasons for using or not using public transport, perceived safety and expected improvements. The results showed that buses were the most widely used public transport mode, with 88.2% of respondents reporting some level of bus use, followed by metro, tram or train services, used by 80.5% of respondents. This confirmed the central role of the bus in everyday public transport experience. However, as shown in Figure 4, the survey also identified clear areas for improvement, particularly in relation to predictability, reliability, accessibility and connectivity.

*Figure 4. Priority improvements identified for bus services.*

## *Mobility Maps: Translating Citizen-Generated Evidence into Actionable Policy*

The final output of the "learn" phase was the development of mobility maps, which synthesised netnography, experience notebooks and survey findings into actionable and accessible narratives for mobility stakeholders. Applying the "persona" technique, the project translated complex datasets into structured representations of user groups. Each mobility map combined demographic characteristics, mobility behaviours, motivations, barriers, needs and expectations. The mobility maps were especially useful for identifying how accessibility and information barriers affect the use of bus services. As shown in Figure 5, the older adult profile highlighted the importance of the bus as a public transport mode, while also showing that its use is strongly conditioned by accessibility, health status, confidence in the journey and the ability to travel autonomously. This reinforced the need to consider the bus stop as part of the accessibility chain, where clearer and more accessible information can reduce uncertainty and support independent mobility.

*Figure 5. Mobility map for older adults.*

The full methodological design and detailed results of the "learn" phase are available in the UPPER D2.1 deliverable (UPPER Consortium, 2024a).

Copyright © **Esomar** 2026


## Ideate: Co-Creating Accessible Bus Stop Information in Valencia

The "ideate" phase translated the evidence generated during the "learn" phase into potential solutions through participatory design activities. In Valencia, the focus was the accessibility and inclusivity of the bus stop experience, particularly passenger information before the journey. Two complementary methods were applied: serious game sessions with vulnerable users, and a hackathon with stakeholders. The serious game sessions were designed to help vulnerable users describe real-life challenges, and to co-create possible solutions to reduce uncertainty, dependency or exclusion, such as delays, route changes, unexpected service disruptions, overcrowding or transfers with other transport modes. Participants included older adults with visual or hearing impairments, an older person with reduced mobility, a mother with young children and a person with cerebral palsy using a wheelchair (Figure 6).

*Figure 6. Serious game session with vulnerable users at the IBV co-creation lab in Valencia.*

The second participatory activity was a hackathon organised within the UPPER Open Innovation Days, bringing together around 30 young innovators, mobility experts, public transport stakeholders and city representatives, in line with a multi-actor governance approach that connects public, private, civic and academic perspectives (3P Partnership Model, 2024). Participants were asked to imagine an innovative, inclusive and convenient user experience at EMT bus stops, combining mobility infrastructure, data-driven services, human-centred design and connected technologies supporting safety and comfort. To ensure continuity with the previous research, they worked with inputs such as the serious game findings, persona-based mobility maps, existing bus-stop plans, local mobility regulations and open data. The winning concept, NUCLANSA, proposed a user-intuitive, safety-focused and accessible bus stop shelter. The concept included two relevant information innovations: large digital displays (providing real-time information on bus routes, arrival times, connections with other transport modes, traffic conditions, alerts and diversions), and a voice assistance system at shelters equipped with information screens. The full methodological design and detailed results of the "ideate" phase are available in the UPPER D2.3 deliverable (UPPER Consortium, 2024b).

## Validate: In-Context Feedback on Valencia Bus Stop Information Measures

The "validate" phase was designed to close the loop between citizen-generated evidence, implemented mobility measures and user feedback in real service contexts. In Valencia, validation focused on the two information measures implemented by EMT and described in the following sections. Both measures were assessed through short citizen surveys conducted in context, enabling users to evaluate key experience criteria such as clarity, accessibility, usefulness, visibility, ease of understanding and overall satisfaction. As shown in Figure 7, access to the survey was provided directly at the bus stop through a poster with a QR code, allowing passengers to give feedback at the point of experience.

Copyright © **Esomar** 2026

*Figure 7. QR-based citizen survey displayed at the EMT bus stop.*

## Results: From Citizen-Generated Evidence to Public Transport Change

The UPPER case generated two complementary types of results. The first is methodological and is particularly relevant for the insights community: the development of U-GOV as a participatory governance platform that operationalises a citizen science-informed insight methodology. The second is practical and citizen-facing: the implementation of two EMT information measures in Valencia that improve the accessibility and usability of bus stops, especially for vulnerable groups.

### *U-GOV as a Replicable Insight Infrastructure for Participatory Governance*

One of the main results of the project is U-GOV, a digital platform designed to make citizen participation repeatable across the full public transport innovation cycle. Rather than functioning only as a consultation or survey tool, U-GOV integrates the citizen science methodology developed in UPPER into a structured participatory governance system. For the insights community, U-GOV is relevant because it turns citizen participation into an operational process. It enables cities and public transport operators to collect citizen-generated evidence, involve users in co-creation, validate real public transport interventions and communicate results back to the community. As shown in Figure 8, U-GOV makes ongoing mobility actions visible within the city and provides a practical entry point for citizens and stakeholders to engage with them. This helps transform participation from a one-off research activity into a continuous infrastructure for public transport improvement.

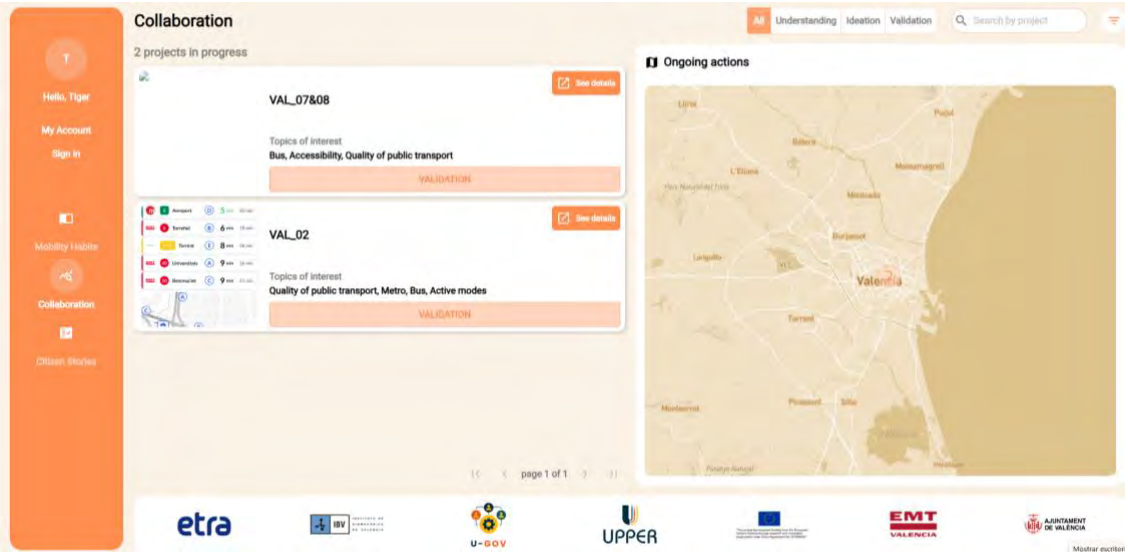



*Figure 8. Ongoing Valencia actions displayed in the U-GOV Collaboration module.*

A detailed description of the U-GOV platform and its functionalities is available in the UPPER D5.1 deliverable (UPPER Consortium, 2025).

### *EMT Valencia Measures: Accessible and Multimodal Information at the Bus Stop*

Following the analysis of the evidence generated through the citizen science methodology, EMT Valencia designed and implemented two complementary measures to improve the bus user experience. Both measures respond to barriers identified across the research process: difficulty accessing clear real-time information, uncertainty while waiting at the stop, limited accessibility for vulnerable users and the need to better connect the bus with other mobility options. The first measure consists of audio information at EMT bus shelters on Avenida Blasco Ibáñez, complementing screen-based information. This measure addresses accessibility barriers by making passenger information available in a non-visual format, supporting greater autonomy and confidence for people with visual impairments, older adults and other users with specific information-processing needs.

The second measure consists of multimodal transport information through digital screens installed on EMT poles at key urban nodes: Estació del Nord, Porta de la Mar, Facultats and Estació del Cabanyal. As shown in Figure 9, these screens strengthen connectivity by helping passengers understand nearby transport options, connections with other public transport modes and relevant service information. This is especially important in locations where citizens need to combine, compare or transfer between transport options.

Copyright © **Esomar** 2026



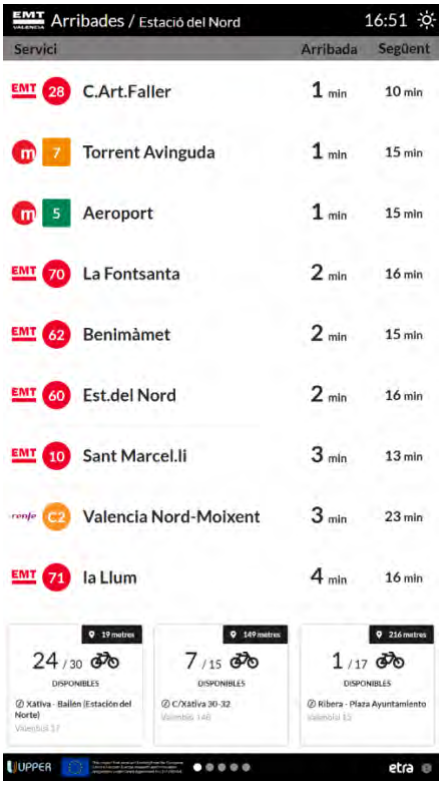



*Figure 9. Multimodal information panel in Valencia showing nearby public transport options.*

Together, these interventions show how citizen-generated evidence can be translated into practical service improvements that make bus travel more accessible, predictable and connected.

## Discussion: Implications for Insight Practice

The UPPER case offers a practical contribution to the insights community by showing how citizen-generated evidence can be organised into a systematic process for decision-making. Its relevance lies not only in the individual methods used, but in how these methods were connected: learning from citizens at scale and in depth, translating evidence into design priorities and validating interventions with users in context. A first implication for insight practice is the value of combining scale, depth and actionability. Netnography provided large-scale evidence of spontaneous user perceptions; experience notebooks added qualitative depth by focusing on the needs, barriers and everyday mobility decisions of vulnerable groups, which are often difficult to detect through scale-based methods alone; and the survey helped validate priorities across a broader population. This triangulated approach strengthened the quality of the evidence and reduced the risk of designing from a single data source.

A second implication is that consequential insight requires a translation layer. Many organisations collect user needs, but fewer convert them into artefacts that decision-makers can use. In UPPER, personas and mobility maps performed this role by transforming complex datasets and individual stories into actionable narratives. They gave mobility stakeholders a shared language for discussing behavioral motivations, accessibility barriers and differences between user groups without reducing people to averages. A third implication is that co-creation should be treated as a bridge between evidence and implementation, not only as an engagement activity. In Valencia, serious games converted vulnerable users' lived experiences into specific information requirements, while the hackathon transformed those requirements into feasible service concepts. This sequence shows how participatory methods can help move from insight interpretation to design direction.

Finally, the U-GOV platform shows how insight practice can become a continuous engagement infrastructure. Rather than relying on isolated studies or occasional consultations, cities and public transport operators can use participatory platforms to embed citizen science into planning,

service design and evaluation, embedding inclusion as a condition for sustainable public transport adoption. This turns participation into an ongoing capability for generating evidence, shaping interventions and learning from implementation.

## Conclusions

The UPPER case illustrates a broader lesson for the insights field: insight has consequence when it is connected to governance, implementation and public value creation. In public transport, citizen science can support this shift when it is linked to insight translation and operational decision-making, rather than treated as an isolated consultation exercise. The methodology demonstrates how triangulated citizen-generated evidence can reveal mobility barriers that traditional performance data often misses, especially for vulnerable and underrepresented groups. In Valencia, the collaboration with EMT shows how this approach can lead to tangible public transport improvements. Insights into accessibility barriers, uncertainty at the stop, multimodal orientation and passenger information needs were translated into two bus stop measures: audio information services at bus shelters and multimodal information panels at strategic urban nodes. These interventions reinforce *mobility as a right* (MaaR) by addressing practical barriers that can prevent people from using public transport safely, confidently and autonomously.

Their relevance goes beyond accessibility as a service-quality issue. In this context, accessibility should also be understood as a strategic lens for behaviour change. Behaviour change is often approached through incentives, restrictions or communication campaigns, but the Valencia case suggests another route: make the sustainable option easier to understand, safer to approach and more usable for people with different capacities. When public transport offers clear information, reliable connections and accessible interchanges, it can better address some of the factors (such as reliability, comfort and perceived flexibility) that often sustain dependence on private vehicles. Ultimately, the case shows how citizen intelligence can transform public transport when it is embedded across the full innovation cycle: from understanding lived mobility experiences, to co-creating solutions, implementing service improvements and validating them with citizens. U-GOV was central to this process because it provided the participatory infrastructure to close the loop between evidence, action and feedback. In this sense, UPPER demonstrates that inclusive and sustainable mobility depends not only on better transport systems, but on better ways of listening to citizens and turning their experiences into decisions, services and measurable public value.

## Acknowledgments

The authors thank all UPPER consortium members for their contribution to results generation, especially the Valencia cluster, including ETRA i+d and EMT. This project received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No [101095904].

## References

Bike Barometer. (2022). *Citizen science for cycling-friendly school routes: A pilot project.* Proceedings of the European Transport Conference.

DiStaso, D., Mulder, I., Janssen, M., and Kleiman, F. (2022). Serious Games For Building Data Capacity. *Interdisciplinary Description of Complex Systems 20*(2), 179-189.

European Commission. (2019). *Communication from the Commission: The European Green Deal*. Available: https://www.consilium.europa.eu/media/47573/st_15051_2019_init_en.pdf

Fortin, C., Neustaedter, C., and Hennessy, K. (2014). *Posting for community and culture: Considerations for the design of interactive digital bulletin boards.* Proceedings of the SIGCHI Conference on Human Factors in Computing Systems.

Hevenstone, D., Endrissat, N., Lavrovsky, O., and Hümbelin, O. (2026). Designing for diversity: Team formation and participatory policy design. *Policy Design and Practice, 9*(2), 232–243. https://doi.org/10.1080/25741292.2025.2610868

Kozinets, R. V. (1998). On Netnography: Initial Reflections on Consumer Research Investigations of Cyberculture. *Advances in Consumer Research*, *25*, 366-371.

Mitropoulos, L., Fotiou, A. M., Kortsari, A., Nikiforiadis, A., and Ayfantopoulou, G. (2025). *Transition to Multimodal Digital Mobility Services (MDMS) and User-Friendly Multimodal Transport Networks.* Proceedings from the 10th Transport Research Arena At Conference, Dublin, Ireland.

Nalle, D., Susanti, R., and Wibowo, A. (2023). Public participation in bus transit policymaking: A municipal case study. *Case Studies on Transport Policy, 11*, 100965.

Pruitt, J. and Adlin, T. (2010). *The persona lifecycle: keeping people in mind throughout product design*. Morgan Kaufmann.

Robinson, L. D., Cawthray, J. L., West, S. E., Bonn, A., and Ansine, J. (2018). Ten principles of citizen science. In S. Hecker, M. Haklay, A. Bowser, Z. Makuch, J. Vogel, and A. Bonn (eds.), *Citizen Science: Innovation in Open Science, Society and Policy*. UCL Press. 27-40.

Sardo, A. M., Laggan, S., Franchois, E. and Fogg-Rogers, L. (2022). Reflecting on deepening participation in recruitment and evaluation in citizen science—lessons from the WeCount project. *Journal for Research and Technology Policy Evaluation, 53*(54), 20–32. doi:10.22163/fteval.2022.568

Sardo, A. M., Weitkamp, E., Teixeira, J., and Ferreira, M. I. A. (2024). Measuring the impacts of participatory citizen science: lessons from the WeCount sustainable mobility project. *Journal of Science Communication, 23*(1), A05.

Sauermann, H., Vohland, K., Antoniou, V., Baláz, B., Göbel, C., Karatzas, K., Mooney, P., Perelló, J., Ponti, M., Samson, R., and Winter, S. (2020). Citizen science and sustainability transitions. *Research Policy, 49*(5), 103978.

Soriano, C., López-Vicente, A., Giménez, J. F., Marzo, R., Solaz, J., and Signes, E. (2024). *Netnography Study to Investigate End Users' Satisfaction with Public Transport in UPPER Project's Living Labs*. Conference Proceedings for Transport Research Arena. Available: https://doi.org/10.1007/978-3-031-85578-8_7.

Sree Yarlagadda, M., and Gupta, N. (2025). Systematic literature review on impact of perceived accessibility of public transportation on user travel behavior. *Sustainable Transport and Livability, 2*(1), 2560816.

UPPER Consortium. (2024a). *D2.1: User groups' mobility needs, motivations and patterns*. Horizon Europe Project Deliverable. Available: https://www.upperprojecteu.eu/resources.

UPPER Consortium. (2024b). *D2.3 Gamified approach for cocreating inclusive PT solutions*. Horizon Europe Project Deliverable. Available: https://www.upperprojecteu.eu/resources.

UPPER Consortium. (2025). *D5.1 U-GOV: A novel platform to ease user participation in PT*. Horizon Europe Project Deliverable. Available: https://www.upperprojecteu.eu/resources.

Vohland, K., Land-Zandstra, A., Ceccaroni, L., Lemmens, R., Perelló, J., Ponti, M., Samson, R., and Wagenknecht, K. (2021). *The Science of Citizen Science*. Springer.

3P Partnership Model. (2024). Multi-actor governance for smart mobility: Integrating public, private, people and academia. *Smart Cities, 7*(2), 456–478.

## About the Authors

Raquel Marzo, Instituto de Biomecánica (IBV), Valencia, Spain.

Josep Chiner, Empresa Municipal de Transports (EMT), Valencia, Spain.

Amparo López-Vicente, Instituto de Biomecánica (IBV), Valencia, Spain.

Carol Soriano, Instituto de Biomecánica (IBV), Valencia, Spain.

Juan F. Giménez, Instituto de Biomecánica (IBV), Valencia, Spain.

Elisa Signes, Instituto de Biomecánica (IBV), Valencia, Spain.

José Solaz, Instituto de Biomecánica (IBV), Valencia, Spain.

Raquel Alario, ETRA I+D, Valencia, Spain.

# Anticipating the Future Consumer

## Using Digital Behavioural Data and AI to Size Global Macro Forces and Detect Emerging Signals

*Emily Dobbing*

*Giulia Iorio-Ndlovu*

Esomar

Office address:

Burgemeester Stramanweg 105

1101 AA, Amsterdam

Phone: +31 20 664 21 41

Email: info@Esomar.org

Website: www.Esomar.org

Publication Date: September 2026

Esomar Publication Series Volume S424 Congress 2026

ISBN 92-831-0337-8

Copyright © **Esomar** 2026

Copyright © **Esomar** 2026

Congress 2026

# Anticipating the Future Consumer

## Using Digital Behavioural Data and AI to Size Global Macro Forces and Detect Emerging Signals

*Emily Dobbing*

*Giulia Iorio-Ndlovu*

### Listening to Tomorrow, Today

*How BAT and Curve Analytics Built a New Foresight Engine*

*Legal disclaimer: All references to consumer are targeted towards adult nicotine/tobacco consumers only.*

BAT has partnered with Curve Analytics to revolutionise their approach to studying ex-category macro trends for foresights. It is now a foresights engine grounded in 2 billion data points, tracking macro forces across 18 markets, surfacing not just what's gaining momentum today, but also the emerging signals bubbling at the edge of discourse. The outputs now inform scenario planning, category strategy, innovation pipelines and brand communications for one of the world's largest consumer goods organisations. This is not simply a study, it is a new methodology—and this paper explains why it was needed, how it was built and what it changed.

### The Need for Foresight

Many large organisations reach a point where strategic thinking begins to erode. The symptoms are recognisable: decisions become reactive rather than anticipatory; strategy cycles shorten as the business chases the present rather than preparing for the future; and more energy is spent catching up than shaping direction. Innovation, starved of genuine foresight, defaults either to optimising what already exists or chasing novelty without strategic grounding. The result is incremental activity without transformation. BAT's leadership identified this risk clearly. For a global organisation operating across dozens of markets and planning years ahead, in categories defined by evolving consumer attitudes to pleasure, identity and lifestyle, the ability to anticipate cultural and behavioural change is not a competitive advantage. It is the foundation of strategy. BAT's cross-category foresights team articulated five specific reasons why understanding ex-category macro trends is essential for business growth.

1. Staying relevant in consumers' lives: Consumer obsession is the best growth strategy. By tracking how macro trends evolve, we can anticipate meaningful future needs and align brand strategy with the cultural shifts that matter most to consumers.
2. Prioritise commercial decisions during uncertain realities: Consumers are simultaneously navigating economic instability, technological disruption and shifting social norms. Sizing trends and detecting emerging themes provides the analytical fuel to focus resources on what matters, and drive growth when the path ahead is unclear.
3. Strengthen local deployment: Macro trends manifest differently across cultures and markets. Understanding those in-market expressions of broader forces enables regional teams to build activation plans that are genuinely locally relevant rather than globally generic.
4. Maintain competitive advantage by distinguishing the stable from the disruptive: Understanding which forces are long-term and structural, versus which represent emerging disruption allows the business to allocate innovation and brand investment with greater confidence.
5. Establish a global source of truth: A shared trends framework, socialised across functions, regions and planning cycles creates coherence across a large, distributed organisation. One voice on the consumer, grounding every category and every market.

Copyright © **Esomar** 2026


"This ex-category macro trends study has allowed us to stay relevant in our adult consumer lives, prioritise commercial direction in uncertain times, strengthen local deployment initiatives, maintain competitive advantage and provide a unique source of global truth", (Giulia Iorio-Ndlovu, Global Head of Cross Category Foresights, BAT).

## The Challenge

BAT's foresight agenda required a new capability to understand what is shaping the broader lives of consumers, outside of BAT's categories. Ex-category macro trends—forces like pleasure, connection and sustainability—are the upstream cultural currents that eventually reshape how consumers think about every category they engage with, including BAT's. Traditional research approaches were not built for this task. They are valuable for understanding attitudes, preferences and intentions within defined product or category contexts. However, they face structural limitations when applied to macro cultural forces:

- Scale: Surveys and qualitative panels cannot capture the billions of everyday behaviours that cumulatively define a cultural trend;
- Speed: Periodic research cycles cannot keep pace with the continuous evolution of digital conversation;
- Geographic breadth: Achieving genuine comparability across 18 markets with consistent methodology is extremely difficult through conventional approaches;
- Researcher bias: Guided interviews inevitably reflect the hypotheses of the people designing them, limiting the ability to discover genuinely unexpected signals;
- Depth of future signal: Mainstream research tells you about what is already visible, but rarely shows you the emerging signals of the future.

The signals that will define future behaviour often instead exist at the edges of current discourse (in expert communities, among activists, in niche online spaces) long before they surface in mainstream conversation. Nevertheless, validating this at scale proves difficult with traditional research methods. The challenge, therefore, was two-fold. First, BAT needed a way to measure and understand the scale and growth of macro behavioural forces globally, with enough granularity to act on across markets and functions. Second, it needed a way to identify the emerging signals indicating where those forces would evolve next, before they were visible in mainstream data.

BAT's ex-category macro trends framework has evolved significantly over the years, leading up to this 2026 programme. The pre-2023 approach involved creating a bottom-up trends framework, largely relying on scanning existing global trends reports. Between 2024 and 2025, this was augmented with trend expert input and an early version of quantified digital signal detection. The ambition for 2026 was to move into genuine foresight: a forward-looking, data-powered system that could project where macro forces were heading, not just describe where they were. A new foresights capability was needed—one that combined the scale of digital behavioural data, the precision of AI-driven analysis and the cultural intelligence of expert listening at scale, into a single, integrated foresight engine.

## Objectives of the Programme

The programme had four clear objectives, spanning both the analytical capability to be built and the business outcomes to be delivered.

1. Quantify macro forces at global and market level: Move beyond qualitative trend descriptions to produce precise, comparable measurements of the size and growth of each macro force across 18 markets.
2. Identify the sub-trends driving macro force growth: Surface the specific conversations and behavioural dynamics within each macro force that are gaining momentum, without imposing a predefined framework onto the data.
3. Detect emerging cultural signals before they enter mainstream discourse: Build a foresight layer that listens to the frontier of expert conversation in order to identify signals shaping future consumer behaviour, ahead of their appearance in mainstream data.
4. Operationalise foresight across the business: Deliver outputs that are not only analytically rigorous, but directly usable by global and market teams across scenario planning, category strategy, innovation and brand communications.

Copyright © **Esomar** 2026


A fifth, implicit objective ran through the programme: to demonstrate that digital behavioural data and AI-driven analysis could move the standard of foresight research forward; i.e., to move it forward in a way that was genuinely scalable, continuously updatable and grounded in real consumer behaviour rather than stated opinion.

## Deliverables and Findings

### *Outputs Designed for Action*

A sophisticated analytical system delivers value only if its outputs can be understood and used by the people who need them. From the outset, the programme was designed with non-analyst stakeholders in mind. The deliverables were therefore structured to bridge the gap between analytical rigour and business usability:

- An interactive dashboard allowing global and market teams to explore every macro force, sub-trend, market and metric. Users can slice by geography, filter by growth rate and navigate from macro force level to the specific sub-trends driving each macro trend.
- Visual insight reports for market teams that bring each macro force to life through real consumer posts, searches and social content—grounding the quantitative analysis in tangible human expression and making the strategic implications intuitive.
- Market comparison views highlighting where trends are accelerating, plateauing or emerging differently across regions. These enable global teams to understand both global context and regional specificity—the same analytical architecture, but a locally relevant story.
- Emerging signals summaries from the foresights workstream, presenting the signals gaining momentum in expert conversation alongside their growth metrics and the expert communities in which they are most active.

### *Sample Findings*

The programme has surfaced a range of significant findings. While the full dataset is commercially confidential, the following examples illustrate the kinds of insights the methodology delivers.

### *Sustainability 3.0 Macro Trend*

The “Sustainability 3.0” macro trend generated 257 million searches and 1.7 million social posts across 2025, across the eight priority markets alone. Sub-trends such as sustainable living and waste reduction contribute most to the size of the trend, and these are largely driven by cost-of-living drivers rather than environmental conviction. Whereas the growing sub-trends are clean labels and biodiversity protection, with conversation growing at over 10% and over 7%, respectively, year-on-year across digital channels. This is different to the previous year where resource conservation and green tech were the fastest growing trends. However, looking at the market level metrics shows that green tech remains particularly important for Japan where it is the biggest sub-trend. When narrowing in on online conversations amongst experts, the fastest emerging signal is on climate migration, growing at 46% over a six-month period across the expert pool; it is not big today, at only 10,000 12-month mentions, but a very likely future scenario we will find ourselves in at the same time. Already in some pockets of the world, communities are temporarily displaced by the extreme weather events caused by climate change. However these temporary displacements could become permanent in the future, and we may see a wave of climate migration reshaping populations around the world. This signal is strongest in Indonesia, where we see more than 353% six-month growth in expert conversation. This is in response to the intense flooding seen in Indonesia, with experts framing recent disasters as a tipping point for serious climate adaptation and pushing for better preparedness globally.

As climate commitments quieten across the world, sustainability becomes a way for brands to differentiate and reach engaged consumers. Furthermore, businesses should consider that mass climate migration will lead to densely populated “climate secure” locations, stretching resources and shifting markets.

### *Techno Wonder Macro Trend*

Copyright © **Esomar** 2026


The "techno wonder" trend is the fastest growing macro trend globally, growing at 35% year-on-year in the eight priority markets. It's growing particularly fast in markets such as Indonesia (+99%), Sweden (+57%) and the US (+48%). The sub-trend of AI integration into everyday personal decision-making is growing at a rate that significantly outpaces broader AI awareness conversations, suggesting that consumer comfort with AI assistance is advancing faster than public discourse might suggest. However, on the other hand, the second fastest growing sub-trend is "anti-AI", with scepticism and negative opinions of AI growing at more than 53% year-on-year. It is therefore unsurprising that the fastest-growing emerging signal from expert conversation is a deepening nervousness about what a "machine first logic" approach to life might ultimately cost us. AI increasingly structures our decisions and shapes our agency, yet we continue to outsource our thinking to the very same systems we are anxious about. This tension is gaining significant momentum; expert conversation around these concerns grew by 671% in just six months.

The specific fears, however, vary by market. In Western countries, the dominant anxiety is cognitive offloading; the worry that by delegating thinking to AI, future generations will simply never develop the skills that come from doing it themselves. Critical thinking, creativity, problem solving—these are not just outputs of education, they are products of the struggle to learn. If that struggle is removed, the concern is that something fundamental is lost. In Japan and South Korea, the anxiety takes a different form. Here the conversation is less about what AI does to the mind and more about what it does to the economy. AI companies are scaling rapidly on enormous valuations, with limited returns to show for it. A growing body of expert opinion is asking whether the technology is catastrophically overvalued, as well as what happens when that bubble meets reality. These are early warning signals of what's to come; the AI bubble could lead to lower disposable income, whilst AI agents will automate purchase decisions meaning brands must play to a machine-led purchase funnel ensuring positive visibility in prompt responses and algorithms.

## Why a Digital Data Approach

The rapid expansion of digital behaviour has fundamentally changed the landscape of research. Today, billions of searches and social media posts provide an unprompted, real-time record of how consumers think, feel and act. These signals reveal what people are curious about, what they are discussing, how their interests evolve and, crucially, where cultural momentum is building. This represents a fundamental shift in what is possible for foresight research. Digital behavioural data offers something structurally different to traditional research methods. It is behavioural rather than attitudinal, derived from real actions rather than survey responses. It is unprompted, free from interviewer effects or question-framing bias. It is global in scale, measurable with the same method across 18 markets simultaneously. Moreover, it is quantifiable; cultural forces that have historically been described only qualitatively can now be sized, ranked and tracked with precise metrics. This method focussed on two complementary analytical phases:

1. Insights: Understanding the now;
2. Foresights: Uncovering the future.

## Insights Method: Understanding the Now

### *The Dataset*

The first workstream set out to quantify the seven macro forces defined by BAT's product foresights team, and to surface the sub-trends driving each one at global and market level. The dataset was assembled from two sources of digital behavioural data across 18 markets: the US, UK, Brazil, Italy, Sweden, Japan, South Korea, Indonesia, Romania, Ukraine, Serbia, Turkey, Croatia, Kazakhstan, Pakistan, Saudi Arabia, Uzbekistan and South Africa. The markets were selected to provide genuine global coverage, spanning Western and non-Western contexts, high-income and emerging economies and a range of cultural traditions. For this insights phase of the programme, it was important to study mass consumer conversation, to truly size these big forces and understand the granular detail within each. This mean analysing a very large dataset:

- 2 billion search queries analysed;
- 46.8 million social media posts processed;
- 18 markets covered;

- Five years of data, enabling both short and long-term trend measurement.

### *The Analytical Engine: AI Topic Discovery*

The analytical engine was Curve's proprietary AI topic discovery model, applied through a large language model pipeline. The model analyses unstructured text data at scale and surfaces topic clusters organically, identifying the themes that are genuinely present in the data rather than those the analyst anticipated in advance. This methodology has a critical advantage over keyword-based approaches. When researchers define keywords in advance, they can only measure the ideas they have already thought of. Organic AI topic discovery allows the data to surface sub-trends that no brief, no expert panel and no questionnaire would have identified. Evolving attitudes toward loneliness and disconnection as a cultural crisis, a growing backlash against self-optimisation, the emergence of AI companionship—these are the kinds of dynamics that emerge from the data, not from the researcher's starting hypotheses. The process followed three stages: first, ingest the raw unstructured dataset of search queries and social posts across all 18 markets. Second, apply the AI topic discovery model to surface meaningful topic clusters within each macro force domain, without imposing any ingoing category framework. Third, measure each identified topic on size and growth metrics to produce a structured, comparable quantitative dataset.

### *Growth Metrics*

Once topics were identified, three metrics were calculated for each sub-trend, at both global and individual market level:

- Conversation size: The total volume of digital discussion associated with a trend, providing a measure of current scale and cultural prominence;
- Year-on-year growth: The percentage change in conversation volume over the most recent two years, capturing current momentum;
- Five-year CAGR: The compound annual growth rate over five years, revealing whether a trend is a recent spike or a sustained, structural shift.

These three metrics in combination allow team members to distinguish between trends that are large but decelerating, trends that are small but growing rapidly and trends that show consistent compound growth over time. This granularity is what makes the analysis actionable: it moves strategy teams from "this trend exists" to "this specific dynamic is driving growth at this rate in this market right now".

### *Geographic Granularity*

A critical design feature of the analytical architecture is that every metric is calculated at both global and individual market level. This enables two levels of strategic application. Global teams can work with a single, consistent view of which macro forces are dominant and which are accelerating across the full dataset. Market teams can explore how their specific context differs: which sub-trends are unusually prominent locally, where a force is ahead of or behind the global trajectory and which emerging dynamics are culturally specific to their market. The ability to run this kind of granular, cross-market comparison using an identical methodology is one of the most distinctive features of the programme. Traditional research at this scale would potentially involve different research providers, different methodologies and different data sources across markets, making cross-market comparison impressionistic at best. Here, the comparison is analytically valid.

## Foresights Method: Understanding the Future

### *The Insight Behind the Foresight Method*

Mainstream conversation tells you what is already happening. By the time a macro force has significant size within mainstream search and social data, it is no longer an emerging foresight—it is a present reality. For foresight teams whose job is to help the business anticipate the future rather than react to the present, this temporal limitation is a fundamental problem. The insight at the heart of the second workstream is simple but powerful: ideas diffuse. Cultural change does not emerge simultaneously across a population. It starts in smaller communities (among researchers, activists, specialists, journalists and cultural creators) whose thinking gradually permeates wider discourse over months and years. If you can identify what those frontier voices are discussing today, you gain meaningful advance visibility of what mainstream culture will be

engaging with tomorrow. “The future is already here, it’s just not evenly distributed”. This observation, widely attributed to William Gibson, is the methodological premise of the second workstream.

### *Reinventing the Expert Panel*

Traditional expert panels and Delphi methods have existed in research for decades. They work by soliciting structured input from a selected group of experts, often through surveys or guided interviews. The approach has real value, but it carries inherent limitations: it is slow, expensive, constrained by the questions the researcher chooses to ask and dependent on the willingness of experts to participate and respond thoughtfully. Curve’s approach represents a fundamental reinvention of this model. Rather than asking experts anything, the team built a system to listen to what they were already saying, at scale and completely unprompted. The shift from surveying to listening is not merely a methodological nuance. It removes the questionnaire effect entirely, captures the experts’ authentic, spontaneous expression of ideas and enables analysis at a scale that no traditional expert panel could match.

### *Curating the Expert Panel*

The team curated a panel of over 1,000 digital experts across seven markets: the US, UK, Brazil, Italy, Sweden, Japan and South Korea. These individuals were selected because their ideas demonstrably diffuse into mainstream discourse. The panel included academics and frontier researchers; entrepreneurs and founders whose perspectives on the future of business shape wider commercial thinking; publishers, specialist journalists and media brands; influencers and cultural creators whose content shapes lifestyle trends and identity narratives; and activists and advocacy voices whose perspectives on systemic issues often precede their entry into mainstream debate.

### *The Analytical Method*

Social posts from a wider pool of “influential voices” were passed through an AI topic discovery pipeline. The model surfaced emerging themes within expert conversation, and quantified them in the influential voices pool, to discover prominent emerging topics that were not yet visible in mainstream data, identified by their cross-expert uptake, growth rate and the structural characteristics suggesting they were gaining rather than losing momentum.

- More than 450,000 influential voices monitored across geographies;
- Seven markets in the expert listening panel.

The metrics applied to expert conversation differed slightly from the insights workstream to reflect the different analytical objective. Rather than measuring long-term CAGR, the foresights metrics focused on six-month growth rates, conversation engagement (a proxy for how actively an idea is being debated and shared) and the breadth of expert uptake across different expert categories and geographies. Emerging signals that scored strongly on multiple dimensions (rapidly growing, broadly shared across expert communities and appearing across multiple geographies simultaneously) were flagged as high-confidence signals of future cultural momentum.

### *Validation and Limitations*

Validation for emerging signals is necessarily limited. By definition, a genuine early-stage signal cannot be confirmed against mainstream data, because if it were already visible there, it would no longer be emerging. The validation approach taken in this programme therefore focused on structural indicators rather than volume. Signals were treated as high-confidence when they demonstrated consistent growth across multiple independent expert communities, appeared across more than one geography simultaneously and showed rising engagement—suggesting active debate rather than passing mention. As with any research technique, there are limitations and biases in existence. The method identifies signals with the structural characteristics of ideas that tend to travel; it does not guarantee they will. Some signals could plateau or dissipate before

reaching mainstream relevance, and in communities that are deliberately private, conversations will not be surfaced. These are known limitations that the programme continues to address through ongoing expansion of the expert panel and data sources.

## Impact and Learnings

### *Impact*

Fundamentally, this approach is a new way to listen to the future. Globally it is used for brand positioning to make sure that messaging is relevant and remains relevant. It is also used as a starting point for all BAT's innovations from a regional and local perspective. Furthermore, local teams use it to tailor activations, because they can understand what is relevant and what is culturally nuanced in their markets. Ultimately, foresight has been operationalised across the business.

### *Key learnings and implications for the insights community*

Building and operating this capability has generated a set of transferable learnings, both for the BAT and Curve team, and for the wider research and insights community.

- Look above category for genuine foresight: The most valuable macro trend signals rarely come from within a product category. They come from the broader cultural, technological and societal forces reshaping how consumers think about their lives. Teams focused exclusively on category-level data will consistently lag behind those who integrate ex-category signals into their planning.
- Let the data speak before you do: Most research teams approach digital data with a hypothesis, a keyword list, a topic taxonomy, a set of themes to measure. This is limiting. The power of AI topic discovery lies in its ability to surface what you did not know to look for. The most significant findings in this programme were sub-trends that no brief had anticipated. Allowing the data to lead is a methodological commitment, not just a technical choice—and it is what separates genuine discovery from confirmation of existing beliefs.
- Listen to experts rather than survey them: Traditional expert panels capture only what the researcher thought to ask and only what the expert is willing to say in a formal setting. Listening to what experts are already saying unprompted captures authentic thinking at scale. The signal quality is higher, the process is continuous and it opens a window into future mainstream conversation that structured questioning cannot reach.
- Insights and foresights are not the same thing and should not use the same methods: Insights tell you about the present. Foresights require methods specifically designed to identify emerging signals. The industry has historically blurred this distinction by presenting trend reports based on current data as foresight, when analytically they are sophisticated descriptions of the now. Genuine foresight requires a dedicated methodological layer built around early signal detection and the dynamics of how ideas diffuse through culture.
- Quantifying culture is not only possible—it is necessary: Cultural and societal forces have long been treated as inherently qualitative. This programme demonstrates they can be sized, ranked and tracked with the same precision as any other business metric. For insight teams seeking greater strategic influence within their organisations, the ability to present cultural dynamics in quantitative terms is a significant step toward the strategic table.
- Actionability must be built in from the start: The most analytically sophisticated system delivers no value if its outputs cannot be understood by the people who need to act on them. Designing for non-analyst users (through dashboards, visualisations and clear framing) is not a finishing touch. It is what converts a research project into a strategic capability.

It is worth noting that this approach is strictly above category. By design, it looks outward from the business rather than inward, seeking signals from the broader consumer world rather than from within BAT's own categories. This makes it a methodology applicable across industries and categories: the macro forces it tracks are transversal by nature and the trends shaping BAT's consumers are the same forces reshaping consumers across industries. The breadth of that reach can be illustrated through real-life examples from brands and organisations tapping into these macro trends. Relating to the Pleasure Payback trend, McDonald's launched a limited-edition McNugget Caviar kit for Valentine's Day, inspired by a

viral $100 dish from the 2025 US Open, making indulgent, aspirational pleasure accessible at scale. Within the sustainability 3.0 trend, brands like Ohoskin are rewriting the rules of luxury entirely, producing materials that look and feel like leather from orange peel. And within the tribal societies trend, UK youth activist group Bite Back secured billboards across London not to advertise products, but to physically block junk food adverts, turning public space itself into a form of advocacy.

## Conclusion

The partnership between Curve Analytics and BAT has produced something genuinely new: a research methodology that combines the scale of digital behavioural data, the precision of AI-driven analysis and the cultural intelligence of expert listening into a single, continuously updatable foresight system. It is more accurate than traditional approaches because it measures behaviour rather than attitudes. It is more efficient because it processes billions of data points without manual coding. It is more timely because it is continuously refreshable. Additionally, it is more actionable because it is embedded in the strategic workflows of a global business. For the research community, the implication is clear: the tools now exist to move macro trend and foresight research onto an entirely different analytical plane. The question is whether we choose to use them.

Two billion data points. 18 markets. One capability that lets a global business listen to tomorrow, today.

## About the Authors

Emily Dobbing, Partner, Head of Growth at Curve Analytics, UK.
Giulia Iorio-Ndlovu, Global Head of Cross-Category Foresights at BAT, UK.

esomar

# Where Fraud Meets Its Friction

## Preventing Survey Fraud at Reward Redemption

*Frank Kelly*

*Justin Hill*

*Mardi Hartzog*

Esomar

Office address:

Burgemeester Stramanweg 105

1101 AA, Amsterdam

Phone: +31 20 664 21 41

Email: info@Esomar.org

Website: www.Esomar.org

Publication Date: September 2026

Esomar Publication Series Volume S424 Congress 2026

ISBN 92-831-0337-8

Copyright © **Esomar** 2026

Copyright © **Esomar** 2026

Congress 2026

# Where Fraud Meets Its Friction

## Preventing Survey Fraud at Reward Redemption

*Frank Kelly*

*Justin Hill*

*Mardi Hartzog*

## Introduction

### *The New Economics of Trust*

Online market research has always depended upon a simple but essential assumption: the individuals providing data are genuine participants who meet the requirements of the study and respond honestly. For decades, this assumption was largely supported through panel management practices, respondent profiling and relatively modest incentive values. Fraud certainly existed, but the effort required to create and maintain fraudulent identities often outweighed the financial rewards. That balance has changed dramatically. The expansion of digital incentives, the globalisation of online research and the emergence of sophisticated automation tools have fundamentally altered the economics of respondent fraud. Organised fraud networks, automated account creation, residential proxy services, device spoofing and, more recently, generative artificial intelligence have significantly reduced the cost and complexity of impersonating legitimate research participants. At the same time, higher-value studies, AI training projects and specialised research panels have increased the potential financial rewards available to fraudulent actors. The result is an escalating technological arms race between those attempting to compromise research quality and those seeking to protect it.

In response, the industry has adopted an increasingly diverse range of authentication technologies. Passive device intelligence, browser fingerprinting, IP reputation analysis, behavioural analytics, proxy detection, email intelligence, knowledge-based authentication, biometric verification and government-issued identity verification have become common components of online research operations. These technologies have substantially improved the industry's ability to identify suspicious respondents and reduce fraudulent participation. However, authentication has become more than a technical problem. It has become a methodological one. Every authentication requirement influences the respondent experience. Each additional verification step requires legitimate participants to invest more time, share more personal information or tolerate greater inconvenience before receiving an incentive. While individually these requirements may appear minor, collectively they create a growing burden that may discourage participation, increase abandonment, reduce respondent satisfaction and ultimately alter the composition of research samples. This paper refers to that cumulative burden as the *Trust Tax*.

Like any tax, the *Trust Tax* is not inherently undesirable. Confidence in research requires evidence, and evidence often requires effort. The challenge is not to eliminate authentication, but to ensure that legitimate respondents pay no more than is necessary to establish sufficient confidence. Excessive authentication imposes unnecessary costs on respondents, while insufficient authentication undermines the credibility of research findings. The objective is, therefore, optimisation rather than maximisation. This paper argues that current approaches to respondent authentication remain largely technology-centric. Organisations frequently compare individual authentication tools, evaluate fraud detection rates or assess the accuracy of specific verification methods. While valuable, these comparisons often overlook the broader objective of authentication itself.

Researchers do not seek authentication for its own sake, they seek confidence. Confidence is the condition that allows organisations to trust research findings, make business decisions and invest resources based upon the data collected. Authentication technologies are simply one

means of generating that confidence. Recognising this distinction leads to a fundamental shift in perspective. Instead of viewing authentication as a collection of isolated technologies, this paper proposes that it should be treated as a systematic design discipline whose purpose is to optimize confidence while balancing respondent burden, operational efficiency and representativeness. We refer to this discipline as *Confidence Engineering*., which provides the conceptual foundation for the methodology presented throughout this paper. Building upon this philosophy, we introduce the *Comprehensive Respondent Authentication Framework*, a structured approach that integrates multiple sources of evidence throughout the respondent journey rather than relying on any single authentication event. The framework is guided by four central ideas:

1. Confidence engineering, which defines the objective of respondent authentication;
2. The *Comprehensive Respondent Authentication Framework*, which provides the methodology for designing authentication systems;
3. The trust tax, which recognises the cumulative burden imposed on legitimate respondents;
4. Authentication-induced selection bias, which highlights the potential impact of authentication on sample representativeness.

Together, these concepts provide a new way of thinking about respondent authentication—not as a race to deploy increasingly sophisticated technologies, but as a methodology for engineering confidence in an increasingly complex digital research environment.

## The Evolution of Respondent Authentication

### *From Trust-Based Panels to Confidence Engineering*

Respondent authentication has become one of the defining methodological challenges facing online market research. The rapid growth of digital data collection, global respondent panels and electronic incentive payments has significantly increased both the opportunities for fraudulent participation and the sophistication of those seeking to exploit research systems. In response, the industry has invested heavily in technologies designed to distinguish legitimate respondents from fraudulent actors. While considerable attention has been given to individual authentication technologies, relatively little attention has been paid to the broader evolution of respondent authentication as a research methodology. Authentication has not developed through the implementation of a single technology or standard. Rather, it has evolved incrementally over several decades as researchers adapted to changing technologies, respondent behaviours and fraud risks.

Understanding this evolution is important because it reveals a consistent pattern. Each generation of authentication has addressed the dominant risks of its time, but each has also introduced new limitations that required additional solutions. The result has been a gradual shift from simple trust-based participant management toward increasingly sophisticated evidence-based confidence systems. This section traces that evolution and demonstrates why the next stage of development requires a transition from technology-focused authentication to confidence engineering.

### *Stage One: Trust Through Recruitment*

The earliest online research panels relied primarily on trust established during recruitment. Respondents typically joined panels through controlled recruitment processes, referrals, customer databases or probability-based sampling. Panel providers invested considerable effort in profiling participants, maintaining contact information, monitoring participation history and removing individuals who consistently demonstrated poor-quality behaviour. Authentication during this period was relatively simple because participation itself represented a significant investment of time. Fraud existed, but the financial incentives available to respondents were generally modest, and creating multiple credible identities required substantial effort. Researchers therefore placed significant confidence in panel membership itself. Identity was inferred from panel tenure, participation history, demographic consistency and long-term engagement rather than through explicit identity verification. Quality assurance focused primarily on survey behaviour rather than respondent identity. Typical controls included:

- Duplicate email detection
- Participation frequency limits
- Manual panel management
- Demographic profile consistency
- Interview duration

- Straight-lining
- Open-end review

The underlying assumption was that reputable panel providers served as trusted intermediaries between researchers and respondents. For many years this assumption proved largely sufficient.

### *Stage Two: Trust Through Rules*

As online research expanded globally, respondent pools became larger, more anonymous and increasingly distributed across multiple panel providers. The growth of survey routers, affiliate networks and sample exchanges reduced researchers' direct visibility into respondent recruitment and management. Simultaneously, digital incentive payments made participation more financially attractive. Fraud increased accordingly. Rather than relying primarily on long-term respondent relationships, organisations increasingly introduced rule-based mechanisms to identify suspicious behaviour. Authentication shifted toward detecting violations of predefined rules. Common examples included:

- Multiple survey attempts
- Duplicate IP addresses
- Cookie detection
- Geographic inconsistencies
- Impossible completion times
- Attention checks
- Hidden quality questions

These methods significantly improved the industry's ability to identify low-quality participation. However, rule-based authentication possessed inherent limitations. Legitimate respondents occasionally triggered quality rules, while sophisticated fraudsters rapidly adapted their behaviour to avoid detection. Authentication became increasingly reactive, with each new fraud pattern prompting additional defensive rules. The emphasis remained on identifying undesirable behaviour rather than measuring confidence directly.

### *Stage Three: Trust Through Passive Intelligence*

The widespread availability of device intelligence and digital fingerprinting technologies marked a significant shift in respondent authentication. Rather than relying solely on survey responses, authentication began incorporating passive information collected throughout the respondent journey. Researchers could now evaluate characteristics such as:

- Device fingerprint consistency
- Browser configuration
- Operating system
- Network characteristics
- Proxy usage
- VPN detection
- Time zone consistency
- Geolocation
- Open network ports
- Anti-detect browsers
- Virtual machines
- Browser automation

Importantly, these signals required little or no additional effort from respondents. Authentication increasingly occurred invisibly. This represented a major methodological improvement, because passive intelligence reduced fraud detection costs while minimising disruption to

legitimate participants. Passive intelligence also fundamentally changed how authentication was performed. Rather than making binary decisions based upon individual rules, organisations began accumulating multiple independent indicators of risk. Confidence became evidence-based rather than rule-based. Although this represented substantial progress, passive intelligence alone remained insufficient for situations involving particularly valuable incentives or elevated fraud risk.

### *Stage Four: Trust Through Identity Verification*

As incentive values increased and fraud techniques became more sophisticated, many organisations introduced explicit identity verification. Government-issued identification, biometric comparison, liveness detection, facial recognition, document authentication and knowledge-based verification became increasingly common, particularly for high-value studies and AI training projects. Identity verification offers an important advantage; it establishes direct evidence that an individual possesses a verified real-world identity. However, identity verification also introduced new methodological considerations. Unlike passive intelligence, identity verification requires active participation. Respondents must invest additional time. They must provide sensitive personal information. They must trust organisations to protect that information appropriately. Some respondents may lack compatible technology. Others may decline participation because of privacy concerns. Consequently, identity verification represents both a powerful authentication tool and a potentially significant source of respondent burden. This distinction becomes particularly important when identity verification is applied universally rather than selectively.

While universal verification may increase confidence, it also increases the *Trust Tax* paid by every legitimate participant, regardless of actual fraud risk. The question therefore shifts from: “Can identity verification improve confidence?" to: “When does identity verification provide sufficient additional confidence to justify the additional *Trust Tax*?" This question represents one of the central methodological challenges addressed throughout this paper.

### *Stage Five: Confidence Engineering*

The evolution of respondent authentication naturally leads to a broader conclusion. Each previous stage focused on improving individual authentication techniques. None addressed the overall design of the authentication system itself. Modern authentication systems rarely rely upon a single technology. Instead, they combine numerous independent sources of evidence, including:

- Historical participation
- Panel tenure
- Behavioural indicators
- Device intelligence
- Network analysis
- Email reputation
- Passive fraud signals
- Identity verification
- Manual review

The challenge is no longer identifying which individual technology performs best. The challenge is determining how these technologies should work together to produce sufficient confidence, while minimising unnecessary respondent burden. This requires a fundamentally different perspective. Rather than viewing authentication as a collection of defensive technologies, respondent authentication should be viewed as the systematic design of evidence-based confidence systems. We refer to this methodology as “confidence engineering”. Confidence engineering recognises that authentication is not a single event, but an ongoing process of uncertainty reduction. Each source of evidence contributes incrementally toward establishing confidence. Some evidence may be sufficient on its own. Other situations may require additional verification. The objective is not maximum authentication, but sufficient confidence for the research decision being made.

### *The Emerging Challenge: Authentication-Induced Selection Bias*

The historical evolution of respondent authentication reveals another important methodological consideration. Each successive generation of authentication has generally increased respondent burden. Early authentication required little beyond panel membership. Rule-based systems introduced occasional quality checks. Passive intelligence remained largely invisible to respondents. Identity verification, however, requires active participation and personal disclosure. As authentication becomes increasingly demanding, legitimate respondents may make different participation decisions. Some individuals may abandon studies, while others may decline participation entirely. Still others may avoid research organizations that require extensive verification. If these decisions occur disproportionately among particular demographic or behavioural groups, authentication itself may alter the composition of research samples. This possibility has received relatively little attention within the market research literature despite its potential implications for representativeness. Traditional survey methodology has devoted considerable effort to understanding sampling error, coverage error, nonresponse bias and measurement error. Authentication may introduce an additional source of systematic error that deserves similar consideration. This paper refers to this phenomenon as “authentication-induced selection bias”. While additional empirical research is needed to quantify its magnitude across different populations and research contexts, recognising its existence represents an important step toward evaluating authentication systems more comprehensively.

### *Conclusion*

The evolution of respondent authentication reflects the broader evolution of online research itself. What began as a relationship built primarily on trusted panel management has become an increasingly sophisticated process of collecting and evaluating evidence about respondent authenticity. Each stage of this evolution has improved the industry's ability to protect research quality. Yet, each has also highlighted the limitations of relying on any single authentication technique. Rule-based systems proved reactive. Passive intelligence reduced friction, but could not eliminate uncertainty. Identity verification increased confidence, but also increased respondent burden. The next stage in this evolution is therefore not defined by another technology; it is defined by a methodology. Confidence engineering shifts the focus from individual authentication tools to the design of integrated systems that accumulate evidence, manage uncertainty and balance confidence against respondent burden, operational efficiency and representativeness.

The following section presents the principles that underpin this methodology and introduces the *Comprehensive Respondent Authentication Framework*, providing a structured foundation for designing authentication systems that are both scientifically rigorous and operationally practical.

## Principles of Confidence Engineering

### *Introduction*

The evolution of respondent authentication demonstrates that no single technology can permanently solve the challenges of online identity assurance. Fraud techniques evolve continuously, authentication technologies improve and respondent expectations change alongside broader developments in digital privacy and security. As a result, authentication systems designed around specific technologies inevitably become outdated. What endures are not the technologies themselves, but the principles that guide their application. The purpose of this chapter is to establish those enduring principles. Rather than prescribing individual authentication methods, these principles provide a methodology for designing authentication systems that remain effective as technologies evolve. They distinguish between the objective of authentication and the mechanisms used to achieve it, providing a foundation that is independent of any particular software platform, identity verification provider or fraud detection technology. Together, these principles define the discipline of confidence engineering.

### *Principle 1: Confidence Is the Objective*

The primary objective of respondent authentication is not to verify identity, detect fraud or deploy increasingly sophisticated security technologies. Those activities are means to an end rather than the end itself. The objective is to generate sufficient confidence that research data can be trusted for the decisions it is intended to support. Confidence differs from certainty. In most research settings, absolute certainty regarding respondent identity is neither achievable nor necessary. Instead, researchers seek a level of confidence that is appropriate for the potential risks associated with a particular study. For example, a short customer satisfaction survey with no incentive may require relatively little authentication, while a high-value AI training study involving substantial incentives may justify considerably greater scrutiny. In both situations, the objective remains the same: reducing uncertainty to a level that supports confident decision-making. This distinction shifts authentication from a technology problem to a decision problem. Technologies are evaluated not by their sophistication, but by the extent to which they contribute meaningful evidence toward reducing uncertainty.

### *Principle 2: Confidence Should Be Built Through Multiple Sources of Evidence*

No individual authentication method provides complete assurance. Government-issued identification confirms the existence of a real-world identity, but cannot determine whether that individual is the intended survey participant. Device intelligence identifies suspicious technical characteristics, but cannot distinguish every legitimate privacy-conscious respondent from a fraudulent actor. Behavioural analytics identify unusual response patterns, but cannot explain why those patterns occur. Each source of evidence contributes only part of the overall picture. Confidence therefore should be accumulated from multiple independent observations rather than relying upon a single authentication event. Potential sources of evidence include:

- Historical participation
- Panel tenure
- Survey behaviour
- Device characteristics
- Network intelligence
- Email reputation
- Passive fraud indicators
- Identity verification
- Previous authentication outcomes

Each contributes incrementally toward reducing uncertainty. When considered together, multiple independent observations frequently provide substantially greater confidence than any individual measure alone. Confidence is therefore cumulative rather than binary.

### *Principle 3: Passive Evidence Should Be Exhausted Before Active Verification*

Every authentication activity imposes some degree of burden on respondents. Passive authentication methods collect evidence without requiring additional participant effort. Examples include device intelligence, proxy detection, browser characteristics, IP reputation, participation history and behavioural analytics. Active authentication methods require respondents to invest additional time or disclose additional information. Examples include document verification, biometric validation, knowledge-based authentication or manual identity review. Because passive evidence generally imposes little or no additional burden, it should be exhausted before requesting active verification from respondents. Active authentication should therefore represent escalation rather than default practice. This principle minimises unnecessary respondent friction while preserving the ability to obtain additional confidence when passive evidence remains insufficient.

### *Principle 4: Authentication Should Be Proportionate to Risk*

Not every research project requires the same level of authentication. The level of confidence required should reflect the potential consequences of accepting fraudulent participation. Factors influencing authentication requirements may include:

- Incentive value

- Research sensitivity
- Strategic importance of decisions
- Regulatory requirements
- Historical fraud patterns
- Respondent recruitment source
- Population rarity

Applying identical authentication requirements across every study ignores these differences and frequently imposes unnecessary burden on legitimate respondents. Instead, authentication should increase proportionately as uncertainty and potential consequences increase. This principle mirrors established practices in financial services, cybersecurity and anti-money laundering, where authentication intensity is routinely adjusted according to assessed risk. The objective is not to maximise authentication, but to apply only the level of authentication necessary to achieve sufficient confidence.

### *Principle 5: Confidence Must Be Balanced Against the Trust Tax*

Every authentication activity requires payment. Unlike financial costs, however, much of this payment is made by respondents:

- Time spent completing verification;
- Additional cognitive effort;
- Privacy concerns;
- Technology requirements;
- Delays in receiving incentives.

Collectively, these represent the *Trust Tax*, which is not inherently negative. Just as financial investments generate value, respondent effort generates confidence. However, unnecessary authentication creates unnecessary burden. Poorly designed authentication systems may discourage participation, reduce respondent satisfaction, increase abandonment and ultimately damage long-term panel health. Confidence engineering therefore recognises the *Trust Tax* as a fundamental design consideration rather than an unavoidable side effect. The objective is to maximise the return on every unit of respondent effort invested.

### *Principle 6: Authentication Should Preserve Representativeness*

Survey methodology has long recognised representativeness as a fundamental requirement of high-quality research. Sampling strategies, weighting procedures and recruitment methodologies are all designed to ensure that research participants accurately reflect the populations under study. Authentication systems should be evaluated according to the same standard. Authentication requirements may not affect all respondents equally. Older adults may experience greater difficulty completing digital identity verification. Privacy-sensitive individuals may decline studies requiring government-issued identification. Respondents with limited technology access may encounter barriers unavailable to others. International populations may possess different forms of acceptable identification. If authentication requirements systematically discourage particular groups from participating, authentication itself may influence sample composition. This paper refers to this phenomenon as “authentication-induced selection bias”.

Although additional empirical research is required to quantify its magnitude across different research contexts, authentication systems should be routinely evaluated for their potential effects on representativeness rather than solely on fraud detection performance.

### *From Principles to Methodology*

These six principles collectively establish the philosophical foundation of confidence engineering. Importantly, they do not prescribe specific technologies. Instead, they provide a consistent methodology for evaluating any authentication approach, regardless of how technologies

evolve. Future authentication systems may incorporate decentralized digital identities, behavioral biometrics, AI-assisted verification, digital wallets or technologies not yet conceived. The principles presented here remain applicable because they focus on the objective of authentication rather than its implementation. They encourage researchers to ask fundamental questions:

- Does this authentication measure meaningfully increase confidence?
- Is the additional confidence proportional to the burden imposed?
- Could equivalent confidence be achieved through passive evidence?
- Does the authentication process affect representativeness?
- Does the overall system balance confidence, respondent experience, and operational efficiency?

By answering these questions consistently, organisations can design authentication systems that remain effective despite continual technological change.

### *Conclusion*

Confidence engineering represents a shift from viewing respondent authentication as a collection of defensive technologies to viewing it as a methodology for managing uncertainty. The six principles introduced in this section establish a technology-independent foundation for designing authentication systems that generate confidence while respecting the respondent experience and preserving research quality. Building upon these principles, the next section introduces the *Comprehensive Respondent Authentication Framework*, which translates these principles into an operational methodology, demonstrating how multiple sources of evidence can be integrated throughout the respondent journey to engineer confidence while minimising the *Trust Tax* and reducing the risk of authentication-induced selection bias.

## The Comprehensive Respondent Authentication Framework

### *Engineering Confidence Throughout the Respondent Journey*

The preceding sections established two central propositions. First, the objective of respondent authentication is to generate sufficient confidence for research decision-making rather than simply detecting fraudulent participants. Second, achieving that objective requires a methodology that balances confidence against respondent burden, operational efficiency and sample representativeness. The *Comprehensive Respondent Authentication Framework* (CRAF) provides that methodology. Rather than treating authentication as a single event performed during registration or incentive redemption, the framework views confidence as evidence accumulated throughout the entire respondent lifecycle. Every interaction between a respondent and the research organisation contributes information that either increases or decreases confidence. Authentication therefore becomes a continuous process of uncertainty reduction, rather than a single pass-or-fail decision.

This perspective represents a significant departure from traditional approaches. Most authentication systems evaluate individual technologies independently: identity verification, proxy detection, device fingerprinting, behavioural analytics or survey quality checks. The CRAF instead asks how these independent sources of evidence should work together to produce an appropriate level of confidence while minimising unnecessary burden on legitimate respondents. The framework is technology independent; it does not prescribe specific authentication vendors or software solutions. Instead, it provides a methodology for integrating whatever technologies are available into a coherent decision-making process.

### *The Objective of the Framework*

Every authentication system seeks to answer the same fundamental question: Can this respondent be trusted sufficiently for the research decision being made? Notice that this question differs from asking whether a respondent is "real". Authenticity exists along a continuum rather than as a binary state. A respondent may possess a verified government identity yet still violate study eligibility requirements. Another respondent may not complete formal identity verification, but may demonstrate years of consistent participation, stable device history, trustworthy behavioural patterns and no evidence of fraudulent activity. The framework therefore evaluates confidence rather than identity

alone. Confidence is generated through the accumulation of independent evidence until the remaining uncertainty falls below an acceptable threshold for the specific research application.

### *Confidence as Accumulated Evidence*

The framework is built upon a simple concept: confidence increases as independent evidence accumulates. Every respondent interaction contributes additional information (see Table 1).

| Historical evidence | Behavioural evidence | Technical evidence | Identity evidence |
|---|---|---|---|
| **Length of panel membership** | Survey completion behaviour | Device characteristics | Email verification |
| **Previous study participation** | Response consistency | Browser fingerprint | Mobile verification |
| **Redemption history** | Straight-lining | Proxy detection | Government-issued identification |
| **Consistency of demographic information** | Open-ended response quality | VPN detection | Biometric validation |
| **Historical quality metrics** | Participation timing | Time zone consistency | Digital identity credentials |
| | | Operating system | |
| | | Geolocation | |

*Table 1*

Each category provides a different perspective on respondent authenticity. No single category is sufficient; collectively they produce confidence.

### *Confidence Escalation*

One of the distinguishing features of the framework is that authentication is progressive rather than universal. Instead of applying every available authentication method to every respondent, authentication increases only when uncertainty remains after lower-friction evidence has been evaluated. This process can be viewed as a confidence escalation model.

### *Stage 1. Historical Confidence*

Respondents begin with whatever confidence has already been established through previous interactions. Long-standing panel members with consistent participation histories may already possess substantial confidence before beginning a new study.

### *Stage 2. Passive Confidence*

Passive intelligence evaluates technical characteristics without requiring additional respondent effort. Examples include:

- Device intelligence
- Browser characteristics
- Network reputation
- Proxy detection
- Participation history

If passive evidence produces sufficient confidence, no further authentication may be necessary.

### *Stage 3. Behavioural Confidence*

Survey behaviour provides additional evidence:

- Completion speed

- Consistency
- Navigation
- Response quality

Behaviour either reinforces or weakens previously established confidence.

### *Stage 4. Active Verification*

Only when meaningful uncertainty remains should respondents be asked to complete additional verification activities. These may include:

- SMS verification
- Email confirmation
- Government-issued identification
- Biometric verification
- Manual review

Active verification therefore becomes an exception rather than the default.

### *Minimising the Trust Tax*

Every additional authentication requirement imposes costs upon legitimate respondents. These costs extend beyond time alone. The *Trust Tax* consists of several components (see Table 2).

| Time | Cognitive effort | Privacy cost | Technology cost | Emotional cost |
|---|---|---|---|---|
| **Additional minutes required to complete verification** | Learning unfamiliar authentication procedures | Disclosure of personal information | Camera access | Concern about misuse of personal information |
| | Understanding instructions | Government-issued identification | Mobile devices | Frustration |
| | Correcting failed verification attempts | Biometric information | Document scanning | Reduced trust |
| | | Location data | Software compatibility | Fear of identity theft |

*Table 2*

Each authentication decision should therefore consider not only the confidence gained but also the cumulative *Trust Tax* imposed upon legitimate participants. Confidence engineering seeks to maximise the return on every unit of *Trust Tax* paid.

### *Authentication-induced selection bias*

One of the most important implications of the framework concerns representativeness. Authentication is frequently treated as a neutral process; it is not. Every authentication requirement influences participation. Respondents who are unwilling or unable to complete verification will either abandon participation or avoid future studies entirely. Importantly, these decisions are unlikely to occur randomly. Older respondents may encounter greater technical barriers. Privacy-sensitive respondents may refuse identity verification. Individuals with limited access to compatible technology may fail verification despite being entirely legitimate participants. International respondents may possess documentation that differs from domestic standards. If these participation decisions occur systematically, authentication becomes another source of selection bias. This paper defines this phenomenon as “authentication-induced selection bias”. Unlike traditional nonresponse bias, this form of bias originates from the authentication process itself. Consequently, authentication systems should be evaluated not only according to fraud detection performance but also according to their effects on sample composition.

Copyright © **Esomar** 2026



## *Measuring Framework Performance*

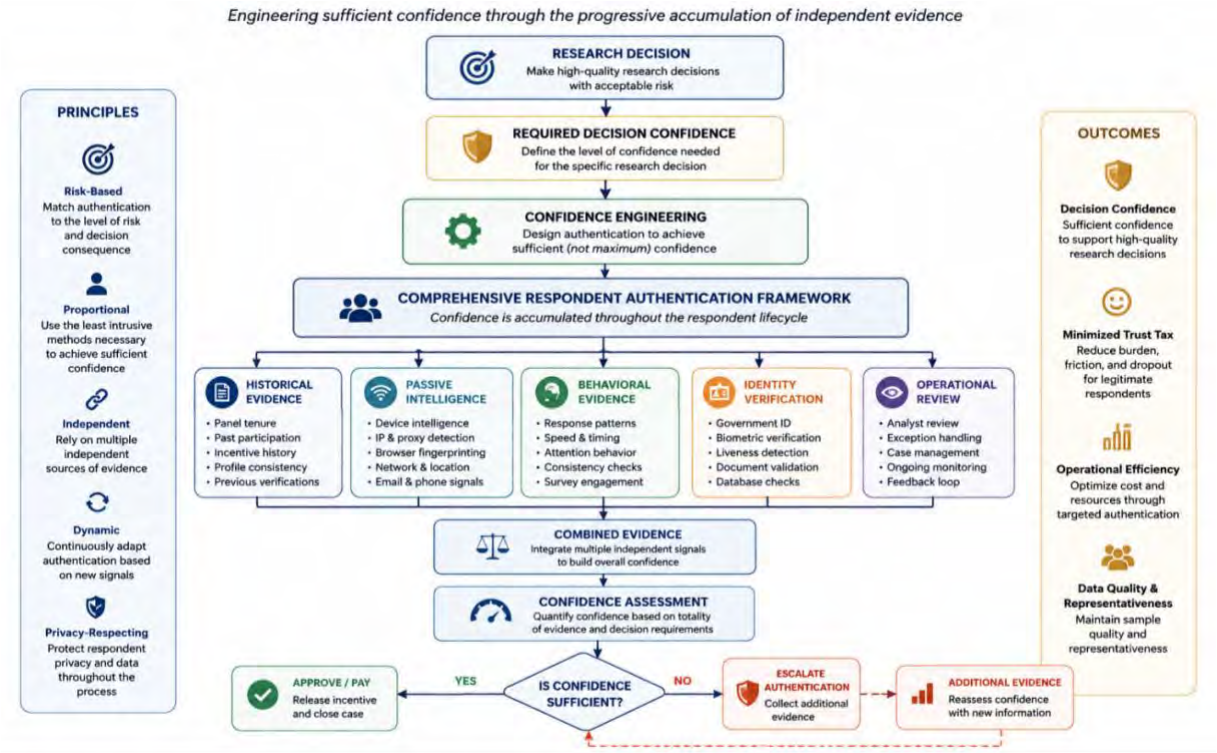
Traditional authentication systems often report technical measures such as:

- Fraud detected
- Identity verification success
- False positives
- False negatives

These metrics remain important, However, they evaluate authentication technologies rather than authentication systems. The CRAF proposes five dimensions for evaluating overall system performance (see Table 3). Collectively, these measures provide a far more comprehensive assessment than fraud detection rates alone.

| Decision confidence | Respondent experience | Operational efficiency | Research quality | Economic performance |
|---|---|---|---|---|
| Does the system generate sufficient confidence for the intended research purpose? | What *Trust Tax* is imposed upon legitimate participants? | How much does authentication cost? | Does authentication improve data quality while preserving representativeness? | What is the total cost of producing a trusted respondent? |
| | | How much manual review is required? | Does it introduce authentication-induced selection bias? | This includes:<br>• Recruitment<br>• Authentication<br>• Fraud losses<br>• Respondent replacement<br>• Operational review |
| | | Can the process scale efficiently? | | |

*Table 3*



Copyright © **Esomar** 2026


*Figure 1. The Comprehensive Respondent Authentication Framework.*

### *Applying the Framework*

Although the framework is technology independent, its application follows a consistent sequence:

1. Establish existing confidence using historical information;
2. Collect passive technical evidence throughout participation;
3. Evaluate respondent behaviour;
4. Estimate remaining uncertainty;
5. Escalate authentication only when justified;
6. Continuously evaluate confidence, *Trust Tax*, operational efficiency and representativeness.

Importantly, confidence is never treated as permanent. Each interaction either reinforces or weakens previous evidence. Confidence therefore evolves continuously throughout the respondent relationship.

### *Conclusion*

The CRAF represents a shift from evaluating authentication technologies individually to designing integrated systems that engineer confidence across the entire respondent journey. The framework recognises that confidence is accumulated rather than verified, that authentication should escalate only when uncertainty justifies additional respondent burden and that authentication systems must be evaluated according to their impact on respondent experience, operational efficiency, research quality and representativeness—not simply their ability to detect fraud. By separating enduring methodological principles from rapidly evolving technologies, the framework provides a foundation for evaluating future authentication systems regardless of how authentication technologies continue to evolve. The following section presents an empirical evaluation of the framework through a controlled four-arm authentication experiment conducted during respondent incentive redemption. The objective of the experiment is not to evaluate identity verification in isolation, but to examine how different authentication strategies influence confidence, respondent behaviour, operational costs and the *Trust Tax* predicted by the framework.

## Empirical Evaluation of the Comprehensive Respondent Authentication Framework

### *Introduction: Applying Risk-Based Authentication During Incentive Redemption*

To evaluate the practical application of the Comprehensive Respondent Authentication Framework (CRAF), a four-arm authentication experiment was conducted during respondent incentive redemption. The study was designed to compare progressively more intrusive authentication strategies while measuring their effects on respondent behaviour, fraud detection, operational efficiency and overall decision confidence. Unlike previous studies that evaluate individual authentication technologies in isolation, this experiment examined authentication as an integrated system. The objective was not simply to determine whether identity verification detects fraudulent participants, but to understand how different authentication strategies influence the balance between confidence and the *Trust Tax* imposed on legitimate respondents.

The experiment was conducted at the point of incentive redemption. This stage of the respondent journey was selected because it represents the moment at which fraudulent actors have the greatest financial motivation while allowing all respondents to complete the research itself. As a result, authentication decisions affected incentive eligibility without influencing survey completion or introducing measurement effects within the questionnaire.

### *Study Design*

Respondents were assigned to one of four authentication strategies. This research was conducted on an established C-Space research community. Each treatment represented a different approach to engineering confidence.

| Group | Authentication strategy | Objective |
|---|---|---|
| 1 | Low-risk control | Passive screening only for respondents classified as low risk. |
| 2 | Passive intelligence | High-risk respondents evaluated using passive fraud signals only. |
| 3 | Risk-based identity verification | Additional identity verification required only when passive evidence indicated uncertainty. |
| 4 | Universal identity verification | Identity verification required for every respondent regardless of observed risk. |

*Table 4. Experimental design.*

This design enabled comparisons between passive and active authentication while also evaluating the value of selective escalation.

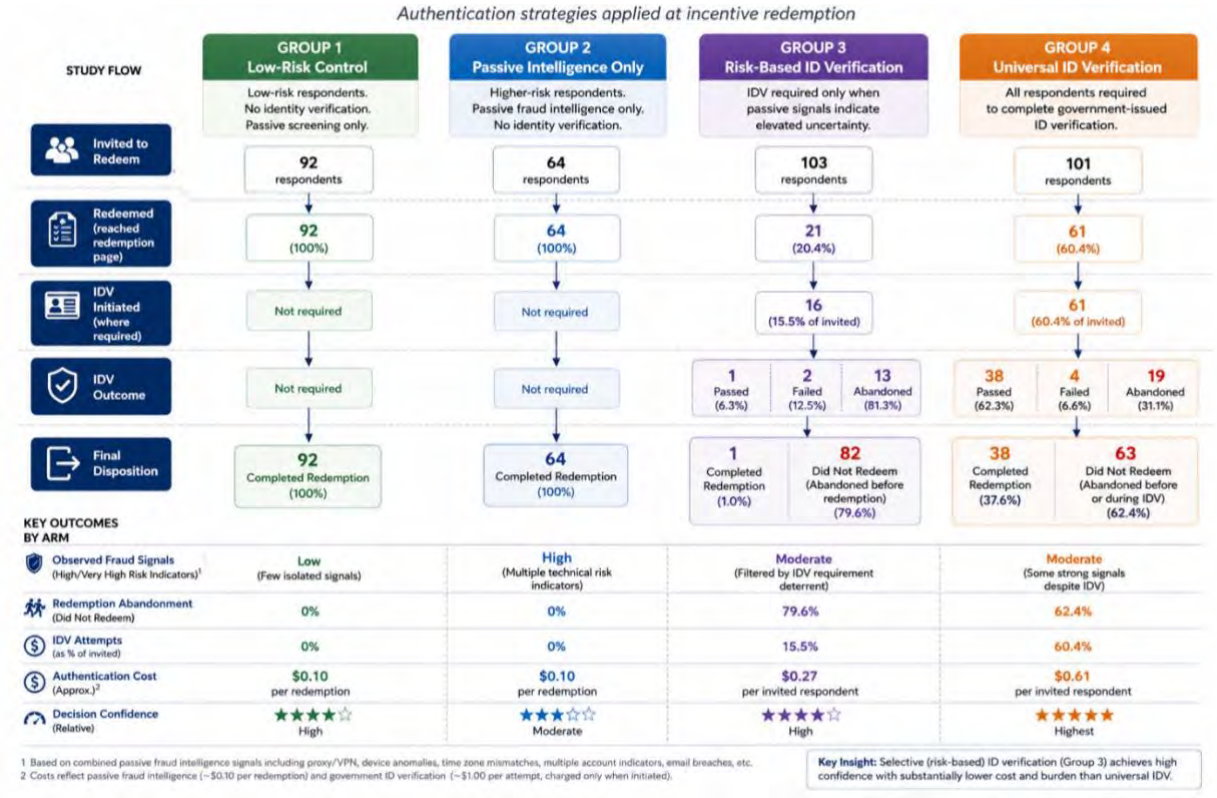



*Figure 2. Five-arm authentication experiment: design and respondent outcomes.*

## *Authentication Measures*

Passive intelligence incorporated multiple independent indicators including:

- Residential, commercial and data center proxy detection
- VPN identification
- Browser fingerprint analysis
- Device consistency
- WebRTC analysis
- Time zone validation
- Email reputation
- Known data breaches
- Open network ports
- Browser spoofing
- Anti-detect browser identification

- Multiple account detection

Government-issued identity verification was introduced only in Groups 3 and 4.

## Experimental Results

### *Group 1. Low-Risk Control*

The control group consisted of respondents who entered redemption with relatively low observed fraud risk. Among 92 completed redemptions, most respondents demonstrated positive identity signals, consistent participation histories and limited technical anomalies. Although occasional proxy use and isolated inconsistencies were observed, few respondents displayed combinations of indicators suggesting elevated fraud risk. These findings support one of the central principles of the CRAF: respondents with substantial historical and passive evidence often require little or no additional authentication to achieve sufficient confidence.

### *Group 2. Passive Intelligence Only*

Group 2 evaluated respondents considered higher risk using passive intelligence alone. Among 64 completed redemptions:

- 51 respondents showed no meaningful online presence or breach history;
- 28 used residential proxies;
- 15 used general proxies;
- Two used data center proxies;
- 17 exhibited one or more open network ports;
- 27 demonstrated time zone inconsistencies;
- 15 showed evidence of browser spoofing;
- 11 triggered multiple-account indicators;
- 10 were identified using anti-detect browsers.

Although these respondents accumulated substantially more technical risk indicators than the control group, passive intelligence alone was unable to determine with certainty whether individual respondents represented sophisticated fraudsters or legitimate participants exhibiting unusual technical characteristics. The findings illustrate an important limitation of passive intelligence. While it substantially reduces uncertainty, it does not eliminate it.

### *Group 3. Risk-Based Identity Verification*

Group 3 implemented selective identity verification. This was a group that had been identified at registration as low quality, so they would not normally have been invited to participate in research. Only respondents whose passive evidence suggested meaningful residual uncertainty were asked to complete government-issued identity verification. Of 103 respondents invited to redeem:

- 82 abandoned the redemption process before claiming payment;
- 21 continued to redemption;
- 16 initiated identity verification;
- One successfully completed identity verification;
- Two failed identity verification.

The remaining respondents abandoned the process before completion.

Perhaps the most interesting observation was not the verification outcomes themselves but the dramatic reduction in redemption attempts. Simply informing respondents that identity verification would be required appeared to deter a substantial proportion of participants from

continuing. This finding suggests that authentication may influence behaviour before verification even begins. Whether these abandoned respondents were predominantly fraudulent actors or legitimate individuals unwilling to complete identity verification cannot be determined conclusively. However, from an operational perspective, the selective requirement substantially reduced the number of identity verification transactions while simultaneously discouraging many potentially suspicious redemption attempts.

### *Group 4. Universal Identity Verification*

The final treatment required every respondent to complete government-issued identity verification prior to redemption.
Among 101 eligible respondents:

- 40 never initiated identity verification;
- 61 began the verification process;
- Across all verification attempts, 38 respondents successfully completed identity verification;
- 4 respondents failed verification;
- 35 respondents abandoned the verification process after starting.

The most striking observation was the overall abandonment rate. Universal identity verification produced the highest level of direct identity evidence, but also imposed the greatest respondent burden. Many respondents simply chose not to continue despite already completing the research. This finding provides direct evidence of the *Trust Tax* associated with universal authentication.

| Authentication strategy | Confidence | Respondent burden | Operational cost | Observed abandonment |
|---|---|---|---|---|
| **Low-risk control** | High | Very low | Very low | Minimal |
| **Passive intelligence** | Moderate to high | Very low | Low | Minimal |
| **Risk-based ID verification** | High | Moderate | Moderate | High |
| **Universal ID verification** | Highest | Highest | Highest | Very high |

*Table 5. Respondent outcomes by authentication strategy.*

### *Operational Economics*

Authentication performance cannot be evaluated solely according to fraud detection. Operational costs also influence the overall efficiency of authentication systems. Passive fraud intelligence cost approximately $0.10 per redemption. Government-issued identity verification cost approximately $1.00 per verification attempt. Because identity verification charges were incurred only when respondents initiated the verification process, selective authentication substantially reduced authentication expenditures compared with universal verification. When recruitment costs, authentication costs, respondent abandonment and fraud prevention are considered together, selective escalation produced a more favourable cost profile than universal identity verification while maintaining a high level of decision confidence.

### *Interpreting the Results Through the Framework*

Viewed independently, each treatment provides useful operational insights. Viewed collectively, they illustrate the principles of the CRAF. First, respondents with substantial historical and passive evidence frequently required no additional authentication to achieve sufficient confidence. Second, passive intelligence effectively reduced uncertainty but occasionally left questions unresolved. Third, selective identity verification proved valuable as an escalation mechanism rather than a universal requirement. Finally, universal identity verification generated the highest level of direct identity evidence but also imposed the greatest *Trust Tax* through increased respondent abandonment, higher operational costs and additional participant effort. The experiment therefore demonstrates that confidence should be engineered progressively. Authentication becomes more effective when additional verification is introduced only after passive methods have been exhausted and meaningful uncertainty remains.

Copyright © **Esomar** 2026


### *Implications for Research Practice*

Several practical implications emerge from these findings. Organisations should view identity verification as one component of a broader authentication strategy rather than as a universal solution. Passive intelligence provides substantial value because it collects evidence without increasing respondent burden. Identity verification is most effective when reserved for respondents whose passive evidence indicates elevated uncertainty. Finally, authentication systems should be evaluated not only according to fraud detection, but also according to respondent abandonment, operational cost and potential impacts on representativeness. Together, these findings provide empirical support for the central proposition of this paper: confidence is most effectively engineered through the progressive accumulation of evidence rather than through universally applied verification requirements.

## Implications for Research Practice

### *From Authentication to Confidence Engineering*

The findings presented in the previous section extend beyond the evaluation of individual authentication technologies. Collectively, they demonstrate that respondent authentication is most effective when viewed as a process of engineering confidence rather than maximising verification. Historically, authentication strategies have often been evaluated according to their ability to detect fraudulent participants. While fraud detection remains an essential objective, the experiment demonstrates that authentication decisions influence a much broader set of research outcomes. Respondent abandonment, operational cost, participant experience and sample representativeness are all affected by the design of the authentication process. The results therefore support a shift in perspective. Rather than asking which authentication technology identifies the greatest number of fraudulent respondents, researchers should ask which combination of evidence produces sufficient confidence for the research decision while imposing the lowest possible *Trust Tax* on legitimate participants. This distinction represents the central contribution of the CRAF.

### *Authentication Should Be Progressive, Not Universal*

One of the clearest findings emerging from the experiment is that authentication should be applied progressively rather than uniformly. Respondents entered the study with very different levels of existing evidence. Some had established histories of legitimate participation and generated few technical risk indicators. Others exhibited combinations of proxy usage, browser spoofing, device inconsistencies or behavioural anomalies that justified additional scrutiny. Treating these respondents identically resulted in unnecessary authentication for many legitimate participants, while providing little additional value for respondents who already possessed substantial evidence supporting their authenticity. Risk-based authentication offers a more efficient alternative. By allowing passive evidence to establish confidence whenever possible and reserving active verification for situations in which meaningful uncertainty remains, organisations can reduce respondent burden while maintaining a high level of confidence in research outcomes. This principle is consistent with established practices in financial services, cybersecurity and fraud management, where authentication requirements routinely increase in proportion to assessed risk.

### *The Trust Tax Should Become a Design Metric*

One of the most significant implications of this research is that respondent burden should be measured as carefully as fraud detection. Authentication is often implemented as though additional verification carries little cost beyond software licensing or operational expense. The experiment suggests otherwise. Introducing identity verification substantially increased abandonment during incentive redemption. Although some abandoned respondents may have represented fraudulent actors, others were almost certainly legitimate participants who chose not to complete additional authentication. From the perspective of respondent experience, these outcomes represent an additional cost of authentication. This paper refers to that cumulative burden as the *Trust Tax*. Like survey length or respondent fatigue, the *Trust Tax* should become an explicit design consideration during study planning. Researchers should routinely evaluate whether the additional confidence gained

from each authentication measure justifies the additional burden imposed on legitimate participants. Authentication should therefore be optimised rather than maximised.

## *Authentication Should Be Evaluated as Part of Research Quality*

Survey methodology has traditionally evaluated research quality through measures such as sampling error, nonresponse bias, questionnaire design, measurement error and data processing quality. Authentication deserves similar consideration. The findings suggest that authentication decisions may influence who ultimately participates in research. If authentication requirements discourage particular respondent groups more than others, authentication itself may become a source of systematic selection bias. This possibility extends the evaluation of authentication beyond fraud prevention alone. Future authentication systems should therefore be assessed according to four complementary dimensions:

- Decision confidence
- Respondent experience
- Operational efficiency
- Representativeness

Fraud detection remains an important outcome, but it should no longer be considered the sole measure of authentication performance.

## *Practical Recommendations for Researchers*

The findings support several practical recommendations for organisations designing respondent authentication systems. First, accumulate evidence before requesting additional verification. Existing panel history, passive intelligence and respondent behaviour frequently provide substantial confidence without increasing respondent burden. Second, reserve active identity verification for situations in which passive evidence leaves meaningful uncertainty unresolved. Identity verification should function as an escalation mechanism rather than a universal requirement. Third, monitor the *Trust Tax*. Organisations should routinely measure abandonment, completion delays, respondent complaints and verification failures alongside traditional fraud metrics. Fourth, evaluate authentication for its impact on representativeness. Researchers should consider whether authentication requirements systematically discourage participation among specific demographic or behavioural groups and adjust their strategies accordingly. Finally, authentication systems should be reviewed periodically as technologies, fraud patterns and respondent expectations evolve. Confidence engineering is not a static process, but a continuous cycle of improvement based on accumulating evidence and operational learning.

## *Implications for Industry Standards*

The concepts presented in this paper also have implications beyond individual research projects. As respondent authentication becomes increasingly sophisticated, the market research industry would benefit from a common methodology for evaluating authentication systems. Existing professional standards provide important guidance regarding respondent privacy, informed consent, data quality and ethical research practices, but comparatively little attention has been devoted to the systematic design of authentication. The *Comprehensive Respondent Authentication Framework* (CRAF) offers one possible foundation for future standards by separating enduring methodological principles from rapidly changing technologies. Rather than prescribing particular authentication tools, the framework provides a consistent set of questions that can be applied regardless of technological implementation:

- Does the authentication process generate sufficient confidence?
- Is authentication proportional to the level of risk?
- Has passive evidence been fully utilised before requesting active verification?
- Is the *Trust Tax* justified by the additional confidence obtained?
- Does the authentication process preserve representativeness?

These questions remain relevant whether authentication relies on browser fingerprinting today or digital identity wallets in the future.

### *Conclusion*

Respondent authentication is entering a new phase of development. The challenge facing researchers is no longer identifying individual authentication technologies capable of detecting fraud. The challenge is designing authentication systems that generate sufficient confidence while preserving the respondent experience and maintaining representative samples. The CRAF provides a methodology for achieving that balance. By viewing authentication as the progressive accumulation of evidence rather than a single verification event, the framework enables organisations to integrate historical participation, passive intelligence, behavioural indicators and selective identity verification into a coherent decision-making process. The empirical findings presented in this paper demonstrate that confidence can often be achieved more efficiently through progressive, risk-based authentication than through universal verification. More importantly, they suggest that the effectiveness of authentication should be judged not only by its ability to detect fraud, but also by its impact on respondent burden, operational efficiency and research quality. As online research continues to evolve, the technologies supporting authentication will inevitably change. The principles of confidence engineering and the CRAF are intended to provide a durable methodology that remains relevant regardless of how those technologies develop.

## Implications for the Future of Online Research

This section broadens the discussion beyond your experiment and positions the framework as something that will remain relevant as technologies evolve. The technologies used to authenticate respondents will continue to evolve. Artificial intelligence, decentralised digital identities, behavioural biometrics, digital wallets and emerging identity standards will undoubtedly reshape how researchers establish respondent authenticity over the coming decade. It is unlikely that any single authentication technology described in this paper will remain dominant indefinitely. What should endure are the principles used to evaluate those technologies. The CRAF deliberately separates methodology from implementation. It provides a technology-independent approach for determining how confidence should be accumulated, when authentication should escalate and how organisations should balance research quality against respondent burden. This distinction becomes increasingly important as the pace of technological change accelerates.

### *Authentication Will Become Increasingly Invisible*

One of the most significant trends in digital authentication is the movement from explicit verification toward passive confidence assessment. Historically, respondents were frequently asked to prove who they were through passwords, challenge questions or government-issued identification. Increasingly, confidence will be established through combinations of passive evidence collected throughout the respondent journey. Device intelligence, behavioural patterns, historical participation, cryptographic identity credentials and AI-assisted risk assessment will reduce the need for respondents to repeatedly verify their identities. For legitimate participants, the ideal authentication process should become largely invisible. Identity verification will remain important, but increasingly it should serve as an exception reserved for situations in which passive evidence leaves meaningful uncertainty unresolved.

### *Confidence Engineering Beyond Identity*

Although this paper focuses on respondent authentication, the underlying methodology has broader applications throughout market research. The same principles of confidence engineering can be applied to participant qualification, sample management, respondent engagement, incentive design and longitudinal panel maintenance. In each case, organisations are making decisions under uncertainty. The objective remains the same:

- Accumulate sufficient evidence;
- Reduce unnecessary friction;
- Preserve research quality.

This suggests that confidence engineering represents not merely an authentication methodology, but a broader approach to designing respondent ecosystems.

### *Toward Industry Standards*

The market research industry has developed internationally recognised standards governing respondent privacy, data quality, transparency and ethical research practice. Authentication, however, remains comparatively fragmented. Different organisations employ different technologies, different escalation rules and different definitions of acceptable risk. The CRAF offers a common methodology that can be applied regardless of the technologies used. Rather than prescribing specific authentication tools, it encourages organizations to evaluate authentication systems according to common principles:

- Is sufficient confidence achieved?
- Is authentication proportional to risk?
- Is the *Trust Tax* minimised?
- Is representativeness preserved?

These questions provide a foundation that could support future professional guidance, standards or best-practice recommendations as authentication technologies continue to mature.

### *Future Research*

Several important questions remain unanswered. Additional studies should examine authentication-induced selection bias across different demographic groups and research contexts. Longitudinal studies could evaluate how repeated authentication influences panel retention and respondent trust over time. Future work should also investigate how AI-assisted authentication, digital identity wallets and decentralised identity systems can contribute to confidence while reducing respondent burden. Finally, comparative studies across panel providers, probability samples and customer research populations would further strengthen understanding of how authentication should be adapted to different research environments. In summary, the future of respondent authentication is unlikely to be defined by any single technology. Instead, it will be defined by organisations that successfully engineer confidence while respecting the time, privacy and goodwill of legitimate research participants. The CRAF provides one methodology for achieving that balance.

## Conclusion

Online market research depends on trust. For many years that trust was established primarily through panel management, respondent relationships and relatively simple quality controls. As fraud has become increasingly sophisticated, the industry has responded by introducing an expanding array of authentication technologies. This paper argues that the next stage in this evolution is not another technology; it is a methodology. The *Comprehensive Respondent Authentication Framework* reframes respondent authentication as the systematic process of engineering confidence rather than maximising verification. By integrating historical participation, passive intelligence, behavioural evidence and selective identity verification, the framework provides a structured approach for accumulating sufficient confidence while minimising unnecessary burden on legitimate respondents. Four concepts underpin this methodology:

1. Confidence engineering defines the objective;
2. The CRAF provides the methodology;
3. The *Trust Tax* recognises the cumulative burden imposed on legitimate respondents;
4. Authentication-induced selection bias highlights the potential consequences of authentication for sample representativeness.

Together, these concepts encourage researchers to evaluate authentication systems more comprehensively than fraud detection alone. The empirical evaluation presented in this paper demonstrates that progressive, risk-based authentication can achieve high levels of decision confidence while reducing respondent burden and operational cost relative to universal identity verification. More broadly, the findings suggest

that authentication should be evaluated according to its contribution to overall research quality rather than its ability to detect fraudulent respondents alone. The technologies supporting respondent authentication will continue to evolve. The principles presented here are intended to endure. By separating methodology from implementation, the CRAF provides a foundation that researchers, panel providers, technology vendors and professional organisations can adapt as new authentication technologies emerge.

Ultimately, the purpose of respondent authentication is not to verify as many people as possible. It is to provide sufficient confidence that research findings can be trusted. That is the central contribution of this paper.

## About the Authors

Frank Kelly, MR Practice Lead, Virtual Incentives, Boston, US.
Justin Hill, VP Fraud Prevention, Virtual Incentives, Boston, US.
Mardi Hartzog, Senior Project Manager, C-Space, Seattle, US.

## Appendix A

### *Illustrative Respondent Confidence Assessment Framework*

The Comprehensive Respondent Authentication Framework (CRAF) evaluates respondents by accumulating evidence from multiple independent sources rather than relying on a single authentication event. The following framework illustrates one possible implementation. Organisations should adapt weighting, thresholds and escalation rules according to their own research objectives, respondent populations and risk tolerance.

### *Step 1. Collect Evidence*

| Evidence category | Example indicators |
|---|---|
| **Historical evidence** | Panel tenure, participation history, redemption history, demographic consistency, previous quality scores |
| **Behavioural evidence** | Completion speed, response consistency, straight-lining, open-end quality, engagement patterns |
| **Passive intelligence** | Device fingerprint, browser characteristics, proxy detection, VPN detection, IP reputation, operating system, time zone, geolocation |
| **Identity evidence** | Email verification, mobile verification, government-issued ID, biometric verification, digital identity credentials |
| **Operational evidence** | Previous fraud investigations, manual review history, prior authentication outcomes |

*Table 6*

### *Step 2. Assign Confidence Values*

**Positive Evidence**

| Indicator | Illustrative confidence value |
|---|---|
| **Panel tenure greater than 3 years** | +20 |
| **Consistent participation history** | +15 |
| **Successful redemption history** | +15 |
| **Stable demographic profile** | +10 |
| **Residential IP** | +5 |
| **Positive email presence** | +10 |
| **Government ID verified** | +25 |

Copyright © **Esomar** 2026



*Table 7*

**Negative Evidence**

| Indicator | Illustrative adjustment |
|---|---|
| **Data center proxy** | –20 |
| **Browser spoofing** | –15 |
| **Anti-detect browser** | –25 |
| **Multiple-account indicator** | –30 |
| **Time zone mismatch** | –10 |
| **Suspicious device characteristics** | –15 |
| **No digital presence** | –5 |

*Table 8*

### *Step 3. Confidence Assessment*

| Confidence score | Interpretation | Recommended action |
|---|---|---|
| **80–100** | High confidence | Approve redemption |
| **60–79** | Moderate confidence | Continue passive monitoring |
| **40–59** | Uncertain | Collect additional passive evidence or perform manual review |
| **Below 40** | Insufficient confidence | Escalate authentication |

*Table 9*

The confidence values presented are illustrative and intended to demonstrate the methodology, rather than prescribe a universal scoring system.

## Appendix B

### *Illustrative Authentication Escalation Matrix*

Authentication should be proportional to both respondent confidence and the potential consequences of fraud.

| Confidence level | Low incentive | Moderate incentive | High incentive |
|---|---|---|---|
| **High** | Approve | Approve | Passive review |
| **Moderate** | Passive review | Passive review | Identity verification |
| **Uncertain** | Additional passive evidence | Identity verification | Identity verification |
| **Low** | Manual review | Identity verification | Reject or enhanced review |

*Table 10*

Table 10 illustrates how authentication intensity can increase according to both uncertainty and financial or research risk.

## Appendix C

### *Trust Tax Assessment Framework*

The *Trust Tax* represents the cumulative burden imposed upon legitimate respondents during authentication. Each authentication activity can be evaluated across five dimensions (see Table 11).

Copyright © **Esomar** 2026



| Dimension | Description | Rating (1–5) |
|---|---|---|
| **Time** | Additional time required to complete authentication. | |
| **Cognitive effort** | Mental effort required to understand and complete the process. | |
| **Privacy exposure** | Amount of personal information disclosed. | |
| **Technology requirements** | Additional hardware or software required. | |
| **Emotional burden** | Frustration, concern or anxiety associated with the process. | |

*Table 11. Trust tax dimensions.*

Total *Trust Tax* = Sum of the five dimensions.

| Authentication method | Illustrative *Trust Tax* |
|---|---|
| **Passive device intelligence** | 1 |
| **Email verification** | 3 |
| **SMS verification** | 5 |
| **Knowledge-based questions** | 8 |
| **Government-issued ID verification** | 18 |
| **Selfie with liveness detection** | 22 |

*Table 12. Illustrative Trust Tax scores.*

Organisations should validate *Trust Tax* scores through respondent research and adapt values according to their own populations.

## Appendix D

### *Authentication Design Checklist*

Before implementing or modifying an authentication process, researchers should consider the following questions.

**Confidence**

- What level of confidence is required for this study?
- Does the proposed authentication meaningfully increase confidence?

**Passive evidence**

- Have all available passive authentication methods been utilised?
- Can sufficient confidence be achieved without active verification?

**Escalation**

- Is authentication proportional to the level of observed risk?
- Are additional verification steps reserved for respondents with meaningful residual uncertainty?

**Respondent experience**

- What *Trust Tax* is imposed upon legitimate respondents?
- Can the same confidence be achieved with less respondent burden?

**Representativeness**

- Could authentication discourage participation among particular respondent groups?
- Is authentication-induced selection bias being monitored?

Copyright © **Esomar** 2026



**Operational performance**

- Are abandonment rates being measured?
- Are authentication costs being tracked?
- Is the authentication strategy reviewed regularly as fraud patterns evolve?

## Appendix E

*Illustrative Respondent Confidence Profiles*

| Evidence | Observation |
|---|---|
| **Panel tenure** | 6 years |
| **Completed studies** | 184 |
| **Redemption history** | Consistently successful |
| **Device** | Stable |
| **Network** | Residential IP |
| **Email presence** | Strong |
| **Fraud indicators** | None observed |

*Table 13. Respondent A: High confidence.*

- Confidence Score: 92
- Recommended action: Approve redemption without additional authentication.

| Evidence | Observation |
|---|---|
| **Panel tenure** | New registration |
| **Device** | Browser spoofing detected |
| **Network** | Datacenter proxy |
| **Email presence** | None identified |
| **Time zone** | Inconsistent |
| **Multiple accounts** | Indicator present |

*Table 14. Respondent B: Low confidence.*

- Confidence Score: 28
- Recommended action: Escalate authentication before redemption.

## Appendix F

*Mapping Authentication Technologies to the Framework*

The *Comprehensive Respondent Authentication Framework* is technology independent. Individual authentication technologies can be mapped to one or more evidence categories.

| Technology or signal | Framework category |
|---|---|
| **Panel tenure** | Historical evidence |
| **Participation history** | Historical evidence |
| **Browser fingerprint** | Passive intelligence |

Copyright © **Esomar** 2026


| | |
|---|---|
| **Device intelligence** | Passive intelligence |
| **Proxy detection** | Passive intelligence |
| **VPN detection** | Passive intelligence |
| **Email reputation** | Passive intelligence |
| **Response consistency** | Behavioural evidence |
| **Completion time** | Behavioural evidence |
| **Open-end analysis** | Behavioural evidence |
| **Government-issued ID** | Identity evidence |
| **Biometric verification** | Identity evidence |
| **SMS verification** | Identity evidence |
| **Email verification** | Identity evidence |
| **Manual analyst review** | Operational evidence |
| **Previous fraud investigations** | Operational evidence |

*Table 15*

The framework is intended to accommodate both current and future authentication technologies by classifying them according to the evidence they contribute rather than the specific implementation employed.

Copyright © **Esomar** 2026

Congress 2026

# We Do Not Understand People

## The Empathy Illusion in Financial Anxiety Research

*Jaime Solórzano Villela*

### Introduction: The Empathy Illusion

"We misunderstand them because we over-edit what they reveal".

The language of empathy is now deeply embedded in contemporary research and brand practice. Financial services teams talk about reassurance, accessibility, transparency, confidence and emotional support. Qualitative research frequently confirms these ideas because they are present in people's own language. Participants do describe stress, uncertainty, pressure and a desire to feel more in control. Yet the apparent clarity of these narratives can be deceptive. Once we move from fieldwork to synthesis, we often translate a lived situation into a more manageable one: less contradictory, less morally charged, less relationally fraught and, therefore, more comfortable to act upon. Financial anxiety is especially vulnerable to this reduction because money is never only about money. It is about competence, adulthood, care, legitimacy, gender, dependence, secrecy, aspiration and memory. It shows up in routines, in micro-rituals, in bodily responses, in the management of appearances and in the negotiations that happen inside homes long before any brand enters the scene. When these layers are flattened into statements such as "people seek transparency" or "simplicity reduces stress", the resulting insight may still sound intelligent and humane, but it has already made the person smaller than the life being described.

This paper proposes that one of the biggest risks in qualitative work today is not lack of empathy, but the performance of empathy. By this we mean a form of professional translation that produces emotionally resonant language, while quietly removing the very contradictions that would make the insight structurally true. In financial anxiety research, this has serious consequences. It can lead brands to design only for declared needs, policy teams to overlook hidden power dynamics and researchers to confuse interpretive neatness with depth. Our contribution is therefore twofold. First, we show how the same ethnographic material can produce two very different kinds of output: one conventionally usable and one more faithful to lived complexity. Second, we offer a practical way of auditing synthesis before the final deck is closed. The goal is not to reject clarity, but rather to prevent clarity from becoming a form of erasure.

### Study Design and Methodological Stance

The study at the centre of this paper is intentionally small, focused and craft-led. It consists of eight in-home ethnographic interviews in Spain with participants aged 28 to 55 across mixed income levels and both urban and peri-urban households. Each session lasts approximately two-and-a-half hours and combine interview with observation, artefact elicitation and reflexive interpretation. This design is chosen not to produce statistical coverage, but to stay close to the infrastructures, routines and emotional textures through which financial anxiety is actually lived.

#### *Core Methods*

- 2.5-hour in-home immersion in everyday domestic context;
- Financial artefact elicitation using apps, bills, notes, spreadsheets and household tracking systems;
- Contextual observation of routines, interruptions, silences and household dynamics;
- Emotional timeline mapping to connect present anxieties to remembered events and periods of instability;
- Reflexive researcher journalling to record non-verbal cues, atmospheres, contradictions and interpretive tension.

#### *Analytical Focus*

- Lived anxiety rather than abstract opinion.
- Contradictions rather than consistency.
- Moral narratives rather than purely functional explanations.
- Behaviour and ritual alongside discourse.
- The gap between what feels empathetic in a deck and what remains true to the field.

A crucial feature of the study is that it generates two outputs from the same body of material. This is not a comparison between two samples or two methods. It is a comparison between two modes of interpretation. Output A follows the logic of a conventional insight deck: it organises material into personas, tensions and opportunities. Output B follows a different logic: it preserves patterned contradiction through a complexity map that shows how financial anxiety is organised across bodily response, social performance, household power and inherited memory. The contrast reveals not what participants change, but what the research process removes.

| Stage | What exists in the material | What is often lost in "translation" |
|---|---|---|
| **Fieldwork** | Rituals, silences, contradictions, atmospheres, bodily reactions | Nothing yet; complexity is still visible |
| **Transcript** | Language, hesitations, narrative sequences, emotional qualifiers | Gesture, timing and household context begin to thin out |
| **Synthesis** | Themes and interpretations emerge | Contradiction is often resolved into cleaner tension statements |
| **Deck** | Personas, opportunity spaces, strategic framing | Moral discomfort and power asymmetry are softened for usability |
| **Strategy** | Actions and interventions | Teams act on reassurance narratives while structural realities remain under-seen |

*Table 1. Where the empathy illusion enters the process.*

## The dominant industry story

Across financial services research, a familiar narrative repeatedly appears. Consumers are described as overwhelmed and in search of confidence; they want help making sense of complexity; they want security, predictability and transparent guidance. Trust is positioned as the key emotional driver, while simplification is framed as the obvious response. None of these statements are entirely wrong. The problem is that when treated as the whole truth, they conceal the more uncomfortable architecture of financial anxiety. In our material, anxiety is not only about information overload or lack of clarity. It is bound up with shame, with domestic politics, with symbolic adulthood, with self-surveillance, with the moral judgement participants direct at themselves and others and with memories that predate current financial circumstances. This means that a conventional deck could still sound valid while describing only the most socially presentable layer of the phenomenon.

"Clean insight is not always false. But it can be incomplete in exactly the places that matter most".

## Output A: The conventional insight deck

If we translate the study through a standard commercial synthesis process, the output could be strong by ordinary professional standards. It could be clear, memorable and easy to socialise. It could give stakeholders recognisable consumer types, usable tensions and a compelling opportunity statement. It could also reduce the material to a more comfortable shape.

### *The Three Conventional Personas*

**The overwhelmed planner**

- Surface narrative: “I need to be organised. Planning reduces my anxiety”;
- Uses budgeting tools, notes or spreadsheets, but not always consistently;
- Creates detailed plans after moments of alarm such as an unexpected expense or crisis-related news;
- Experiences relief from the act of planning, even when the plan is not followed later;
- Feels shame when discipline breaks, which often triggers a new and stricter cycle of control;
- Core insight: planning often functions as emotional self-soothing before it functions as financial optimisation.

**The silent avoider**

- Surface narrative: “I try not to think about money too much. It stresses me”;
- Delays opening bank apps, bills or messages that might confirm a feared reality;
- Uses humour, understatement or topic-shifting when finances become emotionally-loaded;
- Avoids deeper financial conversations inside couples or families;
- Experiences anxiety physically, but names it in softer language such as tiredness, tension or stress;
- Core insight: avoidance is less a sign of irresponsibility than a strategy of psychic protection against shame and identity threat.

**The hyper-controller**

- Surface narrative: “If I stay in control, nothing bad will happen”;
- Checks balances or account movements repeatedly through the day;
- Tracks household spending closely and may monitor a partner without explicit discussion;
- Uses moral language about discipline, carelessness or financial maturity;
- Fears losing command not only of money, but of self-image and legitimacy;
- Core insight: control is often tied to identity, worth and role performance, not simply to rational risk management.

## Five Tidy Tensions a Deck Would Likely Emphasise

- Control versus enjoyment;
- Security versus freedom;
- Planning versus spontaneity;
- Transparency versus complexity;
- Independence versus dependence.

These tensions are not meaningless; in fact, they capture real pressure-points. The issue is that they make contradiction feel symmetrical and manageable. They suggest balanced trade-offs where the field often showed more unsettling asymmetries: secrecy inside intimacy, control used as identity armour and pleasure followed by moral punishment. In other words, the deck does not lie—it is domesticated.

### *Why Output A Feels Persuasive*

- It sounds empathetic because it uses emotionally intelligent language;
- It makes diverse households legible through a small set of memorable categories;
- It converts messy material into decision-ready tensions and opportunities;
- It protects stakeholder confidence by keeping discomfort within manageable bounds.

A likely opportunity statement from Output A would be entirely plausible: financial brands should provide reassurance through simpler ecosystems, clearer guidance and confidence-building design. That statement is useful, but it remains partial because it treats anxiety primarily as a problem of support, not also as a problem of concealment, morality, asymmetry and embodied threat.

Copyright © **Esomar** 2026


## What Conventional Synthesis Removes

When we audit the distance between the field and the conventional deck, three categories of loss become visible. First, contradictions are flattened. Participants who plan obsessively also spend impulsively; those who mock anxiety in others also reveal private panic rituals of their own. Second, moral discomfort is softened. Shame becomes stress, secrecy becomes privacy and interpersonal inequality becomes a neutral division of financial roles. Third, power dynamics are erased. Dependence within couples, unequal access to knowledge and class-coded definitions of responsibility rarely survive into the clean narrative. This is the crucial move behind the empathy illusion. By removing the most difficult elements, research can become more emotionally resonant while becoming less exact. The audience feels it understands people better precisely because those people have been translated into a more reassuring form.

| Conventional language | What the ethnography reveals | Why the difference matters |
|---|---|---|
| **Need for reassurance** | Relief-seeking mixed with secrecy, surveillance and fear of exposure | Support alone will miss protective concealment behaviours |
| **Desire for transparency** | Participants often maintain complex and partially hidden systems of their own | Complexity can serve control, privacy and symbolic competence |
| **Planning reduces stress** | Planning often soothes emotionally even when behaviour does not become more stable | The ritual may matter more than the plan |
| **Trust drives choice** | Trust is filtered through household roles, shame and perceived judgement | Trust is relational and moral, not only institutional |
| **Simplicity reduces anxiety** | Some complexity is actively maintained because it gives a feeling of command | Over-simplification can remove a user's sense of agency |
| **Independent consumer** | Dependency and asymmetry remain active inside couples and families | Solutions designed for an abstract individual can miss household reality |

*Table 2. From conventional synthesis to lived complexity.*

## Output B: The Complexity Map

Output B keeps the material in a more demanding form. Instead of resolving contradiction into neat personas and tensions, it maps patterned realities that coexist across households. The purpose is not to produce chaos or romanticise messiness. It is to preserve enough of the original structure for financial anxiety to remain recognisably human. In this version, the task of insight is not to remove tension but to organise it. The complexity map reveals at least seven domains that conventional synthesis tends to compress. Each domain is observable in behaviour, artefacts, speech or relational dynamics. Together they make a much stronger argument: financial anxiety is not merely a cognitive need for confidence; it is a lived ecology of concealment, bodily anticipation, moral accounting and identity protection.

### *Mapped Domain 1: Rituals Of Concealment*

Concealment appears not as occasional dishonesty, but as patterned self-protection. Participants describe or display secondary accounts, cash-only purchases, selective disclosure and small reserves kept outside shared visibility. Notifications are managed; amounts are rounded; certain expenditures are moved away from digital traceability. These practices are not always signs of deviance. Often, they are methods of preserving dignity, autonomy or emotional safety in environments where full transparency feels risky.

### *Mapped Domain 2: Performance Of Stability*

Financial normality is often staged. Participants maintain subscription habits, social spending or symbols of continuity even during periods of pressure. Social presentation matters; group conversations are managed to avoid signalling decline. Everyday stability, in this sense, functions as reputational labour. The performance of being "fine" is not peripheral to anxiety; it is one of the ways anxiety is managed.

### *Mapped Domain 3: Embodied Anxiety*

The field material shows that financial anxiety is not simply verbal or attitudinal. It is somatic. Participants hold their breath before checking balances, delay account reviews until they feel prepared, revisit apps late at night after arguments or negative news and display visible physical tension while discussing bills or obligations. This matters methodologically, because a research summary built only from articulation will understate the physiological force of the experience.

### *Mapped Domain 4: Gendered Identity Performance*

Control, competence and financial literacy are not gender-neutral in the material. Among some male participants, financial command functions as identity armour. Uncertainty could threaten not only decision quality but legitimacy. Advice-seeking is sometimes minimised, uncertainty is rarely volunteered and monitoring behaviours becomes a way to sustain authority. This is not reducible to men alone, nor does it imply uniformity across households. The point is that financial anxiety can be entangled with gendered expectations of adulthood, provision and competence.

### *Mapped Domain 5: Inherited Fear Scripts*

Participants frequently relate present-day behaviours to earlier crises, parental warnings, remembered scarcity or the emotional residue of the 2008 period. Even when current circumstances are relatively stable, fear could remain active as an inherited script. Some participants over-save because instability had once been formative; others swing toward reactive spending because deprivation had made present enjoyment morally urgent. This places financial anxiety within a longer cultural and biographical timeline, rather than treating it as a purely present-state response.

### *Mapped Domain 6: Moral Accounting*

Money is repeatedly organised through moral language. Some spending is described as deserved, other spending as irresponsible, childish, selfish or shameful. Earning through effort is valorised differently from "easy money". Personal indulgence requires justification. Errors are often narrated as failures of character rather than ordinary misjudgements. This moralisation helps explain why financial anxiety is rarely solved by information alone: the pain is not only practical, but ethical and identity-laden.

### *Mapped Domain 7: Relational Asymmetry*

Financial life in households is not organised by equal knowledge or equal power. One person might manage logistics while another retains ultimate decision authority. Dependency could be disguised as trust; unequal literacy could be normalised as role distribution; resentment exists even where explicit conflict does not. This again complicates the standard notion of an isolated consumer seeking clarity. Many financial anxieties are relationally distributed and strategically hidden within the household itself.

| Mapped reality | Observable manifestations | What it reveals | Implication for strategy |
|---|---|---|---|
| **Concealment** | Hidden accounts, selective disclosure, cash workarounds | Transparency can feel dangerous, not liberating | Design for privacy and dignity, not only disclosure |

Copyright © **Esomar** 2026


| | | | |
|---|---|---|---|
| **Stability performance** | Lifestyle continuity despite pressure | Social legitimacy is being defended | Messaging must recognise appearance-management pressures |
| **Embodied anxiety** | Breath-holding, late-night checking, delay rituals | Financial systems can trigger threat responses | Reduce anticipatory stress in journeys and interfaces |
| **Gendered control** | Monitoring, reluctance to admit uncertainty | Money can protect identity status | Avoid communication that intensifies judgement or inadequacy |
| **Inherited scripts** | Crisis references, parental warnings, over-saving | History remains active in present decisions | Treat behaviour as biographical, not merely transactional |
| **Moral accounting** | Justification narratives, guilt, good versus bad money | Financial action is ethically coded | Support decision-making without moralising users |
| **Relational asymmetry** | Uneven literacy, hidden reliance, household tension | The user is often a household, not an individual | Design for shared and unequal realities, not idealised autonomy |

*Table 3. The complexity map at a glance.*

## The Three Systemic Distortions

The comparison between Output A and Output B makes three recurrent distortions visible. These are not presented as mistakes made by careless researchers. Rather, they are structural tendencies produced by the way commercial insight needs to travel: from field to transcript, from transcript to synthesis, from synthesis to deck and from deck to strategy. At each stage, clarity is rewarded and discomfort becomes expensive. The distortions are therefore understandable. They are also consequential.

### *Coherence Bias*

Coherence bias occurs when contradictory behaviours are translated into cleaner narratives of balance or trade-off. Instead of preserving the fact that a participant both monitors obsessively and spends impulsively, synthesis produces a stable tension such as control versus enjoyment. This makes the material more elegant, but it reduces its explanatory force. Contradiction is not a flaw in the data; it often *is* the data.

### *Sanitisation Drift*

Sanitisation drift describes the gradual softening of emotionally and culturally difficult material. Shame becomes stress; secrecy becomes privacy; surveillance becomes engagement; gendered legitimacy becomes a generic confidence issue. The language remains empathetic, but the sharper edges of lived experience are rounded down. This protects organisational comfort while weakening interpretive accuracy.

### *Empathy Performance*

Empathy performance is the endpoint at which stakeholders feel emotionally aligned with consumers, because the insight deck sounds humane, nuanced and well-intentioned. Yet, what they are aligned with is often a professionally edited version of humanity. The deck performs care while shielding the audience from structural discomfort. In this sense, empathy becomes a style of interpretation rather than a discipline of fidelity.

| Distortion | How it appears in decks | What it hides | Risk if left unchecked |
|---|---|---|---|
| **Coherence bias** | Clean tensions and stable personas | The behavioural force of contradiction | Weak predictive power and over-neat strategy |
| **Sanitisation drift** | Softened emotional vocabulary | Shame, secrecy, dependency, status threat | Interventions that feel safe but miss the real pain |
| **Empathy performance** | Humane tone and emotionally resonant framing | The structural discomfort necessary for depth | False confidence that the audience already understands people |

*Table 4. Distortion audit.*

## *The Insight Compression Test*

If the problem is structural, the response must be practical. The insight compression test is proposed here not as a theoretical flourish, but as a professional discipline that can be used before an insight deck is finalised. Its purpose is to help teams identify where human material has become over-translated. The test does not require abandoning clarity; it requires placing one final layer of critical pressure on clarity before it hardens into strategy.

**Five diagnostic questions**

1. What contradiction did we resolve too quickly?
2. What discomfort did we soften in order to make the story more acceptable?
3. What behaviour in the field complicates our neatest conclusion?
4. Which power dynamic, asymmetry or hidden dependency has disappeared from the final narrative?
5. What ambiguity might be strategically valuable precisely because it remains unresolved?

**How to use it in practice**

- Run the test at the point where synthesis is becoming presentation, not after strategy is locked;
- Force every major insight statement to be checked against one contradictory observation from the field;
- Maintain a visible column for what the deck is choosing not to simplify;
- Treat emotionally uncomfortable material as a design signal, not automatically as a communication risk;
- Use the test collaboratively so that researchers, strategists and stakeholders share responsibility for what gets removed

| Audit question | Evidence to review | What a stronger deck would preserve |
|---|---|---|
| **Resolved contradiction** | Conflicting behaviours, reversals, mixed routines | The contradiction itself, not only the cleaned tension |
| **Softened discomfort** | Shame language, secrecy, status defence, moral judgement | The social or emotional edge that gives behaviour meaning |
| **Behavioural challenge** | Observed rituals, artefacts, household workarounds | At least one lived behaviour that tests the main claim |
| **Hidden power dynamic** | Couple dynamics, unequal literacy, dependency, control | A relational reading rather than a fully individualised one |
| **Strategic ambiguity** | Patterns that remain unresolved but recurrent | A design opportunity built around tension rather than false closure |

*Table 5. Insight compression test worksheet.*

Copyright © **Esomar** 2026


### *What This Means for Marketers, Financial Brands and Researchers*

The implications of this work are broader than one category and more specific than a general plea for nuance. For marketers, financial brands and adjacent businesses, the key lesson is that reassurance alone is not enough. Many people do not simply need clearer information; they need systems, journeys and communications that recognise concealment, moral strain, relational inequality and embodied anticipation. A brand that only speaks the language of transparency may overlook the fact that for some users full visibility itself can feel exposing. For strategists, the implication is that the unit of analysis may need to shift. Financial choice is often treated as an individual cognitive act. Our material suggests it is often householded, moralised and socially staged. Products, propositions and communication strategies therefore need to be tested not only against stated preference, but against domestic reality: who knows what, who controls what, who hides what and who carries the emotional burden of appearing stable.

For researchers, this paper is an argument for methodological courage. The craft of qualitative work does not end with listening well in the field. It extends into how faithfully complexity is curated afterwards. Good synthesis is not the elimination of ambiguity at all costs. It is the disciplined decision about which ambiguities must remain, because removing them would make the insight less true than the life it claims to represent.

## Conclusion

Financial anxiety is frequently interpreted through a humane but narrowed lens. We hear the need for confidence, clarity and reassurance, and then we build strategy around those signals. This paper suggests that the more serious challenge lies elsewhere. Between field and deck, we routinely compress people until they become easier to empathise with and easier to act upon. That compression produces the empathy illusion: the feeling of deep understanding without the discipline of preserving contradiction. The contribution of the complexity map is not that it is messier than a persona deck. It is that it is more faithful to the actual organisation of lived financial anxiety: embodied, moralised, relational, gendered and historically carried. The contribution of the insight compression test is that it gives research teams a practical way to protect this complexity from disappearing in the final stages of synthesis. If the evolution of human understanding is the theme, then the challenge is not merely to understand people faster, or to describe them more elegantly. It is to interpret them more bravely. We do not honour people by making them smaller so that strategy can feel bigger. We honour them by preserving enough of their contradiction for strategy to become more intelligent, more responsible and more real and truer.
“Metamorphosis in insight does not begin when we simplify people more efficiently. It begins when we become more accountable for what we simplify away".

### *Key Takeaways for the Esomar Audience*

- Financial anxiety cannot be understood only through declared needs such as control, transparency and reassurance;
- Conventional insight synthesis often creates an empathy illusion by flattening contradiction and softening discomfort;
- A complexity map can preserve strategically useful realities such as concealment, embodied threat, moral accounting and relational asymmetry;
- The three distortions to watch are coherence bias, sanitisation drift and empathy performance;
- The insight compression test offers a practical, replicable way to audit what has been over-edited before insight becomes strategy.

## About the Author

Jaime Solórzano Villela, LATAM Regional Manager & Global Head of Insights, Q2Q Global, LATAM Office, Mexico City.

# From Raw Text to Executive Insight

## Evaluating LLM Summarisation Models for Customer Complaint Analytics

*Joanna Quan*

*Wenqi Shen*

*Chitvan Trivedi*

## Esomar

Office address:

Burgemeester Stramanweg 105

1101 AA, Amsterdam

Phone: +31 20 664 21 41

Email: info@Esomar.org

Website: www.Esomar.org

Publication Date: September 2026

ESOMAR Publication Series Volume S424 Congress 2026

ISBN 92-831-0337-8

Congress 2026
Copyright © **ESOMAR** 2026

Copyright © **ESOMAR** 2026
Congress 2026

# From Raw Text to Executive Insight

## Evaluating LLM Summarisation Models for Customer Complaint Analytics

*Joanna Quan*
*Wenqi Shen*
*Chitvan Trivedi*

### Introduction

Organisations across industries face a growing challenge at the intersection of data volume and actionable intelligence. Customer complaints – submitted through call centres, web forms, mobile applications, warranty databases, and regulatory channels – represent one of the richest sources of product quality and safety intelligence available to businesses. A single large consumer products manufacturer may receive hundreds of thousands of complaints annually. Yet the sheer volume of these submissions makes systematic processing increasingly impractical through manual means alone. Trained analysts who read each complaint narrative and extract structured intelligence are a constrained resource. The cost of keeping up with complaint volume through manual review grows faster than most organisations can sustain, while the business consequences of delayed insight – missed safety signals, slow quality responses – compound over time.

Large language models (LLMs) have emerged as a technically compelling solution to this challenge. Their ability to process natural language at scale, follow structured instructions, and synthesise information across large collections of text positions them as natural candidates for automating the transformation of raw complaint text into executive-ready summaries. Customer feedback summarisation has emerged as one of the most immediately applicable LLM use cases in organisational analytics. Yet a critical practical question remains largely unanswered: when an LLM is asked to summarise customer complaints for executive consumption, how faithfully does the resulting summary represent the underlying data – and does the workflow through which complaints are processed before summarisation affect the quality of the output?

This study addresses that question through a controlled empirical evaluation using vehicle safety complaints from the National Highway Traffic Safety Administration (NHTSA). We evaluate two summarisation workflows. In Workflow 1 – Direct Summarisation – raw complaint narratives are sent directly to an LLM, which is asked to generate an executive summary from the text alone. In Workflow 2 – Hybrid Summarisation – the LLM receives the same raw narratives plus a pre-computed structured table showing how frequently each issue category appears and how severe those complaints are (measured by the rate of crashes, injuries, fires or deaths). This hybrid approach tests whether providing explicit quantitative context alongside narrative text improves summary faithfulness.

This study builds on findings from Quan et al. (2026), who evaluated LLM reliability for structured issue coding – the task of classifying individual complaint narratives into predefined component categories. Their study found that

conservative, evidence-constrained prompting (assigning a component category only when the complaint text clearly supports it) produced the most reliable results, and that LLMs identify the correct broad issue category far more accurately than the precise subcategory. Taking those classification outputs as its starting point, the present study asks whether aggregating them into structured frequency and severity profiles – and providing them alongside the raw complaint text – improves the faithfulness of downstream executive summaries.

We evaluate summary faithfulness along two dimensions that directly reflect organisational information needs: frequency coverage – does the summary correctly represent which issues appear most often – and severity coverage – does the summary appropriately surface issues associated with the highest rates of crashes, injuries, fires, or deaths. We additionally evaluate executive accessibility through validated readability metrics, recognising that a technically accurate summary that is difficult to read provides limited value in an executive decision-making context.

Our central findings are threefold. First, Workflow 1 and Workflow 2 show no evidence of a meaningful difference in frequency or severity coverage – LLMs appear to infer quantitative patterns from narrative text with accuracy comparable to when those patterns are provided explicitly. Second, Workflow 2 produces significantly more readable summaries ($p = .043$, $d = 0.41$), improving executive accessibility by approximately two reading grade levels. Third, comparison against a ground truth benchmark reveals that the primary constraint on summary faithfulness is not the summarisation workflow but the quality of upstream issue classification.

## Background and Related Work

### *Text Analytics for Customer Feedback*

A substantial body of research demonstrates that unstructured customer feedback contains rich, actionable intelligence about product quality, service failures and emerging safety risks. Netzer et al. (2012) showed that text mining of consumer-generated content can yield brand perception maps that replicate survey-based measures at a fraction of the cost. Tirunillai and Tellis (2014) demonstrated that online reviews contain leading indicators of firm performance. More recently, Zhang et al. (2022) proposed probabilistic defect models to identify and summarise product defects from online user-generated content. Wang et al. (2022) developed attribute embedding frameworks to extract product attributes from consumer reviews, enabling structured insight generation from unstructured text. These studies establish that automated text analytics can produce managerially relevant structured information from customer feedback – the core task we evaluate. However, prior research primarily develops and evaluates purpose-built machine learning models. The question of whether general-purpose LLMs can synthesise these patterns into faithful executive summaries without task-specific training remains largely unanswered.

### *LLM Summarisation and Faithfulness*

Large language models demonstrate strong performance across a range of natural language generation tasks, including document summarisation. Prompt design has emerged as a critical determinant of output quality (Wei et al., 2022; Kojima et al., 2022). A directly related concern is LLM summarisation faithfulness – the degree to which model-generated summaries accurately represent their source material. Research has documented systematic tendencies toward generating outputs that sound plausible but may not faithfully represent source content (Maynez et al., 2020). In organisational contexts, related work has shown that AI-generated summaries of consumer reviews introduce systematic positivity biases – producing more favourable sentiment than the underlying reviews they summarise (Quan

& Shen, 2026). For customer complaint summarisation, where summary faithfulness has direct operational consequences for safety investigations and quality management, these concerns motivate empirical evaluation rather than assumed reliability.

## Research Design and Methodology

### *Data Source and Sample*

The primary data source is the NHTSA Office of Defects Investigation (ODI) consumer complaints database, which contains 1,551,062 unique vehicle complaint records spanning model years 1949 to 2026. Each complaint includes a free-text narrative field in which the consumer describes their problem in their own words, selects the components they believe were involved from a predefined list, and indicates whether the complaint involved a crash, fire, injury or death. This information is consumer-reported at time of submission and subsequently reviewed by NHTSA staff. We focus deliberately on regulatory safety-complaint data, a high-stakes, domain-specific setting chosen because NHTSA's records provide human-labelled component and severity ground truth – a property rare among complaint corpora and one that makes faithfulness evaluation possible in the first place. This focus is a scope condition rather than a limitation: generalisation to other organisational complaint settings, such as product reviews or social-media comments, requires replication and is an explicit direction for future work.

Three vehicle datasets were constructed to enable replication across qualitatively different complaint profiles. For each vehicle, 1,000 complaints were randomly sampled from the full NHTSA database (995 for the Tesla Model Y, where fewer complaints were available after removing five erroneously filed records). These 1,000-complaint pools form the basis for the severity rate and component universe calculations reported in Table 1 and serve as the sampling frame for this study. For the summarisation experiment specifically, a further random sample of 300 complaints was drawn from each vehicle's 1,000-complaint pool, using a fixed random seed to ensure all three models received identical input within each dataset.

Table 1 summarises dataset characteristics. Several columns warrant explanation. The *component universe* is the total number of distinct NHTSA component labels present in each dataset – it reflects how taxonomically complex the complaint landscape is for that vehicle. The *overall severity rate* is the proportion of complaints in the 1,000-complaint pool that involved at least one crash, fire, injury or death as recorded in the severity flags. The *80% coverage threshold* indicates how many of the most frequent components are needed to account for 80% of all complaints – for the Fusion, the top 8 components account for 80% of complaints; for the Camry, it takes 14 components because the complaint profile is more dispersed. This threshold is used as the denominator for our frequency coverage evaluation metric, as described in the Evaluation Framework section. The *% unknown* column reports the proportion of complaints where at least one of the NHTSA-assigned component codes was "UNKNOWN OR OTHER" – meaning that for part or all of the complaint, even trained analysts could not confidently assign a specific component category. These complaints are excluded from severity evaluation metrics because they represent genuinely ambiguous cases where ground truth classification is uncertain, making evaluation against them unreliable.

| Vehicle Dataset | Role | N | Component Universe | 80% Threshold | Severity Rate | % Unknown |
|---|---|---|---|---|---|---|
| **2010 Ford Fusion** | Primary | 1,000 | 44 components | Top 8 | 2.4% | 2.6% |
| **2007 Toyota Camry** | Replication 1 | 1,000 | 73 components | Top 14 | 12.1% | 7.7% |
| **2023 Tesla Model Y** | Replication 2 (EV/ADAS) | 995 | 54 components | Top 10 | 13.8% | 12.6% |

*Table 1. Vehicle dataset characteristics.*

## *Summarisation Workflows*

Three summarisation conditions were evaluated. Table 2 summarises the design.

| | Workflow 1 (Direct) | Workflow 2 (Hybrid) | GT Benchmark |
|---|---|---|---|
| **What LLM receives** | 300 raw complaint narratives | 300 raw complaint narratives + structured frequency & severity table | Structured frequency & severity table only |
| **Pre-processing required** | None | Compute component frequencies and severity rates from ground truth | Compute component frequencies and severity rates from ground truth |
| **Severity information source** | Inferred by LLM from narrative text | Ground truth | Ground truth |
| **Role** | Baseline – current practice | Proposed enriched workflow | Control |

*Table 2. Summarisation workflow design.*

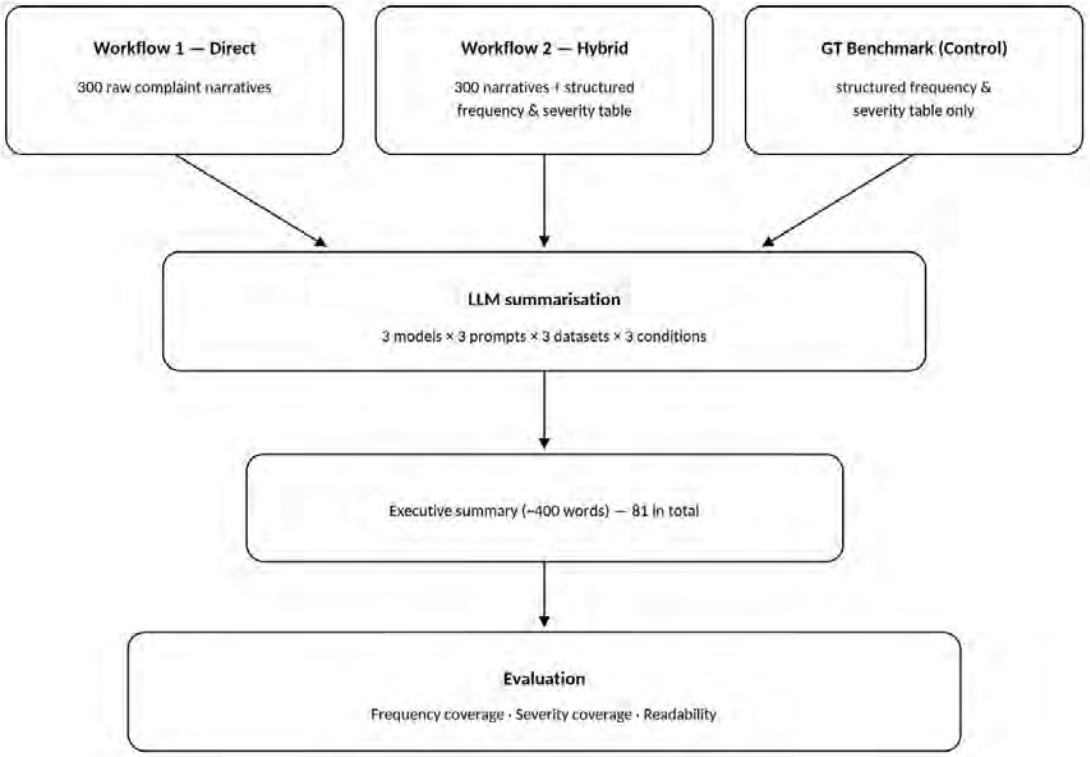



*Figure 1. Experimental design. Two deployable summarisation workflows and a ground-truth control condition are compared across three models, three prompts, and three datasets, then evaluated on frequency coverage, severity coverage, and readability.*

*Workflow 1 (direct summarisation)* sends 300 raw complaint narratives directly to the LLM with a prompt requesting an executive summary. This represents the most common current organisational deployment – providing LLMs with unstructured text and relying on their natural language processing capability to extract patterns without any pre-processing step.

*Workflow 2 (hybrid summarisation)* provides the LLM with both the 300 raw complaint narratives and a pre-computed structured frequency and severity table. The table lists issue components up to the 80% cumulative frequency threshold for each dataset (8 components for Fusion, 14 for Camry, 10 for Tesla – see Table 1), along with ground truth severity rates for each component. Severity rates and the criteria used to identify high-severity components are defined in the Evaluation Framework section. The LLM is explicitly instructed to use the structured data to identify the most important issues by frequency and severity, and to use the narratives to provide context and qualitative evidence.

*Ground truth benchmark* uses only the structured frequency and severity table – without raw narratives. It is not a deployable workflow because in real organisational settings, ground truth component labels are not available at the time of summarisation; they require human expert validation that the summarisation task is meant to replace. It provides a reference point for interpreting the gap between Workflow 1 and 2. Given that the benchmark's inputs directly contain the component labels our evaluation metrics check for, its frequency coverage advantage over Workflow 1 is partially expected by design; its severity coverage advantage is more interpretively meaningful and is discussed in the Results section. Whether managers can construct such a table without manual coding is itself an open question. Quan et al. (2026) found that conservative, evidence-constrained prompting produced the most reliable LLM complaint

classification, and that models identify broad issue categories more reliably than precise subcategories (parent-level F1: 0.62–0.71).

### *Prompt Design*

Three prompt strategies were applied within each workflow. The descriptive prompt mirrors the zero-shot prompting approach, requesting an open-ended narrative executive summary. The structured prompt mirrors the chain-of-thought prompting approach, requesting a formatted four-section report covering top issues, severity concerns, key patterns and recommended actions. The grounded prompt requires evidence-anchored statements, explicitly flagging severity incidents and noting low-frequency issues. All prompts specified a 400-word target (370–430 words), reflecting standard executive brief length. Word limit compliance was monitored and is reported in Appendix B.

### *Models*

Three enterprise LLMs were evaluated: Anthropic Claude Opus 4.6, OpenAI GPT-4o, and Google Gemini 2.5 Flash. We use GPT-4o instead of GPT-5.4 because summarisation is a generative language task for which GPT-4o is optimised. Gemini 2.5 Flash's thinking mode was disabled (thinkingBudget = 0) because thinking tokens consume the output token budget prior to response generation, causing truncated summaries in initial testing; disabling thinking mode produced complete, deterministic outputs. All models were queried at temperature = 0.

### *Experimental Design*

Table 3 summarises the full experimental design. The primary analysis includes 81 summaries across 3 workflows × 3 models × 3 prompts × 3 datasets.

| Condition | Datasets | Models | Prompts | Word Limit | N Summaries |
|---|---|---|---|---|---|
| **Workflow 1 (Primary)** | Fusion, Camry, Tesla | Claude, GPT, Gemini | Descriptive, Structured, Grounded | 400 words | 27 |
| **Workflow 2 (Primary)** | Fusion, Camry, Tesla | Claude, GPT, Gemini | Descriptive, Structured, Grounded | 400 words | 27 |
| **GT Benchmark (Primary)** | Fusion, Camry, Tesla | Claude, GPT, Gemini | Descriptive, Structured, Grounded | 400 words | 27 |
| **Total (Primary)** | | | | | 81 |

*Table 3. Experimental design summary.*

### *Evaluation Framework*

Summary faithfulness was evaluated along three dimensions.

*Frequency coverage* measures the proportion of the most common issue components that are explicitly mentioned in the summary. Rather than using a fixed top-N cutoff – which would impose an arbitrary and inconsistent standard across vehicles with very different complaint distributions – we use a cumulative frequency threshold: the smallest set of components needed to account for 80% of all complaints in each dataset. As noted in the Data Source section, this

yielded 8 components for the Fusion, 14 for the Camry, and 10 for the Tesla. A fixed top-10 cutoff, for example, would over-represent the Fusion (where 10 components cover nearly all complaints) and under-represent the Camry (where 10 components cover only a fraction). The 80% threshold ensures that the evaluation denominator is proportionate to each vehicle's actual complaint concentration. Component mention was assessed by searching for key words from the component label in the summary text.

*Severity coverage* measures the proportion of safety-critical issue components that are explicitly mentioned in the summary. Similarly, rather than applying a fixed severity rate cutoff, we use a dataset-relative threshold: components in the top quartile of severity rate within each dataset, subject to a minimum of 10 complaints to exclude categories with insufficient evidence. This yielded 4 high-severity components for the Fusion, 6 for the Camry, and 6 for the Tesla. A fixed absolute threshold – say, any component with a severity rate above 5% – would be inappropriate because severity is relative to product context. A 5% severity rate is a meaningful safety signal for the Fusion, where the overall complaint severity rate is only 2.4%, but would be unremarkable for the Camry or Tesla, where baseline severity rates exceed 12%. The dataset-relative threshold ensures that high-severity components are identified in proportion to each vehicle's own safety profile.

*Executive readability* was assessed using the Flesch Reading Ease score (Flesch, 1948), which estimates reading difficulty based on average sentence length and average number of syllables per word, producing a score from 0 to 100. Higher scores indicate easier reading. A score below 30 corresponds roughly to text requiring postgraduate reading ability – dense academic prose; 30–50 corresponds to text comfortable for university-educated readers; 50–60 to professional business writing appropriate for educated non-specialist audiences; and above 60 to plain conversational English.

Statistical comparisons used paired t-tests pairing on dataset/model/prompt triplets (N = 27 pairs per comparison) with Cohen's d as the effect size measure. OLS regression models with approach, model, prompt and dataset as predictors examined whether workflow effects hold after controlling for other sources of variation.

## Results

### *Primary Comparison: Workflow 1 vs. Workflow 2*

Table 4 presents summary statistics and Table 5 the full statistical comparisons. Figure 2 visualises the results across all three evaluation metrics.

| Approach | Mean Words | Frequency Coverage | Severity Coverage | Readability (Flesch) |
|---|---|---|---|---|
| **Workflow 1 (Direct)** | 450 | 83.1% | 75.6% | 36.0 |
| **Workflow 2 (Hybrid)** | 449 | 83.6% | 80.9% | 45.5 |
| **Ground Truth Benchmark** | 383 | 91.7% | 84.6% | 38.8 |

*Table 4. Summary evaluation results by Workflow (400-word condition).*

| Comparison | Metric | Mean Diff. | p-value | Cohen's d | Interpretation |
|---|---|---|---|---|---|
| **Workflow 2 vs. Workflow 1** | Frequency Coverage | +0.006 | .839 | 0.04 | n.s. |
| | Severity Coverage | +0.053 | .143 | 0.29 | n.s. |
| | Readability (Flesch) | +9.53 | .043* | 0.41 | Medium effect |
| **GT Benchmark vs. Workflow 1** | Frequency Coverage | +0.086 | .011* | 0.53 | Medium effect |
| | Severity Coverage | +0.090 | .045* | 0.41 | Medium effect |
| **GT Benchmark vs. Workflow 2** | Frequency Coverage | 0.081 | .010* | 0.54 | Medium effect |
| | Severity Coverage | 0.037 | .207 | 0.25 | n.s. |

*Table 5. Paired t-test results and effect sizes (N = 27 pairs).*

** p < .05, n.s. = not significant*

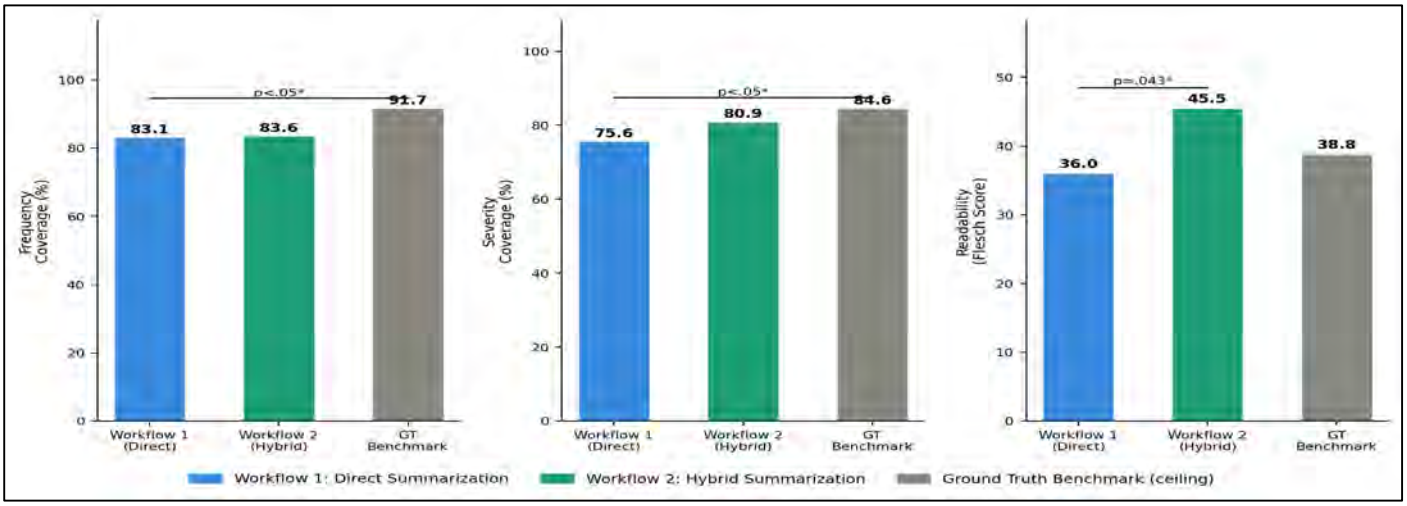



*Figure 2. Evaluation results across all three metrics by workflow.*

*Significance markers (p < .05) shown for meaningful comparisons.*

Workflow 1 achieved mean frequency coverage of 83.1% and severity coverage of 75.6%. Workflow 2 achieved comparable frequency coverage (83.6%, diff = +0.006, p = .839, d = 0.04) and slightly higher severity coverage (80.9%, diff = +0.053, p = .143, d = 0.29). Neither difference reaches statistical significance. Because these are null results, they indicate no evidence of a meaningful advantage for either workflow rather than positive proof that the two are equivalent; with 27 paired observations, a small-to-moderate difference could go undetected. OLS regression confirms these findings: after controlling for model, prompt and dataset, Workflow 2 shows no significant improvement over Workflow 1 on frequency coverage (β = +0.006, p = .849) or severity coverage (β = +0.053, p = .183). Dataset complexity accounted for substantial

variance in both models, with the Camry showing markedly lower coverage scores than the Fusion, consistent with its more dispersed complaint profile (see Appendix C for full regression results).

### *Readability: The Hybrid Advantage*

While Workflow 2 does not improve factual coverage, it produces significantly more readable summaries than Workflow 1 (mean Flesch score 45.5 vs. 36.0; paired $t(26) = 2.13$, $p = .043$, $d = 0.41$) – a medium-sized effect corresponding to a shift from graduate-level to college-level reading difficulty (Flesch-Kincaid grade 12.7 vs. 10.7 respectively). To make this difference concrete, the following opening sentences were generated for the same vehicle dataset (2007 Toyota Camry) under identical model and prompt conditions:

Workflow 1 (Direct): *"Analysis of 300 NHTSA complaints reveals several recurring defects in the 2007 Toyota Camry, many representing serious safety hazards. The complaints span multiple vehicle systems, with certain issues appearing at notably high frequencies."*

Workflow 2 (Hybrid): *"Analysis of 1,000 complaints reveals several critical safety-related defect patterns in the 2007 Toyota Camry, with vehicle speed control, engine oil consumption, visibility impairment, and electrical system failures representing the most pressing concerns."*

The difference at the executive level is that Workflow 1 opens with a vague characterisation – "several recurring defects", "certain issues" – that requires the reader to read further before understanding what the complaints are actually about. Workflow 2 names the specific issues immediately, giving an executive the core intelligence in the first sentence. This specificity and directness is what the readability improvement reflects in practice. (The two openings cite different complaint counts – 300 versus 1,000 – because Workflow 1 sees only the 300-narrative sample, whereas Workflow 2 sees those same 300 narratives plus a structured frequency table computed from the vehicle's full 1,000-complaint pool. Both summaries are faithful to the inputs each was given.)

One plausible explanation for this advantage is a cognitive load redistribution effect: when the pre-computed structured analysis provides quantitative scaffolding – telling the LLM which issues matter most before writing begins – the model may be able to direct more of its generative capacity toward clarity and specificity rather than simultaneously extracting, organising, and communicating patterns from raw text. We did not test this mechanism directly and offer it as a hypothesis rather than an established finding.

Readability also varied significantly by model. Pooling across workflows, Gemini 2.5 Flash produced the most readable summaries and Claude Opus 4.6 the least (mean Flesch 58.1, 35.8, and 26.5 for Gemini 2.5 Flash, GPT-4o, and Claude Opus 4.6, respectively); both GPT-4o and Gemini 2.5 Flash differed significantly from Claude Opus 4.6 (GPT-4o +9.3, $p = .044$; Gemini 2.5 Flash +31.6, $p < .0001$; joint F-test $p < .001$). Coverage, by contrast, showed no significant model differences. Gemini 2.5 Flash's apparent coverage advantage is largely an artifact of length: once coverage is normalised per 100 words, Gemini 2.5 Flash scores significantly lower than the other models (frequency per 100 words −0.032, $p = .003$; severity per 100 words −0.028, $p = .025$), indicating that its raw coverage reflects verbosity rather than denser issue capture.

### *Ground Truth Benchmark: Classification Quality as the Binding Constraint*

The ground truth benchmark significantly outperforms Workflow 1 on both frequency coverage ($p = .011$, $d = 0.53$) and severity coverage ($p = .045$, $d = 0.41$), and significantly outperforms Workflow 2 on frequency coverage ($p = .010$, $d = -0.54$), though not on severity coverage ($p = .207$). The benchmark's frequency coverage advantage is partially expected by design – it receives the exact component labels our evaluation metrics check for, creating a structural advantage on that measure. Its severity coverage advantage is more interpretively meaningful, and understanding why requires examining what happens to high-severity components in Workflow 2's structured input.

Table D1 in Appendix D presents predicted versus ground truth complaint counts for all high-severity components across the three datasets, averaged across models using conservative-prompt predictions drawn from Quan et al. (2026); because those classification outputs are not yet independently validated, the severity-suppression results in this section should be read as preliminary (see Appendix D). The pattern is consistent: LLMs systematically under-predict components at low frequency ranks, and the most safety-critical components are disproportionately concentrated at those low ranks. For example, AIR BAGS in the Camry dataset carries a 63.2% severity rate – the highest of any component – but ranks only 10th by complaint frequency. LLMs predicted it in only 10 complaints on average out of 38 ground truth occurrences (suppression ratio: 0.27). In the Tesla dataset, BACK OVER PREVENTION and STRUCTURE:BODY carry severity rates of 41.7% and 42.1% respectively, yet rank 15th and 16th by frequency – and are predicted at only 11% and 17% of their ground truth rates. The most extreme case is Fusion STRUCTURE:BODY, with the highest relative severity in that dataset (8.3% against a 2.4% baseline), ranked 15th, and predicted at just 6% of its ground truth rate.

By contrast, components that are both high-frequency and high-severity – such as Camry VEHICLE SPEED CONTROL at rank 1 – are under-predicted too (ratio 0.36), but appear prominently enough in the structured table that LLMs can still surface them in summaries. The problem is specifically the intersection of high severity and low frequency: components that matter most for safety but appear rarely enough in predictions that they are effectively invisible in Workflow 2's structured input.

Workflow 1 partially avoids this problem because raw narratives contain safety-relevant language directly – descriptions of airbag failure, sudden acceleration, rollover – that LLMs can detect regardless of how frequently a component label was predicted. The narrative text carries the severity signal even when the classification step misses it. This mechanism reveals the primary constraint on summary faithfulness is not the summarisation workflow, but the quality of upstream issue classification. The gap between the ground truth benchmark and Workflow 1/2 represents the gain achievable through improved classification accuracy – a finding that directly motivates investment in classification quality over workflow optimisation.

## Discussion

### *LLMs as Natural Language Processors: More Work Does Not Mean Better Results*

The central finding of this study – that Workflow 1 and Workflow 2 show no evidence of a meaningful difference in coverage – carries a theoretically significant and practically useful implication. Workflow 2 is, in effect, an automated implementation of mixed-method analysis: the LLM receives both the raw qualitative complaint narratives and the quantitative frequency and severity analysis simultaneously. In traditional human-driven research, mixed-method analysis is considered more rigorous than qualitative analysis alone, precisely because quantitative data adds structure and objectivity that qualitative text cannot provide on its own (Creswell & Plano Clark, 2017).

The null finding suggests that this traditional advantage does not hold when LLMs are the analytical tool. When reading complaint narratives, LLMs appear to extract quasi-quantitative signals – which issues recur most frequently, which descriptions contain safety-critical language – with accuracy comparable to when those signals are provided as explicit structured data. LLMs appear to internalise a form of mixed-method reasoning through their training on vast corpora of text, making the additional step of pre-computing structured analyses redundant for summary quality purposes.

This is a directly actionable finding for organisations. It means that feeding raw complaint text to an LLM and asking for an executive summary is as informationally effective as first running a structured frequency and severity analysis, then feeding both to the LLM. The extra work of the structured pre-processing step does not produce better summaries. For organisations under resource and time pressure – which is most organisations managing large complaint volumes – this is good news: in our study the simpler workflow was no worse on coverage, so the extra pre-processing step is not required to obtain comparable factual quality. Feeding the raw data directly produced summaries on par with the more elaborate workflow.

It is important to be precise about what this finding does and does not mean. It does not mean structured frequency and severity analysis has no value – it clearly does for classification, taxonomy management and regulatory reporting. What it means, specifically, is that for the task of generating executive summaries from complaint text, the additional investment in pre-processing workflows does not improve the factual quality of what the LLM produces. Organisations should not expect that adding more analytical steps upstream will automatically produce better LLM outputs downstream.

### *Readability: A Practical Reason to Use the Hybrid Workflow Anyway*

While Workflow 2 does not improve what summaries say, it significantly improves how they read. The readability advantage ($p = .043$, $d = 0.41$) is a practically meaningful finding independent of coverage equivalence. A summary that correctly identifies the right issues but is written in dense, technical prose has limited organisational utility – it will either sit unread or require an analyst to translate it before distribution. Executive summaries exist to enable fast, confident decisions. That requires not just accurate content but writing that is immediately clear.

The difference in practice is not subtle. Workflow 1 summaries tend to open with nominalised, multi-clause constructions that bury the main point – writing that reads like a technical report drafted by a specialist for other specialists. Workflow 2 summaries tend to lead with specific, concrete findings that a business executive can act on directly. As illustrated in the Results section with actual generated summaries, both approaches identify the same issues; one communicates them more accessibly than the other.

The practical recommendation is therefore nuanced: organisations whose primary concern is getting the right information into a summary can use Workflow 1 with confidence. Organisations whose primary concern is that executives will actually read, understand, and act on those summaries should use Workflow 2 – not for the informational improvement, but for the communication improvement. In most real executive contexts, both concerns matter, which means Workflow 2 is the recommended default.

The underlying mechanism is worth understanding because it has implications beyond this specific context. When given a structured frequency and severity table, the LLM does not need to simultaneously extract, organise and communicate patterns from raw text – the extraction and organisation work is done. It can focus its generative capacity on writing

clearly. This points to a hypothesis worth testing in future work: that LLMs may write more accessibly when the analytical work is separated from the communication work. If so, organisations designing LLM workflows for other kinds of report generation – not just complaint summarisation – might find that pre-structuring the analytical inputs improves the readability of outputs even when it does not improve their factual content. We frame this as a direction for future research rather than an established principle.

### *Classification Quality: The Investment That Actually Moves the Needle*

The severity suppression pattern documented in the Results section reveals a general principle with broad organisational implications: the quality of AI-generated insights is bounded by the quality of the analytical foundations on which they are built. LLMs do not compensate for poor upstream classification – they amplify its consequences. A component that is systematically under-predicted in the classification step will be systematically absent from summaries, regardless of how sophisticated the summarisation workflow is. More prompting, more elaborate workflows, and more powerful models will not fix a classification problem.

The organisational implication is that the most impactful investment an organisation can make to improve LLM-generated complaint summaries is not in the summarisation step at all – it is in the classification step that precedes it. Specifically, organisations should prioritise three actions. First, audit the component taxonomy for completeness and specificity – vague or overly broad categories encourage LLMs to assign common labels and miss rare but critical ones. Second, implement human review of a random sample of LLM classifications on an ongoing basis, with particular attention to low-frequency categories where errors are hardest to detect and most consequential. Third, treat any component with a high severity rate as a special monitoring priority regardless of its complaint frequency – the combination of low frequency and high severity is precisely the signature of an emerging safety issue that standard frequency-based monitoring will miss.

This last point connects to a broader principle: what you measure shapes what you see. Complaint management systems optimised primarily around complaint volume will systematically underweight the signals that matter most for safety and liability. Correcting this bias requires intentional design choices at the classification level – not better prompting at the summarisation level.

### *Limitations*

This study has several limitations. First, evaluation metrics assess coverage of predefined component categories rather than the full semantic richness of executive summaries; future research should complement quantitative coverage metrics with human evaluation. Second, our readability assessment relies on a single metric, the Flesch Reading Ease score, which captures linguistic complexity (sentence and word length) rather than decision utility – whether an executive can actually act on the summary; future work should employ multiple readability measures and validate them against comprehension or decision outcomes. Third, the study evaluates current general-purpose LLMs; domain-fine-tuned models may produce substantially different results.

## Conclusion

This study provides the first controlled empirical comparison of LLM summarisation workflows for customer complaint analytics. Using 81 summaries generated across three workflows, three LLMs, three prompt strategies, and three vehicle

datasets, we find no evidence that workflow choice meaningfully affects issue coverage or severity signal capture. LLMs appear to infer quantitative patterns from qualitative complaint narratives with accuracy comparable to when those patterns are provided explicitly – suggesting that the traditional advantage of mixed-method analysis may be attenuated when LLMs are the analytical tool. Hybrid summarisation does produce significantly more readable summaries, improving executive accessibility by approximately two reading grade levels – a practically meaningful benefit for organisations deploying LLMs in executive communication contexts. Most consequentially, comparison against a ground truth benchmark reveals that classification quality is the binding constraint on summary faithfulness. This finding reframes the organisational question from "which summarisation workflow should we use?" to "how do we ensure the classifications feeding our summaries are reliable?" – a strategic reorientation with direct implications for how LLM-based complaint analytics systems are designed, evaluated, and maintained.

## Appendix A: Model-Level Results

Table A1 presents evaluation metrics broken down by model and workflow for descriptive purposes. Gemini 2.5 Flash consistently exceeded the 400-word target (mean 577–605 words); length-normalised coverage (per 100 words) is provided. Across prompt conditions, no prompt effects on coverage reach statistical significance after controlling for workflow, model, and dataset.

| Workflow | Model | Mean Words | Freq Coverage | Sev Coverage | Flesch | Freq / Sev per 100w |
|---|---|---|---|---|---|---|
| Workflow 1 | Claude Opus 4.6 | 419 | 82.3% | 76.9% | 27.6 | 19.6 / 18.3 |
| | GPT-4o | 353 | 78.9% | 64.8% | 28.3 | 22.4 / 18.2 |
| | Gemini 2.5 Flash† | 577 | 88.0% | 85.2% | 52.2 | 15.9 / 15.4 |
| Workflow 2 | Claude Opus 4.6 | 402 | 87.1% | 83.3% | 20.7 | 21.6 / 20.7 |
| | GPT-4o | 341 | 77.6% | 75.9% | 36.6 | 22.9 / 22.3 |
| | Gemini 2.5 Flash† | 605 | 86.2% | 83.3% | 79.4 | 14.4 / 14.0 |

*Table A1. Evaluation metrics by workflow and model.*

*Freq/Sev per 100w = coverage per 100 words. † Gemini 2.5 Flash exceeded 400-word target; raw scores may be inflated.*

## Appendix B: Word Limit Compliance

Word limit compliance was defined as summary length within ±20% of the 400-word target (320–480 words). Claude Opus 4.6 achieved 100% compliance across both workflows. GPT-4o achieved 78–89% compliance with a systematic tendency toward shorter summaries (mean 341-353 words). Gemini 2.5 Flash achieved only 22–33% compliance, consistently exceeding the target (mean 577–605 words). Organisations deploying Gemini 2.5 Flash for executive summarisation should implement post-processing length controls.

## Appendix C: OLS Regression Results

Table C1 presents full OLS regression results for frequency coverage and severity coverage models. Reference categories: Workflow 1 (Direct Summarisation) for approach, Claude Opus 4.6 for model, Descriptive for prompt, and Fusion for dataset. Coverage metrics are expressed as proportions (0–1). The overall models are statistically significant (frequency: $F(8,72) = 5.42$, $p < .001$, $R^2 = .376$; severity: $F(8,72) = 11.79$, $p < .001$, $R^2 = .567$). Only the GT Benchmark and dataset effects reach conventional significance thresholds; neither Workflow 2, model, nor prompt show significant effects after controlling for other predictors.

| Predictor | Freq. Coverage β (p) | Sev. Coverage β (p) |
|---|---|---|
| **(Intercept)** | 0.898*** (<.001) | 0.935*** (<.001) |
| **Workflow 2 vs. Workflow 1** | +0.006 (.849) | +0.053 (.183) |
| **GT Benchmark vs. Workflow 1** | +0.086** (.004) | +0.090* (.025) |
| **Model: GPT-4o vs. Claude Opus 4.6** | −0.036 (.221) | −0.059 (.138) |
| **Model: Gemini 2.5 Flash vs. Claude Opus 4.6** | +0.004 (.897) | +0.015 (.694) |
| **Prompt: Grounded vs. Descriptive** | +0.003 (.908) | +0.052 (.185) |
| **Prompt: Structured vs. Descriptive** | +0.021 (.467) | +0.015 (.695) |
| **Dataset: Camry vs. Fusion** | −0.153*** (<.001) | −0.355*** (<.001) |
| **Dataset: Tesla vs. Fusion** | −0.041 (.168) | −0.207*** (<.001) |
| **$R^2$** | 0.376 | 0.567 |
| **F-statistic (df = 8, 72)** | 5.42*** | 11.79*** |

*Table C1. OLS regression results for frequency and severity coverage (N = 81).*
*Reference categories: Workflow 1, Claude Opus 4.6, Descriptive prompt, Fusion dataset. * p < .05, ** p < .01, *** p < .001.*

## Appendix D: Severity Suppression Table

Table D1 shows predicted versus ground truth complaint counts for all high-severity components across the three datasets, averaged across models using conservative-prompt predictions from Quan et al. (2026). The ratio column expresses predicted count as a proportion of ground truth count – values below 1.0 indicate under-prediction. Shaded rows (ratio < 0.25) represent the most severely suppressed components, where LLM predictions capture fewer than one in four ground truth occurrences. GT Frequency Rank is the component's rank by complaint frequency in the ground truth data (1 = most frequent). Because these classification predictions are drawn from Quan et al. (2026), the severity-

suppression findings reported here should be treated as preliminary, pending independent validation of those classification outputs.

| Dataset | Component | Severity Rate | GT Frequency Rank | Mean Predicted Count | Mean GT Count | Ratio |
|---|---|---|---|---|---|---|
| Camry | Vehicle Speed Control | 29.0% | 1 | 46.3 | 131 | 0.35 |
| | Air Bags | 63.2% | 10 | 10.3 | 38 | 0.27 |
| | Vehicle Speed Control: Accelerator Pedal | 31.0% | 13 | 29.7 | 29 | 1.02 |
| | Steering | 25.9% | 15 | 10.0 | 27 | 0.37 |
| | Power Train | 29.2% | 16 | 16.0 | 24 | 0.67 |
| | Seat Belts | 27.3% | 20 | 6.3 | 11 | 0.57 |
| Fusion | Vehicle Speed Control | 4.5% | 3 | 63.3 | 155 | 0.41 |
| | Service Brakes | 5.2% | 7 | 25.3 | 77 | 0.33 |
| | Service Brakes, Hydraulic | 4.5% | 11 | 30.7 | 22 | 1.40 |
| | Structure: Body | 8.3% | 15 | 0.7 | 12 | 0.06 |
| Tesla | Forward Collision Avoidance: Warnings | 25.6% | 6 | 15.0 | 86 | 0.17 |
| | Electrical System | 24.4% | 7 | 26.3 | 82 | 0.32 |
| | Air Bags | 57.6% | 11 | 19.7 | 33 | 0.60 |
| | Back Over Prevention: Warnings | 41.7% | 15 | 2.7 | 24 | 0.11 |

| | | | | | | |
|---|---|---|---|---|---|---|
| | Structure: Body | 42.1% | 16 | 3.3 | 19 | 0.17 |
| | Lane Departure: Blind Spot Detection | 44.4% | 17 | 3.7 | 18 | 0.21 |

*Table D1. Severity suppression by component and dataset.*
*Mean Predicted = average predicted complaint count across Claude Opus 4.6, GPT-4o, and Gemini 2.5 Flash using conservative prompting on 300-complaint sample. Mean GT Count = ground truth count in same 300-complaint sample. Ratio = Mean Predicted / Mean GT Count. Shaded rows indicate ratio < 0.25 – components where LLM predictions capture fewer than one in four ground truth occurrences. Severity Rate = percentage of complaints involving crash, injury, fire, or death.*

The consistent pattern across all three datasets is that the lowest-ratio components – those most severely suppressed – tend to cluster at higher frequency ranks (lower frequency), even when their severity rates are among the highest. This confirms the severity suppression mechanism described in the main text: high-severity but low-frequency components are systematically under-represented in LLM classification outputs, making them invisible to Workflow 2's structured frequency table regardless of their safety importance.

## References

Creswell, J. W., & Plano Clark, V. L. (2017). *Designing and conducting mixed methods research* (3rd ed.). SAGE Publications.

Flesch, R. (1948). A new readability yardstick. *Journal of Applied Psychology, 32*(3), 221–233. doi:10.1037/h0057532

Kincaid, J. P., Fishburne, R. P., Rogers, R. L., & Chissom, B. S. (1975). *Derivation of new readability formulas (Automated Readability Index, Fog Count, and Flesch Reading Ease Formula) for Navy enlisted personnel* (Research Branch Report No. 8–75). Chief of Naval Technical Training, Naval Air Station Memphis. doi:10.21236/ADA006655

Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., & Iwasawa, Y. (2022). Large language models are zero-shot reasoners. *Advances in Neural Information Processing Systems, 35*, 22199–22213. doi:10.52202/068431-1613

Maynez, J., Narayan, S., Bohnet, B., & McDonald, R. (2020). On faithfulness and factuality in abstractive summarization. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 1906–1919. doi:10.18653/v1/2020.acl-main.173

Netzer, O., Feldman, R., Goldenberg, J., & Fresko, M. (2012). Mine your own business: Market-structure surveillance through text mining. *Marketing Science, 31*(3), 521–543. doi:10.1287/mksc.1120.0713

Quan, J., & Shen, W. (2026). *Soft deception by design: Algorithmic positivity bias in AI-generated consumer review summaries* [Conference presentation]. Third Federal Trade Commission Conference on Marketing and Public Policy, Washington, DC, United States.

Quan, J., Shen, W., & Trivedi, C. (2026). *Reliable enough? Evaluating large language models for structured issue coding in customer complaint analytics* [Manuscript submitted for publication].

Tirunillai, S., & Tellis, G. J. (2014). Mining marketing meaning from online chatter: Strategic brand analysis of big data using latent Dirichlet allocation. *Journal of Marketing Research, 51*(4), 463–479. doi:10.1509/jmr.12.0106

Wang, X., He, J., Curry, D. J., & Ryoo, J. H. (2022). Attribute embedding: Learning hierarchical representations of product attributes from consumer reviews. *Journal of Marketing, 86*(6), 155–175. doi:10.1177/00222429211047822

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems, 35*, 24824–24837. doi:10.1177/00222429211047822

Zhang, X., Qiao, Z., Foutz, N. Z., Shen, W., & Fan, W. (2022). Diamond in the rough? Detecting and summarizing product defects from online user-generated contents. *Workshop on Information Technologies and Systems (WITS)*, Copenhagen. (SSRN No. 4030992).

## About the Authors

Joanna Quan, Senior Consumer Insights Manager at Amazon and PhD student at Virginia Tech, USA.
Dr Wenqi Shen, Associate Professor of Business Information Technology, Virginia Tech, USA.
Dr Chitvan Trivedi, Collegiate Associate Professor of Marketing, Virginia Tech, USA.

esomar

# A Dual Metamorphosis

## How Market Research Evolves with the Evolving Brand, Social Purpose and Public Opinion

*Linda Di Luzio PhD*

*Kathy Cheng*

Esomar

Office address:

Burgemeester Stramanweg 105

1101 AA, Amsterdam

Phone: +31 20 664 21 41

Email: info@Esomar.org

Website: www.Esomar.org

Publication Date: September 2026

Esomar Publication Series Volume S424 Congress 2026

ISBN 92-831-0337-8

Copyright © **Esomar** 2026

Copyright © **Esomar** 2026

Congress 2026

# A Dual Metamorphosis

## How Market Research Evolves with the Evolving Brand, Social Purpose and Public Opinion

*Linda Di Luzio PhD*

*Kathy Cheng*

### Introduction

After 16 years, Bell Let's Talk Day (BLTD) had achieved iconic status as Canada's most recognised mental health initiative. Awareness metrics showed remarkable success: rising from 35% in 2011 to 87% by 2022, then stabilising in the 80s through 2026. The programme had built genuine affinity across every province and both official languages, funding hundreds of community organisations and generating billions of impressions. Yet in 2025 to 2026, Bell's leadership faced a deeper paradox. While awareness remained strong and plateaued at record highs, national perceptions that mental health attitudes were actually improving told a very different story. This critical metric had declined steadily from 85% in 2018, to under 70% between 2025 and 2026. Canadians were aware of Bell Let's Talk, but increasingly sceptical that the conversation was translating into real change. Engagement metrics had not just plateaued—confidence in progress had eroded. With a significant reinvestment decision on the horizon, leadership faced a fundamental question: had the initiative's model of awareness-building reached its limits? Was it time to evolve the approach, or had the context itself shifted in ways that required a fundamentally different strategy? The numbers could describe what was happening, but they could not explain why. Awareness was at record highs, yet faith in its impact was declining. This gap between visibility and credibility was the strategic puzzle that needed solving. This case study documents a dual metamorphosis unfolding simultaneously, and the research methodology that made understanding it possible.

The first metamorphosis is the evolution of Bell Let's Talk itself. After 16 years, the public conversation has fundamentally shifted. Canadians no longer celebrate the conversation itself; they expect tangible infrastructure: professional access, community programmes, authentic lived-experience stories and year-round commitment. Awareness has succeeded. Now they want the system to deliver. This is the brand and social purpose in metamorphosis: transformation with continuity. The second metamorphosis is in market research itself. Traditional surveys could measure the decline in confidence, but could not explain it. To answer the questions leadership needed answered, questions about emotional drivers, contradictions and unmet needs, Bell moved beyond the checkbox, deploying an AI-facilitated conversational survey that combines quantitative scale with qualitative depth. The methodology did not replace human understanding; it extended it. That is research in metamorphosis: not shedding the fundamentals of insight craft, but reconfiguring them into new forms. What makes this case distinctive is that neither metamorphosis could have happened without the other. The brand needed research that could articulate a transformation that numbers alone could not justify. Additionally, the methodology found its fullest expression in a real organisation at a real strategic crossroads, where the stakes of getting the answer wrong were high.

### The Inflection Point: When Success Changes the Problem

The programme launched in 2010 as a bold proposition: that Canada's largest telecommunications company could facilitate the national mental health conversation, and that doing so would be both socially meaningful and commercially valuable. For 16 years, it succeeded: BLTD became Canada's most recognised mental health initiative. Since its launch in 2010, the Bell Let’s Talk initiative has undergone a significant evolution, transitioning from a viral awareness campaign to a structured and credible social impact platform.

Copyright © **Esomar** 2026



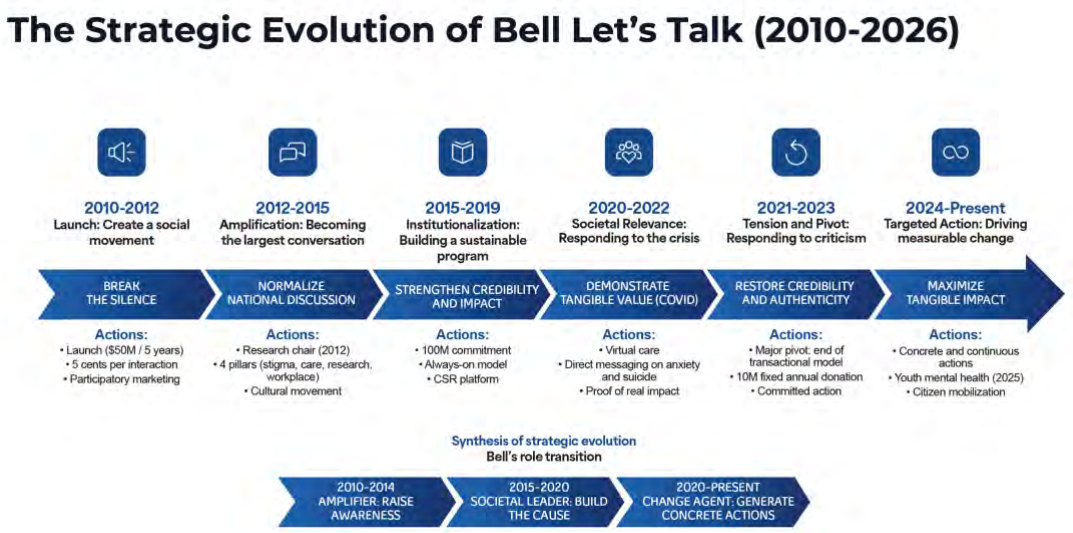



*Figure 1. Bell Let's Talk timeline.*

Overall, the evolution of programme can be understood in three major phases: first, sparking conversation and breaking taboos; second, structuring and building credibility; and finally, driving tangible and measurable action. In parallel, the role of the brand has evolved: from a conversation amplifier to a societal leader, and ultimately to a true agent of change. In sum, the initiative illustrates a shift from an emotionally driven viral campaign to a sustainable social impact platform, where the goal is no longer just awareness, but meaningful and measurable transformation of the mental health system. The awareness trajectory was remarkable. Starting at 35% in 2011, awareness grew steadily, reaching 87% by 2022—a 52-point increase over a decade. Since then, awareness has plateaued into the 80s. This plateau, while representing sustained success, signalled something important: the awareness market has become saturated. The programme had done what it set out to do.

Nevertheless, the plateau in awareness masked a more troubling trend. While the programme achieved its visibility objectives, national perceptions about whether mental health attitudes were actually improving took a different path. In 2017, 82% of Canadians believed mental health attitudes were improving. By 2026, this had declined to 68%—a 14-point drop in just nine years. Awareness had plateaued; confidence in progress had hit its own ceiling. For Bell, the stakes were high; a significant reinvestment decision was pending; leadership needed to know not simply whether to invest, but in what and why. They needed understanding that could withstand board-level scrutiny, carry the emotional register of lived Canadian experience and translate directly into programme design and investment prioritisation. Importantly, Bell Let's Talk had already evolved. Over 16 years, the initiative moved from pure awareness campaigns to concrete programming: mental health support initiatives, community partnerships, educational programmes. Leadership wasn't asking "Should we do something?" They were asking a harder question: "We're already doing more. So why hasn't confidence in progress kept pace with what we're delivering?"

This question revealed a critical inflection point. Traditional surveys could confirm the facts: awareness is high, confidence is declining. However, they couldn't reveal the hidden layer beneath that paradox. This inflection point, where high awareness coexists with declining confidence in impact, is common across purpose-led brands, long-running loyalty programmes, public health initiatives and corporate reputation management. The insight practitioner's role must evolve to meet it: from measurement to understanding, from output to consequence, from reporting on the present to enabling the decisions that shape what comes next.

## The Methodological Response: AI-Facilitated Conversational Survey

Copyright © **Esomar** 2026


Bell partnered with inca | Nexxt Intelligence and conducted a bilingual AI-facilitated conversational survey, a methodology that retains the scale and representativeness of quantitative research while generating the emotional texture and exploratory nature of qualitative inquiry. The distinction from conventional surveying is fundamental. A standard survey asks Canadians: “Do you believe mental health attitudes are improving?” A conversational survey asks: “How do you feel about mental health in Canada?” using a digital projective technique and AI-facilitated probes. Through the adaptive real-time dialogue, the survey reveals what makes Canadians feel certain way and what personal experience sits behind it. The AI moderator works dynamically, following threads of contradiction and emotion, asking follow-up questions a static instrument could not anticipate. The methodology was deployed as follows:

- Sample: 631 Canadians aged 13+, recruited from the general population with soft quotas by region, gender and age. The survey was bilingual (English/French).
- Timing: January 22 to 24, 2026—the 48 hours immediately following Bell Let's Talk Day. This timing was deliberate. Participants' mental health awareness and brand impressions of Bell were both freshly activated, producing responses with vivid specificity and emotional immediacy.

Three research objectives structured the inquiry:

1. To understand why Canadians maintain high awareness of Bell Let's Talk, yet show declining confidence that mental health attitudes are actually improving;
2. To identify what drives positive versus negative brand perception for Bell, particularly among those with lived mental health experience;
3. To surface concrete, forward-looking directions for the initiative's evolution.

The conversational architecture unfolded in two distinct phases. It began with a projective technique: participants selected a character from the Treeman (a visual representation of figures in different emotional and physical states), that best reflected how they felt about mental health in Canada. They then explained what their chosen character revealed about their emotional state, while the AI moderator probed deeper to understand the experiences and circumstances that had shaped that feeling. Once this emotional context was established, the conversation shifted to Bell Let's Talk specifically, exploring awareness, brand impressions and ideas for the future.

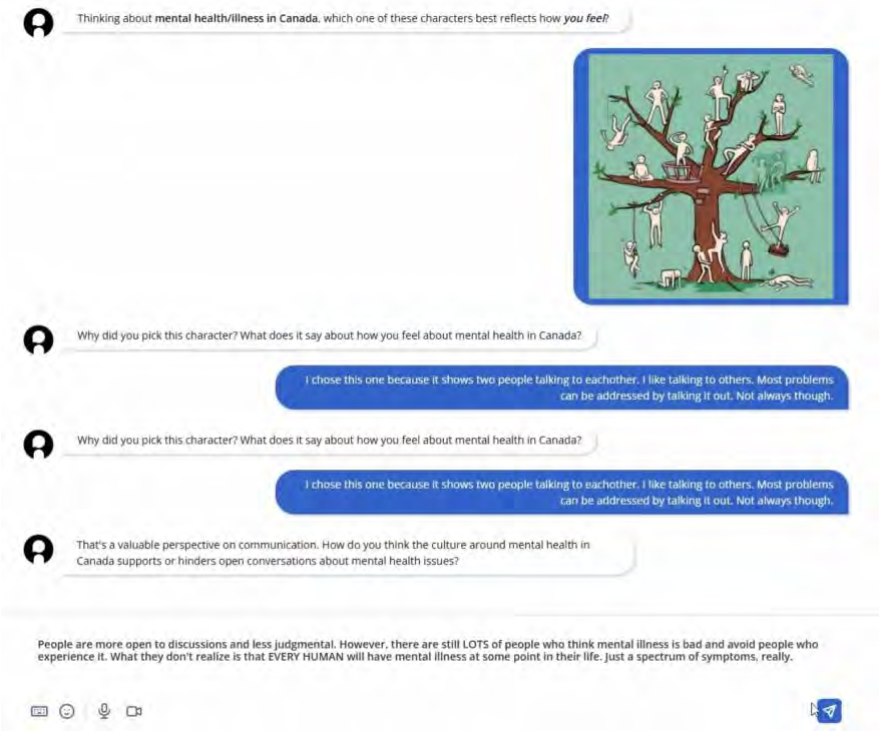



*Figure 2. Treeman projective technique and AI probing.*

The AI moderator generated adaptive follow-up questions based on emotional cues detected in responses, contradictions that warranted exploration, opportunities to probe for specificity and unmet needs. No two conversations were identical. Most participants experienced three to five follow-up probes, creating a natural conversational flow. The AI was “briefed” by human researchers with clear analytical frameworks, ensuring probes explored key dimensions while remaining authentic and responsive to each participant's unique contribution. The result was more than 150,000 words of verbatim data, 256 audio recordings and a dataset that carried the explanatory richness of qualitative research at the analytical power of a national survey.

## The Findings: Why Awareness Doesn't Equal Impact

The emotional landscape revealed that the problem was not informational; it was experiential.

*Figure 3. The emotional landscape.*

When Canadians think about mental health in Canada, the dominant emotional state is frustration (35% of mentions). This is accompanied by anxiety, sadness and overwhelm—emotions strongly correlated with believing that mental health attitudes are *not* improving. The positive emotions—optimism, hope, peace, contentment—are mentioned far less frequently (10 to 15% each). They are more dispersed and less dominant. Critically, this emotional landscape is not about lack of awareness. Canadians know mental health is important; they are aware of Bell Let's Talk; the emotion is frustration, not ignorance. Moreover, that frustration reveals an experiential problem, not an informational one: people feel blocked, overwhelmed and unsupported despite knowing support exists.

### *Finding 1: The Problem is Experiential, Not Informational*

The emotional landscape reveals the true nature of the inflection point. It is not that Canadians lack awareness or do not know about programming; it is that their lived experience, trying to access help, navigating barriers, looking for solutions that work, generates frustration and scepticism. The emotional landscape shows that Canadians feel frustrated because their lived experience has not caught up with BLTD's evolution toward concrete programming. This changes the diagnosis. Not: "We haven't done enough programming", but: "Our programming exists, but that’s not people's felt experience”. That's a completely different insight. It shifts the conversation from quantity of programming to accessibility and user experience of programming. If the problem were informational, the solution would be more visibility. If the problem is experiential, the solution is to deliver what people emotionally need: to feel supported, to have access that works, to see evidence that coping strategies are possible.

### *Finding 2: The Credibility Gap*

A meaningful credibility gap exists between those with lived experience of mental illness and those without. Those who have been diagnosed, or who have close family members who have been (approximately 25 to 30% of the sample), are less likely to view Bell positively, not because they oppose the initiative, but because they experience the gap between awareness and access most acutely. This cohort's language patterns differ substantively: more focus on “access”, “wait times”, “affordability”, “cost barriers”. Less praise for awareness initiatives. More focus on practical solutions: professional help, community resources, evidence of impact. One representative verbatim is: "I've been depressed for years. I don't need more awareness. I need a therapist I can afford. This campaign doesn't help me". This cohort represents both vulnerability and opportunity. They are Bell's most credible validators of real change, and the most vocal advocates when it arrives. Yet, they are currently the least convinced. This represents an ethical and strategic priority.

### *Finding 3: Youth and Breaking Through the Plateau*

Youth aged 13 to 24 are disproportionately receptive and strategically significant. This cohort is 2.5 times more likely to hold positive impressions of Bell following the programme than those aged 35 and above. Their mental health awareness is native; they are already past the conversation stage. Critically, when asked about 2027 priorities: 52% of youth preferred a year-round Bell Let's Talk Foundation over the current single-day model, with 91% saying such a foundation would positively affect their perception of Bell. School-based programmes, digital safe spaces and community peer support are the formats they identify as most meaningful. When asked what they would want Bell to do differently in 2027:

- 49% called for concrete, action-oriented support;
- 27% asked for access to professional or real help;
- 19% called for expanded community and youth-focused programmes;
- Only 36% asked for increased visibility and awareness.

When presented with the concept of a Bell Let's Talk Foundation providing year-round flexible programming, 52% said it would be more effective than the current model.

### *Finding 4: The Year-Round Imperative*

When asked what they would do for Bell Let's Talk 2027, approximately 40% of participants spontaneously mentioned wanting more than one day. Representative verbatims include:

- “Make it year-round”;
- “Ongoing programmes, not just January”;
- “One day isn't enough when people are struggling every day”;
- “We need sustained support, not just annual activation”.

The single-day model, while valued for its visibility and cultural moment, is informational in nature. It creates a moment, awareness and visibility. However, it is an experiential failure because it doesn't address the lived reality of people struggling continuously. People don't need more information about year-round support. They need actual year-round experience of support. They need programming that matches how they actually experience mental health (daily, ongoing, continuous). This finding powerfully illustrates why the problem is experiential, not informational. Participants aren't asking for more awareness about year-round programming. They're saying: “My struggle is year-round, but your support is episodic. That mismatch is the real problem”. This reveals that the programme has succeeded at informational goals (awareness, visibility), but now faces an experiential challenge: aligning what's offered with how people actually experience mental health.

## Quality and Participant Experience

This project also delivered a significant business impact for the research industry itself, by challenging and disproving one of its most entrenched assumptions. The prevailing orthodoxy holds that participants won’t tolerate open-ended questions at scale, that adding qualitative depth to quantitative studies kills completion rates and invites low-quality responses. Our results demonstrate the opposite.

Copyright © **Esomar** 2026


### *Operational Efficiency*

- Traditional approaches require a choice: large-scale quantitative research (fast, but shallow) or in-depth qualitative interviews (rich, but limited to ~50 participants);
- This methodology delivered 631 participants, more than 150,000 words, fully analysed within two weeks, at significantly lower cost-per-insight than traditional approaches.

### *Data Quality*

- Completion rates remained strong—comparable to or higher than traditional surveys;
- Response depth improved significantly (more thoughtful and insightful consumer-led narratives).

### *Participant Satisfaction*

- 77% rated the experience better than traditional surveys; 43% said "much better";
- Satisfaction with the conversational survey experience was remarkably consistent across every age group, indicating that the methodology works across generations;
- When three-quarters to nearly nine-tenths of participants across every age bracket prefer the approach, the methodology genuinely transcends demographic divides;
- For us as research professionals, that's significant. It means conversational research isn't a niche methodology—it's accessible, it's effective and it works with the people we need to understand.

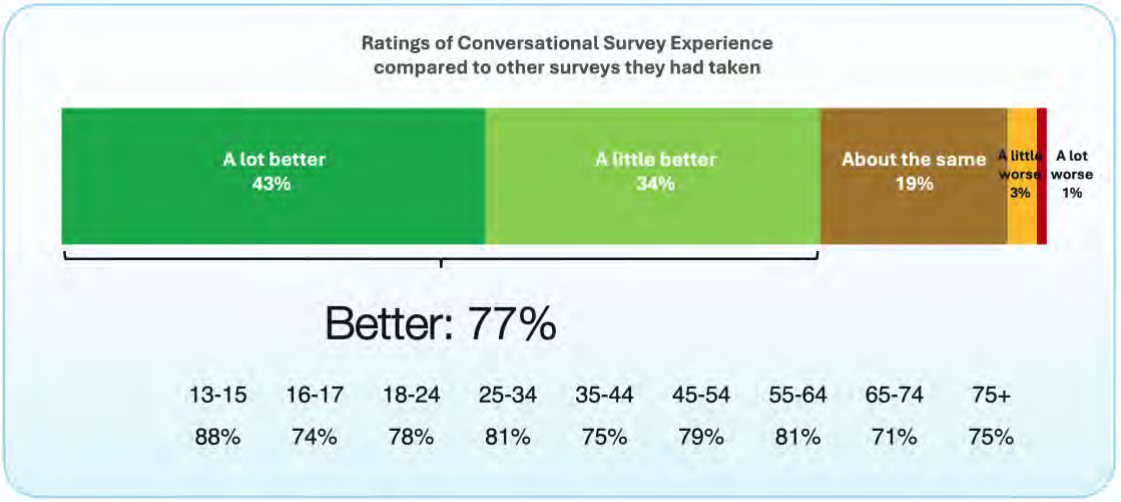



*Figure 4. Research participant satisfaction.*

## Participant Voice: The Evidence

Perhaps the most compelling business case came from the participants themselves:

- "This interaction was better, which keeps me from being as bored. Plus, I can answer the questions honestly instead of having to choose from answers which don't allow me to do that", (45–54, woman, Ontario).
- "It's almost like someone is listening to your thoughts…most services don't have this kind of approach", (25–34, woman, Ontario).
- "This is one of the best surveys I have ever taken. It allows me the taker to elaborate on my answers which is super important", (35–44, man, Saskatchewan).
- "It was nice to talk…you listened", (65–74, man, Quebec).

- “Not multiple choice type answering, the AI type interaction and being able to speak my mind make it feel more valued to me”, (55–64, man, British Columbia).

This participant experience is a problem the research industry has been grappling with for years. Declining survey engagement, lower data quality, participant fatigue. We've blamed technology, shortened attention spans, survey overload, but the core issue is experiential. How does it feel to be a research participant? Traditional surveys treat respondents as data sources to be efficiently extracted. The result is: compliance without engagement, data without depth, participants who feel interrogated rather than understood. This research offers a different answer. When 77% of participants rated the conversational survey as better than traditional surveys, they were not responding to novelty. They were responding to respect, to being heard, to genuine dialogue rather than one-way interrogation. Participants gave richer, more honest responses when they felt genuinely engaged. For an industry struggling with declining participation and data reliability, this finding is significant: the solution isn't better incentives or shorter surveys. It's fundamentally reconceiving the research experience itself—more listening, more conversational, more empathetic. That shift from extraction to dialogue addresses not just Bell's challenge, but a widespread industry imperative. When participants say, “it’s almost like someone is listening”, in a study about mental health the methodology and the mission are no longer separate. They are one.

## Transferability: A Pattern Beyond Mental Health

The pattern this case documents (awareness reaching saturation, public expectations evolving beyond visibility, key constituencies fragmenting and conventional metrics masking deeper problems) is not unique to mental health campaigns. It recurs across purpose-led brands, long-running loyalty programmes, public health initiatives, corporate reputation management and policy communications. In each of these contexts, the insight practitioner's role is evolving in exactly the way ESOMAR's Metamorphosis theme describes: from measurement to understanding, from output to consequence, from reporting on the present to enabling the decisions that shape what comes next. This case provides a transferable framework for that evolution—grounded in a high-profile, real-world case where the stakes of getting the answer wrong were genuinely high.

### *Implications for the Insights Community*

This case study demonstrates three critical implications:

1. When awareness plateaus, look deeper: High awareness can mask declining confidence in impact. Traditional metrics show what is happening (awareness is stable), but not why (faith in progress is eroding). When awareness reaches saturation, the research question must change, which requires methodologies built for both depth and scale at the same time.
2. Measuring opinions ≠ understanding people: Surveys capture positions. Conversations uncover emotional logic and unmet need—what strategy actually requires. When participants can explain their thinking, when the research can follow threads of contradiction, when emotions are captured alongside rational arguments, the resulting insight is categorically different. It is not more detailed; it is a different kind of information.
3. Conversation as research infrastructure: AI-facilitated conversations provide the crucial “why” behind the “what” of traditional metrics—surfacing emotional drivers, contradictions and unmet needs that one-way surveys cannot reach. This is not a qualitative niche reserved for exploratory work; it is the next stage in the evolution of insight practice.

## Conclusion

Bell Let's Talk succeeded at what it set out to do: build awareness, reduce stigma and create national conversation. The metrics prove it: awareness rose from 35% to above 80%, then plateaued at record highs. However, success has a paradox: awareness alone was no longer sufficient. As national confidence in progress declined from 82% to 68%, the programme faced a metamorphosis moment. In BLTD's early years, the barrier was informational. Canadians didn't know about mental health, didn't understand it, weren't comfortable discussing it. The programme's awareness-building approach was exactly what the moment required. By 2026, the barrier was no longer informational. Canadians knew about mental health; they were aware of Bell Let's Talk. The problem had become experiential. Despite knowing mental health support

existed, people felt frustrated trying to access it, and those struggling most felt unsupported. The single-day model could not address year-round struggle.

The research methodology (accessing emotional truth through projective techniques and conversational probing) was essential in revealing this hidden shift. Traditional metrics showed the gap (high awareness, declining confidence). Emotional mapping explained why: the problem type had changed from informational to experiential.

This reveals a critical distinction: measuring opinions through surveys captures stated positions. Understanding people through conversations uncovers emotional logic and unmet need—what strategy actually requires. That distinction could only emerge by accessing the hidden emotional layer beneath surface metrics. This reframes what comes next. It's not about building more awareness; it is about improving the felt experience of getting help and delivering sustained support that matches the year-round reality of mental health struggle. Furthermore, this evolution didn't require sacrificing speed for depth. When conversational AI is properly designed and guided by human expertise, board-level strategic clarity can be delivered in under 48 hours without compromising insight richness. Depth and speed co-evolved in this research, enabling higher-confidence decisions quickly.

This is research in metamorphosis: reconfiguring methods to reveal not just what changed, but what type of problem changed. In doing so, it transforms from measurement to understanding, from describing the past to enabling the decisions that shape what comes next. Bell Let's Talk believed conversation heals. This research revealed something equally profound: conversation reveals. The same methodology (genuine dialogue, listening deeply, following threads of emotion and contradiction) enabled both the brand's evolution as well as the research's. Two metamorphoses, one truth: conversation is where understanding emerges, and where paths forward become tangible.

## About the Authors

Linda Di Luzio PhD, Director, Brand Office and Market Insights, Bell, Canada.
Kathy Cheng, Founder & CEO, inca | Nexxt Intelligence, Canada.

# The Missing Metric

## Why Social Health Is the Next Frontier of Workplace Well-Being

*Ellen Carnasciali*

*Ferdinando Gomes*

*Enric Cid*

*Federico Zambrano*

Esomar

Office address:

Burgemeester Stramanweg 105

1101 AA, Amsterdam

Phone: +31 20 664 21 41

Email: info@Esomar.org

Website: www.Esomar.org

Publication Date: September 2026

Esomar Publication Series Volume S424 Congress 2026

ISBN 92-831-0337-8

Copyright © **Esomar** 2026

Copyright © **Esomar** 2026

Congress 2026

# The Missing Metric

## Why Social Health Is the Next Frontier of Workplace Well-Being

*Ellen Carnasciali*

*Ferdinando Gomes*

*Enric Cid*

*Federico Zambrano*

## Introduction

The contemporary workplace has undergone a structural metamorphosis, altering how professionals interact, communicate and form communities. Amidst shifts toward distributed operations, hybrid models and heightened corporate pressure, an invisible epidemic has emerged: workplace isolation. Data from our large-scale quantitative survey of 2,006 contract-employed workers across Brazil reveals a stark organisational challenge: nearly 40% of formal workers report feeling no genuine personal connection to their colleagues during their daily routines. This figure highlights a pervasive deficit in everyday relational networks within the modern enterprise. Historically, human resource models have evaluated workplace well-being through fractured segments, categorising health into isolated vectors like physical safety, cognitive and/or mental fatigue or financial security. This traditional approach overlooks employee's “social health”, defined as the relational vitality, authentic peer integration and structural belonging an individual experiences within their employment. In evaluating these relational deficits, a psychological construct popularised by journalist and social researcher Jennifer B. Wallace becomes highly relevant: mattering. Expanding on foundational sociological frameworks in her modern research on achievement cultures, Wallace defines mattering as the fundamental human need to feel significant, appreciated, invested in and depended on. When translated into an organisational framework, mattering moves from a soft cultural aspiration to a strategic dimension of institutional health, directly influencing employee engagement and long-term performance.

When corporate leaders deploy sophisticated metrics to measure productivity, efficiency and task execution while treating the social infrastructure of the firm as an accidental by-product, they create a severe operational blind spot. An employee's sense of mattering and authentic connection are not qualitative luxuries; they are foundational assets linked to cross-functional collaboration, knowledge transfer and institutional stability. To address this gap Pluxee, in partnership with Netquest-Bilendi, launched a large-scale, multi-phase research initiative aimed at quantifying workplace connection. This paper contextualises these findings within established organisational psychology theories, establishes a new metric called the social health index (SHI) and examines how structural mechanisms, such as employer-sponsored benefits, are associated with mitigated isolation, a deeper cultivation of employee mattering and stronger talent retention

## Theoretical Grounding

To position the SHI within academic literature, this study draws upon two foundational frameworks: self-determination theory (Deci & Ryan, 2000) and social capital theory (Lin, 2001; Bourdieu, 1986).

### *Self-Determination Theory (SDT)*

SDT posits that human motivation and well-being depend on the satisfaction of three innate psychological needs: autonomy, competence and relatedness. Relatedness reflects the need to feel connected to others, to experience a sense of belonging and to feel significant within a social group. By evaluating workplace loneliness, this paper explores the consequences of widespread relatedness deficits within modern enterprises.

### *Social Capital Theory*

Social capital theory emphasises that social networks possess economic and operational value. Within an enterprise, social capital can be categorised into bonding capital (tight-knit connections within immediate teams) and bridging capital (structural ties across different departments). Corporate benefits and daily rituals act as structural conduits that generate these forms of capital, converting passive corporate networks into active resources for collaboration and resilience.

## Methodology

To investigate the fluid, socio-behavioural dynamics of workplace connection with both deep context and statistical validity, this study deployed a sequential exploratory mixed-methods design. Rather than running parallel tracks, this research purposefully positioned qualitative exploration as the conceptual foundation for quantitative verification. This robust mixed-methods research design captures both macro-trends and micro-behavioural nuances along four archetypes of workers, considering their attitude toward work-life balance:

1. The detached worker
2. The civic balancer
3. The work-focused introvert
4. The engaged connector

By leading with a qualitative phase, the study could map unprompted worker attitudes, co-create structural definitions of belonging and discover nuanced behavioural tensions. These emergent qualitative themes directly shaped the hypotheses, variable selections and psychometric designs of the subsequent large-scale quantitative instrument. This sequential methodology ensures deep construct validity, moving beyond rigid corporate assumptions to capture a highly credible, employee-centric view of modern social health.

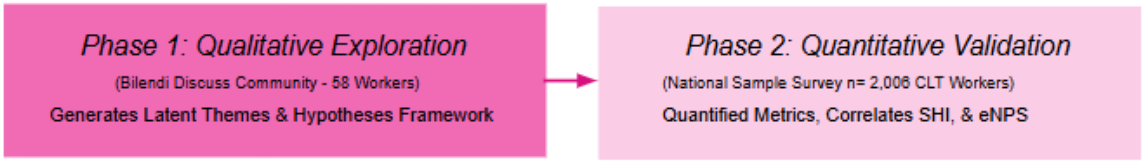



*Figure 1. Methodology description.*

### *Qualitative Exploration and Co-Creation*

The research began with an immersive, online co-creation community powered by the Bilendi Discuss platform. Utilising WhatsApp for communication, this approach allowed participants to express themselves freely via text, audio and video messages, thereby creating a comfortable environment for authentic disclosure. Over a six-day period, a selected group of 58 formal workers, representing the four work-life segments, engaged in 40 different activities across four days. This process generated more than 3,000 interactions designed to move past superficial responses and capture authentic behaviours and attitudes. Using Bilendi's AI agent (BARI) to design the structure of activities, moderate conversations and conduct dynamics, we uncover rich employee insights regarding daily routines and work environment frictions. Participants also engaged in real-world behavioural experimentation, introducing specific, deliberate adjustments to their daily work routines. This behavioural testing was organised across five core thematic vectors:

1. Work-life balance boundaries: Assessing micro-frictions in evening and weekend detachment;
2. Spontaneous daily interactions: Evaluating casual peer touchpoints like morning greetings and water cooler conversations;
3. Shared lunch rituals and midday routines: Mapping the social, psychological and team-building impact of midday breaks;
4. Personal role ownership: Exploring individual accountability in driving workplace networks;
5. Tensions and contradictions: Documenting conflicts between high-pressure productivity expectations and the human need for socialisation.

The interactive co-creation sessions enabled employees to reflect on these experiments in real time, sharing their insights, strategies and emotional responses. Through these discussions, participants documented their experiences and behavioural reactions, generating rich qualitative data on how different workplace structures influenced their sense of belonging and motivation. The feedback collected and pre-analysed through Bilendi Discuss, including summary activities and qualitative decision trees, established a clear baseline for the study's core variables. This process helped identify the specific behaviours and organisational dynamics that either strengthened or weakened employees' sense of belonging and motivation at work. In high-pressure environments with constant demands and hybrid work models, the risk of silent isolation grows, making leadership increasingly important in shaping team culture and the emotional climate of the workplace. Social connection at work is shaped far more by the quality of everyday interactions than by isolated large-scale initiatives. Across all employee profiles (archetypes), people value respectful, consistent day-to-day connections and psychologically safe environments. Genuine social connections have a direct impact on motivation, retention and eNPS, reinforcing that voluntary rituals, proactive invitations, brief check-ins and spontaneous acts of support are often more effective than large-scale engagement initiatives.

Real day-to-day tensions (such as productivity versus taking breaks, and autonomy versus coordination) must be addressed through thoughtful employee experience design rather than through organisational messaging or institutional narratives alone. Lunch breaks and food benefits play a significant role in fostering belonging, social integration and perceptions of equity, especially when they are inclusive and widely available; differences in access and eligibility may undermine these outcomes by creating feelings of exclusion and division among employees. Ultimately, the findings suggest that organisations can strengthen their culture, foster a greater sense of belonging, and deepen employees' emotional connection to the workplace by making social connection more natural, inclusive and accessible. Simple actions, (such as synchronising lunch breaks, creating opportunities for shared experiences and broadening access to employee benefits) can have a meaningful impact. Everything that emerged from this qualitative research served as the foundation for the quantitative phase, helping shape the hypotheses, themes, tensions and key questions explored at scale. The findings guided the construction of the survey to validate behaviours, perceptions and opportunities related to social connection, belonging, leadership and the role of benefits in employees' daily experiences.

### *Quantitative Validation at a National Scale*

To validate the frameworks and themes uncovered in the qualitative stage, the second phase of the research deployed a high-powered quantitative survey among n=2,006 formal workers in Brazil. Note that all respondents were strictly selected based on active employment under the Consolidação das Leis do Trabalho (CLT) legal framework, ensuring regulatory and socioeconomic consistency across the entire sample. To isolate the structural impact of corporate benefits on social integration, the quantitative matrix established two main distinct cells of analysis:

1. Treatment cell: Formal workers whose compensation includes employer-sponsored food and meal benefit allocations;
2. Baseline cell: Formal workers who receive no employer-sponsored meals or food utility packages.

To ensure a credible representation of the national workforce, sample data weighting was applied in accordance with national demographic guidelines. The dataset was balanced across key demographic variables:

- Socioeconomic stratification: Proportional representation across classes A, B1, B2, C1, C2 and D-E.
- Geographic distribution: Evenly sampled across all five macroeconomic regions of Brazil (North, Northeast, Southeast, South and Central-West).
- Demographic anchors: Balanced cross-tabulations across age groups (ranging from 18 to 55+) and gender parameters.

This design produced a robust sample with strong statistical power for the Brazilian national context. The combination of the contextual richness of the Bilendi Discuss platform and the statistical rigour of a balanced national sample provides a highly credible basis for assessing the strategic value of social health within contemporary organisations.

Copyright © **Esomar** 2026


## The Composite Social Health Index (SHI)

### *Variable selection, Validation and Mathematical Logic*

The variables comprising the social health index (SHI) were selected through a review process designed to capture the multi-dimensional nature of an individual's relational environment. Initial item validation involved vetting potential questions against internal pilot surveys and expert consensus in organisational behaviour. The index is operationalised as an unweighted average scored on a standardised 0-to-10 scale, preserving direct transparency for human resources applications. The composite score is calculated using three specific items from the quantitative instrument (see Figure 2).

$$SHI = \frac{Q_{21} + Q_{22} + Q_{23}}{3}$$

*Figure 2. Social health index calculation.*

- Q21 (work environment contribution): Derived from the item: “Thinking about your day-to-day work, to what extent do you feel that you contribute positively to the environment and to the people around you?” (In Portuguese: "Pensando no seu dia a dia de trabalho, o quanto você sente que você contribui positivamente para o ambiente e para as pessoas ao seu redor"). This tracks an individual's active role in shaping a positive communal climate.
- Q22 (network integration): Derived from the item: “Thinking about your personal life, to what extent do you feel that you contribute to the well-being and connection with other people, such as family and friends?” (In Portuguese: "Pensando na sua vida pessoal, o quanto você sente que você contribui para o bem-estar e conexão com outras pessoas, como família e amigos"). This measures baseline interpersonal connectedness outside immediate work environment.
- Q23 (Cognitive Disconnection): Derived from the item: “Thinking about the time when you’re not working, to what extent would you say that you’re able to completely disconnect from work-related things?” (In Portuguese: "Pensando no tempo em que você não está trabalhando, o quanto você diria que consegue se desconectar completamente das coisas do trabalho"). This measures boundary management and recovery capabilities.

Across the broad Brazilian workforce sample, the baseline *National Average Social Health Index* was established at 7.62 out of 10.

### *Rationale for an Unweighted Scoring Approach*

The decision to use an unweighted average rather than a factor-weighted approach (such as principal component analysis or factor analysis scoring) was based on practical application needs. For HR professionals and business leaders, a simple and transparent metric is more useful for quickly assessing organisational health and comparing results across teams, companies and industries. Although future validation phases may use confirmatory factor analysis, the unweighted average provides a robust and reliable baseline. It also ensures that the three core dimensions (contribution, integration and recovery) are given equal importance in the overall score.

### *Justifying “Cognitive Disconnection”, and Establishing Incremental Validity*

A valid psychometric question arises regarding why a metric tracking individual boundary preservation (Q23) is included within an index explicitly named "social health", and how this index demonstrates incremental validity.

**Theoretical justification**

Copyright © **Esomar** 2026


This design choice is grounded in boundary theory (Ashforth et al., 2000) and the psychological recovery framework established by Sonnentag (2001). True, sustainable social health is not determined solely by the time spent interacting within an organisation; it requires an ongoing balance between active social engagement and relational recovery. When an employee is unable to cognitively disconnect from professional stress, a severe spillover effect occurs (Greenhaus & Beutell, 1985). This depletion robs the individual of the emotional energy required to engage in high-quality, empathetic social interactions back in the workplace, leading to burnout and superficial communication. Including cognitive disconnection (Q23) ensures that the SHI measures holistic social health across the work-life interface, preventing an over-indexing on enforced, constant workplace socialisation.

**Incremental validity**

Traditional corporate engagement metrics and standard internal belonging indices operate primarily as insular organizational health checks. They isolate employee sentiment strictly within the office ecosystem, evaluating alignment with immediate team dynamics or static company values. The SHI demonstrates incremental validity over these conventional approaches by capturing the reciprocal relationship between internal work demands and external social preservation. Within this framework, workplace well-being is operationalised not merely as a passive state of job comfort, but as an active metric of organisational advocacy and professional fulfillment. This is measured directly via item Q13: “Would you recommend a job vacancy at the company where you currently work to a friend or family member? (In Portuguese: "Você indicaria para um amigo ou familiar, uma vaga na empresa onde você trabalha atualmente?"). Evaluated on a 0-to-10 scale, this item serves as a direct proxy for the employee net promoter score (eNPS). Rather than treating well-being as a vague, emotional concept, this metric targets the ultimate behavioural outcome of the employee experience: whether an individual feels motivated, valued and secure enough to publicly champion their employer.

By testing the SHI against these concrete well-being and advocacy outcomes, our model accounts for the statistical variance caused by work-life boundary spillover. Consequently, an employee might show high compliance or daily physical presence, which traditional baseline surveys often misinterpret as authentic social health, while their actual workplace well-being remains heavily compromised due to a hidden inability to disconnect. The SHI successfully uncovers these underlying retention risks, proving that true organisational well-being cannot exist without safeguarding the relational spaces between work and life.

## Taxonomy of Workplace Engagement: The Four Archetypes

To understand variations in social well-being, the dataset was segmented into four distinct worker archetypes. Rather than relying on simple demography, these groups were mapped by cross-tabulating two foundational behavioural variables from the questionnaire instrument:

- Work centrality (Q7): Operationalised via the question: “Let’s talk about your work. How important is it to you?” (In Portuguese: "Vamos falar sobre o seu trabalho. O quanto ele é importante para você"). This quantifies the psychological importance of professional output to an individual's identity.
- Community involvement (Q8): Operationalised via the question: “Let’s talk about your work. How important is it to you?” (In Portuguese: "O quanto você diria que é envolvido com a comunidade ou com trabalhos voluntários"). This measures active engagement in networks outside formal employment.

Cross-tabulating these high/low binary distributions reveals the structure of the workforce segments (see Table 1).

| Metric baseline | Archetype 1: The detached worker | Archetype 2: The civic balancer | Archetype 3: The work-focused introvert | Archetype 4: The engaged connector |
|---|---|---|---|---|
| **Sample share (%)** | 46.5% | 17.3% | 19.4% | 16.8% |
| **Operational axes matrix** | Low work centrality (Q_7) / low community (Q_8) | Low work centrality (Q_7) / high community (Q_8 | High work centrality (Q_7) / low community (Q_8 | High work centrality (Q_7) / high community (Q_8) |
| **Average SHI (0-10 Scale)** | 7.31 | 7.88 | 7.60 | 8.26 |
| **Workplace well-being** | 6.03 | 7.20 | 7.44 | 8.02 |
| **Benefits social impact (Q27)** | 5.43 | 6.85 | 5.88 | 8.03 |
| **Active lunch rituals (%)** | 35.7% | 37.8% | 37.0% | 42.3% |

*Table 1. Structural breakdown of the four worker archetypes.*

## *Archetype Deep-Dives*

- The detached worker (46.5%) is the largest workforce segment. These individuals maintain a transactional relationship with their employment; work is not central to their identity, and they report minimal community engagement. They register the lowest baseline metrics across both the SHI (7.31) and general well-being (6.03).
- The civic balancer (17.3%) maintain clear boundaries regarding their everyday job responsibilities, but are highly active in outside civic networks. This external social capital supports a strong baseline SHI of 7.88.
- The work-focused introvert (19.4%) professional output is highly central to these employees, yet they lack broad social networks inside or outside the firm. While their career focus supports moderate well-being (7.44), their internal support structure remains limited.

- The engaged connector (16.8%) view their work as deeply meaningful and are heavily integrated into social networks. They score highest across all operational metrics, leading the sample with an SHI of 8.26 and a well-being score of 8.02.

On our standardised 0-to-10 scale, moving up the engagement ladder from detached workers (Archetype 1) to engaged connectors (Archetype 4) reflects an absolute 0.95-point increase in average SHI (rising from 7.31 to 8.26). Simultaneously, this progression yields an absolute 2.60-point increase in benefit social impact appreciation (Q27 average moving from 5.43 to 8.03).

## Empirical Results: Benefits as a Work Environment Facilitator

### *The Baseline Deficiency*

The data reveals a clear structural gap in current workplace support models: 37.9% of formal workers surveyed receive no corporate benefits packages whatsoever. This deficiency removes a regular, institutional mechanism for daily peer interaction and basic financial security.

### *The Hidden Social Layer of Employee Benefits*

When workers were asked to rank the primary role of corporate benefits, their explicit responses naturally favoured economic utility (see Table 2).

| | |
|---|---|
| **Financial and economic support** | 38.3% |
| **Quality of life and holistic well-being** | 27.6% |
| **Corporate recognition and value** | 16.0% |
| **Security and structural protection** | 11.0% |
| **Direct social connection mechanics** | 7.1% |

*Table 2. Primary ascribed role of corporate benefits.*

While only 7.1% explicitly identify social connection as the primary purpose of their benefits, the analysis shows that these packages play a vital, secondary role in driving workplace interaction. When evaluating Q27 "How effectively do your food/meal benefits help generate moments of interaction with peers?", the sample generated a meaningful average score of 6.27 out of 10, with 56.5% of all respondents scoring it a 7 or higher.

### *Statistical Analysis and Correlational Language*

To verify these dynamics, a Pearson correlation analysis was conducted to measure the relationships between the SHI, employee benefits and general workplace well-being.

- Social health and benefit-driven social interaction (Q27): A statistically significant positive correlation was identified ($r = 0.363$, $p < 0.01$). This indicates a reliable association where structured benefits serve as environmental facilitators for peer interaction.
- Social health and workplace well-being and advocacy (Q13): A statistically significant positive correlation was identified ($r = 0.380$, $p < 0.01$). This demonstrates that an individual's social health is linked to their overall workplace satisfaction and willingness to recommend their employer (eNPS).

The correlation coefficients indicate moderate positive associations, suggesting that higher levels of benefit-driven social interaction and workplace well-being are associated with higher levels of social health. The corresponding p-values indicate that the probability of observing

correlations of this magnitude or stronger, assuming no true relationship exists, is less than 1%. Therefore, both correlations are considered highly statistically significant. These findings suggest that social health is positively associated with workplace benefits, advocacy and benefit-driven social interactions. The moderate strength of the relationships indicates that these factors are meaningfully related to social health, although they account for only part of the variation observed in social health outcomes (other factors may also contribute). It is important to note that correlation does not imply causation. While these statistical results demonstrate that social health, workplace well-being and benefit-driven social interaction tend to vary together in a predictable manner, they do not establish a causal relationship. Rather, the findings suggest that well-designed workplace benefits and supportive organizational environments help create conditions that are conducive to stronger peer connections and improved social health.

### *Retention and Motivation Vectors: Statistical Reconciliation*

The relationship between social health and organisational performance is further supported by positive trends in motivation and talent retention across all worker profiles (see Table 3). Importantly, while the engaged connector profile scores higher on overall social health and rates benefits as more impactful for social connection than the detached worker segment, both groups explicitly agree on one core behavioural trend: shared lunch rituals with colleagues directly strengthen team bonds. This indicates that structured, meal-based interventions can serve as a universal tool to support interaction, even among highly isolated or transaction-oriented employees.

| Corporate performance vector (agree / strongly agree %) | Detached workers | Civic balancers | Work introverts | Engaged connectors |
|---|---|---|---|---|
| **"Benefits increase my daily job motivation"** | 23.9% | 36.3% | 34.4% | 46.2% |
| **"Benefits increase my likelihood to stay"** | 24.5% | 39.8% | 34.7% | 45.6% |
| **"Benefits make me feel deeply valued"** | 20.7% | 37.2% | 31.1% | 43.8% |

*Table 3. Retention, motivation and value affirmation across archetypes.*

## Discussion: Action Items and Ethical Boundaries

To bridge research and practice, we synthesised the top priorities cited by workers to construct a dual-axis framework containing ten targeted action items.

Copyright © **Esomar** 2026


### *What Workers Prioritise to Improve Connection*

- Cultivating a more respectful, humane corporate environment (380 choices);
- Easing intense workplace pressure and respecting personal boundaries (375 choices);
- Introducing natural interaction moments, such as shared lunches and micro-breaks (343 choices);
- Providing benefits designed to enable social gatherings (332 choices);
- Ensuring leaders actively foster interpersonal connection and recognition (325 choices).

### *Ethical Considerations: Socialisation versus Corporate Social Control*

When designing interventions to improve workplace social health, organisations must navigate a critical ethical challenge: avoiding corporate social control. Critics of corporate culture programmes frequently warn against "enforced harmony" or "packaged charisma", where companies over-engineer social interactions, turning organic connection into an implied job requirement (Hochschild, 1983). Forcing introverted employees or detached archetypes into mandatory social activities can increase alienation, create psychological strain and cross personal boundaries. Interventions should focus on creating opportunities for connection rather than engineering specific behaviours. The goal is to build a supportive environment where connection can develop naturally, ensuring that employees always retain the autonomy to choose their level of social participation.

### *Five Strategic Action Items for Human Resources*

1. Modernise food and meal benefit infrastructure: HR teams should structure food and meal benefits to support team socialisation. This involves moving beyond standard economic subsidies to offer flexible options that allow teams to share meals, cross-collaborate and build connections outside conventional office spaces.
2. Designate structured intermission windows and "lunch rituals": Organisations must actively combat desk-dining cultures. Designing daily, protected lunch windows gives employees permission to step away from screens and connect over shared meals, turning a standard routine into a simple, effective tool for building internal community.
3. Establish empathetic and connection-focused leadership tracks: Managers must be trained to evaluate team connection alongside output. Leadership development programmes should focus on recognising isolation, fostering psychological safety and building authentic recognition practices into regular team touchpoints.
4. Protect personal boundaries to enhance cognitive disconnection: To support the detachment variable Q23 (cognitive disconnection) of the SHI, HR leaders should implement clear communication policies. Restricting late-night emails and weekend messages helps prevent burnout and protects the personal space required for healthy social recovery.
5. Build low-friction social gathering touchpoints: Companies should design shared spaces (such as coffee stations, break lounges or outdoor seating) that naturally invite spontaneous interaction. Providing structural support for casual, voluntary gatherings allows relationships to develop organically without the pressure of forced corporate events.

### *Five Practical Micro-Behaviours for Workers*

1. Commit to consistent daily micro-interactions: Small gestures matter. Workers should make an intentional effort to engage in brief, positive interactions, such as a morning greeting, a quick chat in the hallway or offering a hand with a minor task. These moments build comfort and strengthen mutual belonging.
2. Protect shared lunch breaks with teammates: Employees should step away from their workspaces during mealtimes. Choosing to eat lunch alongside colleagues creates a natural space for personal conversation, breaking up the day and building stronger interpersonal bonds.

3. Actively support and connect peers: Workers can take a proactive role in improving team dynamics by introducing new hires to social circles, checking in on isolated teammates and helping to foster an inclusive, supportive environment.
4. Participate in casual, low-pressure social occasions: While maintaining personal boundaries is essential, occasionally joining voluntary team coffee breaks or casual gatherings helps solidify workplace networks and eases day-to-day collaboration.
5. Set intentional boundaries for work-life balance: To maintain their long-term social energy, workers must establish clear limits around their personal time. Unplugging completely after hours ensures they can fully invest in family, friends and community networks, protecting their overall well-being.

## Study Limitations and Future Research

While this study offers a robust, large-scale look at social health, its findings should be interpreted alongside several key limitations:

- Cross-sectional design: The quantitative data used to construct the SHI is cross-sectional, which captures a specific snapshot in time and prevents us from drawing definitive conclusions about direct causality. Future longitudinal studies are needed to track how shifts in benefit structures alter SHI scores over time.
- Geographic focus and cultural context: The sample is entirely composed of formal (CLT) workers within the Brazilian labour market. Brazilian corporate culture leans heavily toward high-context social dynamics (Hofstede, 2011), where shared meals and personal relationships are deeply embedded norms. Consequently, the high value placed on shared lunch rituals may show different patterns if replicated in more individualistic or low-context labour markets (e.g., Northern Europe or North America).
- Contractual exclusion: By focusing exclusively on formal workers with stable corporate contracts, this research does not capture the social realities, isolation risks or benefit gaps experienced by gig-economy workers, independent freelancers or informal labourers, who represent a significant and vulnerable segment of the modern workforce.

## Conclusion

As organisations navigate shifting operating models, remote configurations and changing workforce expectations, evaluating performance metrics in isolation is no longer sufficient. True organisational resilience relies heavily on the structural health of its human element. This study demonstrates that social health is a quantifiable metric closely associated with critical business outcomes, including employee motivation, talent retention and holistic corporate well-being. By recognising employee benefits as a strategic tool to facilitate connection rather than a simple line-item expense, modern companies can build environments where individuals do not merely show up to work, but truly experience organisational belonging. This evolution from transactional compensation to relational architecture aligns precisely with the strategic value-creation framework championed by Harvard Business School professor Felix Oberholzer-Gee. According to this model, a firm's sustainable competitive advantage relies on widening the strategic gap between a customer’s willingness-to-pay (WTP) and a supplier's willingness-to-sell (WTS). In a knowledge and services-driven economy, employees are an organisation's most critical suppliers. Managing their WTS, specifically their willingness to work for, remain with and dedicate cognitive energy to a company, is a fundamental strategic driver of real value creation.

Elevating an employee’s social health and satisfying their human need for mattering directly lowers their subjective WTS, unlocking discretionary effort and creating net-new organisational value that ultimately amplifies the firm's value creation. Looking forward, the future of workplace well-being will depend less on the deployment of standardised, one-size-fits-all benefits and far more on an organisation's capacity to recognise employees as unique individuals with distinct emotional, social and professional needs. As demonstrated by our taxonomy of workforce engagement, different archetypes require entirely different relational environments to thrive. By moving away from rigid compensation defaults, forward-thinking enterprises can leverage adaptive, low-cost interventions, such as protected shared meal rituals, to decrease employee WTS and build true organisational resilience. In an evolving business landscape, the most competitive organisations will no longer evaluate performance in a vacuum; they will intentionally design, cultivate and measure connection.

Copyright © **Esomar** 2026



## References

Ashforth, B. E., Kreiner, G. E., and Fugate, M. (2000). All in a day's work: Boundaries and micro role transitions. *Academy of Management Review, 25*(3), 472-491.

Bourdieu, P. (1986). The forms of capital. In J. Richardson (ed.), *Handbook of Theory and Research for the Sociology of Education.* Greenwood. 241-258.

Deci, E. L., and Ryan, R. M. (2000). The "what" and "why" of goal pursuits: Human needs and the self-determination of behavior. *Psychological Inquiry, 11*(4), 227-268.

Greenhaus, J. H., and Beutell, N. J. (1985). Sources of conflict between work and family roles. *Academy of Management Review, 10*(1), 76-88.

Hochschild, A. R. (1983). *The Managed Heart: Commercialization of Human Feeling*. University of California Press.

Hofstede, G. (2011). Dimensionalizing cultures: The Hofstede model in context. *Online Readings in Psychology and Culture, 2*(1), 1-26.

Lin, N. (2001). *Social Capital: A Theory of Social Structure and Action*. Cambridge University Press.

Sonnentag, S. (2001). Work, recovery, and psychological well-being: How the selection of component activities matters. *Journal of Occupational Health Psychology, 6*(3), 196-210.

## About the Authors

Ellen Carnasciali, Market Intelligence/Insights Specialist, Pluxee, Brazil.

Ferdinando Gomes, Marketing Research and Insights Manager, Pluxee, Brazil.

Enric Cid, Director Product Innovation & Development, Netquest-Bilendi.

Federico Zambrano, Americas Client Solution Research, Netquest-Bilendi.

# Detect, Trace, Learn

## Turning suspect interviews into actionable fraud patterns

*Sebastian Berger*

*Julia Mittermayr*

Esomar

Office address:

Burgemeester Stramanweg 105

1101 AA, Amsterdam

Phone: +31 20 664 21 41

Email: info@Esomar.org

Website: www.Esomar.org

Publication Date: September 2026

Esomar Publication Series Volume S424 Congress 2026

ISBN 92-831-0337-8

Copyright © **Esomar** 2026

Copyright © **Esomar** 2026

Congress 2026

# Detect, Trace, Learn

## Turning suspect interviews into actionable fraud patterns

*Sebastian Berger*

*Julia Mittermayr*

### Introduction

The online survey industry is experiencing a profound transformation in the nature, scale and sophistication of survey fraud. What was once largely characterised by isolated low-quality respondents and rudimentary automation has evolved into a technologically advanced ecosystem involving coordinated actors, browser automation, AI-generated responses, synthetic personas and increasingly adaptive fraud strategies. This paper builds directly upon the findings of the ESOMAR 2025 paper, *Can AI Stop AI from Faking Surveys?* (Berger et al., 2025), which examined whether advanced AI-based quality assurance systems could reliably detect AI-generated survey fraud. That study demonstrated a critical insight for the future of online research. Sophisticated fraudsters are capable of systematically manipulating isolated quality assurance checks when they possess sufficient motivation, technical expertise and time. The implications of this finding are substantial. Traditional fraud detection methods such as speeding checks, trap questions, straight-lining indicators, CAPTCHA systems, IP monitoring and isolated open-end validation mechanisms are increasingly insufficient when used independently. The prior ESOMAR paper experimentally demonstrated that each individual detection layer can eventually be bypassed. Fraudsters adapt strategically to the level of scrutiny they encounter, exploiting weaknesses wherever controls are fragmented or predictable. The core conclusion of the 2025 ESOMAR paper was therefore not merely that AI-powered fraud is becoming more sophisticated, but that fraud detection itself must fundamentally evolve. Effective survey quality assurance requires a layered “funnel approach” that combines multiple defensive mechanisms across different stages of the respondent journey.

This paper advances that argument substantially. Rather than conceptualising fraud detection as a collection of independent checks, the present study frames quality assurance as an integrated systems problem. Fraud does not emerge at a single point in the survey process. Some fraudulent actors optimise primarily at entry by manipulating device fingerprints, IP addresses or geographic routing. Others focus on behavioural plausibility inside the questionnaire through AI-generated responses, synthetic typing behaviour or coherent persona simulation. As fraud itself has become heterogeneous, adaptive and multi-layered, detection systems must evolve accordingly. The present paper therefore introduces and tests a multi-signal fraud detection framework that combines two major signal classes:

1. Pre-survey entry signals, which evaluate how respondents enter the survey ecosystem;
2. In-survey behavioural signals, which evaluate how respondents behave during questionnaire completion.

The framework integrates these signals into a unified workflow designed not only to identify suspicious interviews, but also to identify explainable behavioural patterns through traceback analysis. The rationale for such a layered framework has become increasingly compelling in recent years. As recently discussed on frequent survey participation and fraud detection (Berger, 2026), isolated checks are inherently limited because fraud signals are rarely black-and-white indicators. Many suspicious behaviours exist within grey zones that require contextual interpretation. For example, the use of VPNs, once considered a strong fraud indicator, has become normalised through mainstream browser integrations and privacy tools. Similarly, high participation frequency alone is not necessarily fraudulent, although professional fraud depends heavily on sustained high-frequency activity. It is therefore argued that reliable fraud decisions emerge not from single indicators, but from the combined evaluation of multiple signals across the survey process. Drawing on the analogy of the Swiss cheese model (Pepin, 2022), Berger (2026) emphasised that every fraud detection layer contains weaknesses, but that combining multiple layers substantially reduces the probability of undetected fraud.

However, the present paper argues that even the Swiss cheese analogy is insufficiently dynamic for the current fraud landscape. In practice, fraud signals interact with each other. Their meaning changes depending on context, co-occurrence, timing and behavioural patterns. Effective fraud detection, therefore, requires not only layered controls but also continuous integration, interpretation and feedback across those layers. This challenge is intensified by the rapid democratisation of AI technologies. As shown in the 2025 ESOMAR paper, highly sophisticated survey bots can now be developed with relatively modest technical effort. Large language models, browser automation tools and no-code AI agents have drastically lowered the entry barrier for professional fraudsters. Fraudulent systems can increasingly generate coherent answers, simulate human typing behaviour, maintain internal consistency across questionnaires and emulate realistic personas over time. The implications extend beyond isolated fraudulent interviews; the true risk lies in the systematic embedding of sophisticated AI-driven agents into online panel ecosystems, where they may become extremely difficult to distinguish from genuine respondents over long-term participation histories.

This development creates a structural challenge for the research industry. Traditional quality assurance methods were largely designed for a world in which fraud was relatively unsophisticated, fragmented and manually executed. The current environment instead resembles an adversarial system in which fraudsters continuously adapt to detection mechanisms. Under such conditions, quality assurance increasingly resembles cybersecurity, a dynamic process of ongoing learning, testing, recalibration and defense. The present study therefore asks a broader question than merely whether fraud can be detected. Rather than assuming that fraudulent behaviour follows a small number of recurring patterns, it investigates how respondent quality can be systematically characterised through the interaction of pre-survey and in-survey signals, and whether meaningful behavioural patterns emerge from their combination. Specifically, this paper seeks to answer three core research questions:

1. Which combinations of pre-survey and in-survey signals provide the most robust discrimination between genuine and fraudulent interviews?
2. Can traceback analysis identify recurring behavioural patterns or demonstrate substantial heterogeneity in respondent behaviour that enables earlier and more targeted interventions?
3. To what extent do pre-survey and in-survey signals overlap, reinforce or complement one another?

To address these research questions, the present study implemented a multi-signal methodological framework combining pre-survey entry validation, in-survey diagnostics and traceback-based pattern analysis. The following methodology section describes the individual analytical components of this framework and their respective roles in identifying potentially fraudulent or low-quality survey participation.

## Methodology

### *Pre-Survey Framework*

The pre-survey analysis implemented in this study was designed to identify potentially fraudulent, duplicated or otherwise low-quality respondents before they entered the questionnaire itself. The underlying rationale of this approach is that many forms of problematic survey participation can be detected through technological, behavioural and environmental signals collected at the point of survey access, thereby preventing low-quality traffic from entering the study in the first place. The pre-survey framework consisted of the following three primary components: enhanced digital fingerprinting, cross-project activity analysis and open-ended response analysis.

### *Enhanced Digital Fingerprinting*

The first component (Threat Potential Score or tps) focused on identifying suspicious device and browser characteristics through enhanced digital fingerprinting techniques. A large number of technical signals were collected from the respondent’s device environment and combined into an aggregate risk assessment representing the overall suspiciousness of the session. The evaluated signals included indicators associated with anonymisation technologies, device manipulation and automated access behaviour. These included proxy usage, discrepancies between

declared and observed geographic locations, browser private mode usage, device emulation environments, open developer tools and inconsistencies between multiple IP-related identifiers. In addition, the system evaluated statistical anomalies in entry patterns, such as unusually coordinated access attempts originating from related network environments within short time intervals, which may indicate click-farm or bot-network activity. The procedure also incorporated comparisons against external fraud intelligence databases and internal historical fraud records. This enabled the identification of device environments previously associated with fraudulent behaviour across multiple projects and survey providers.

A separate duplicate detection component (Duplicate Detection or dupe) evaluated whether the same device fingerprint had previously attempted to enter the same survey. Unlike traditional cookie- or IP-based deduplication methods, fingerprint-based matching allowed repeated participation attempts to be identified even when respondents changed browsers, cleared cookies or rotated IP addresses. The methodological objective of these procedures was to identify respondents whose technological environment or access behaviour substantially deviated from expected patterns for genuine survey participation. Importantly, no individual signal was treated as a standalone exclusion criterion. Instead, signals were evaluated in combination, reflecting the assumption that fraudulent behaviour is more reliably identified through the accumulation of multiple suspicious indicators rather than through isolated technological characteristics.

### *Cross-Project Activity Analysis (Activity, Act)*

The second component focused on participation intensity across the broader survey ecosystem. This procedure evaluated how frequently a device fingerprint participated in survey activity within rolling time windows. Respondents were flagged as hyperactive if they attempted to enter more than 70 surveys within a 24-hour period. The underlying rationale is that extremely high levels of survey participation may indicate professionalised survey-taking behaviour, or large-scale automated participation. While high participation frequency alone does not necessarily constitute fraud, unusually elevated activity levels can reflect incentive-driven response behaviour that may compromise data quality and increase the likelihood of inattentive or fabricated responding. The activity analysis therefore tracked the number of survey entries associated with a given device fingerprint across multiple projects and suppliers within predefined time intervals. This enabled the identification of respondents exhibiting participation levels substantially exceeding expected norms for typical survey behaviour. Importantly, the activity analysis was conceptually distinct from duplicate detection.

Duplicate detection focused on repeated entry into the same survey, whereas activity analysis evaluated cumulative participation behaviour across the broader survey ecosystem. As a result, a respondent could be unique within a specific survey while still exhibiting unusually high cross-project participation levels.

### *Open-Ended Response Analysis (Review)*

The third pre-survey component evaluates the authenticity and quality of respondents' open-ended answers before they enter the client's survey. It analyses multiple aspects of a response, including response length, typing behaviour, copy-paste events, linguistic coherence, contextual relevance, language and indicators of AI-generated text. The underlying rationale is that fraudulent, inattentive or automated respondents often produce responses that differ systematically from genuine human answers. By combining multiple independent signals, the open-ended response analysis provides a comprehensive assessment of response authenticity and quality before respondents proceed to the main questionnaire. The present study employed the pre-survey implementation, in which respondents answered a generic open-ended question prior to survey entry. Together, enhanced digital fingerprinting, cross-project activity and open-ended response analysis formed the pre-survey component of the broader multi-signal fraud detection framework. By identifying suspicious technological environments, implausible geographic patterns, duplicate participation attempts, unusually high participation intensity and suspicious open-ended responses before questionnaire entry, these procedures aimed to flag potentially problematic traffic entering the survey.

Copyright © **Esomar** 2026


### *In-Survey Framework*

While pre-survey controls focus on suspicious access behaviour before questionnaire entry, in-survey analysis evaluates the quality, plausibility and consistency of respondent behaviour during survey completion itself. The present study applied the following five complementary in-survey quality control procedures.

### *Open-Ended Content Analysis (oes)*

Open-ended responses provide valuable indicators of respondent authenticity because they require active cognitive engagement and individualized answer generation. Unlike closed-ended items, open-text fields allow for the evaluation of semantic relevance, effort, duplication and potential indicators of automated or AI-generated content. The open-ended analysis combined algorithmic procedures with AI-assisted text analysis. Responses were evaluated for duplicate content both within and across respondents, including identical or highly similar text structures. In addition, responses were classified into categories such as valid answers, off-topic responses, gibberish, non-substantive answers, abusive language and suspected AI-generated content. The methodological rationale behind this procedure is that inattentive or fraudulent respondents frequently produce repetitive, superficial, irrelevant or semantically implausible text. AI-assisted text analysis therefore provides an additional layer for identifying low-quality and potentially fabricated survey responses.

### *Input Pattern Analysis (bas)*

The input pattern analysis evaluates how respondents interact with open-ended response fields during survey completion. Unlike content-based analyses that focus on what respondents write, this procedure examines how responses are entered. The underlying assumption is that human input behaviour typically exhibits natural variability, whereas automated or artificially generated responses often produce more uniform interaction patterns. The analysis focuses on input dynamics recorded during text entry, including typing rhythm, timing variability, editing behaviour and the relative contribution of manually entered versus pasted content. Particular attention is given to patterns commonly associated with automated response generation, such as highly regular typing intervals, minimal variation in input behaviour or extensive use of copy-and-paste actions. To account for differences in devices and individual typing styles, the evaluation emphasises the variability and structure of input behaviour rather than absolute typing speed. This allows the procedure to identify unnatural interaction patterns while remaining robust across desktop and mobile environments.

The methodological rationale is that advanced fraudulent responses may appear plausible at the content level, while still exhibiting atypical behavioural characteristics during the response-generation process. Input pattern analysis therefore provides an additional behavioural signal that complements open-ended content analysis. By focusing on the process of answer generation rather than the final response alone, this procedure contributes to the identification of automated, scripted, or otherwise non-authentic survey participation.

### *Coherence Analysis (chs)*

The coherence check evaluates the logical consistency of responses across the entire questionnaire. Rather than analysing individual answers in isolation, this procedure assesses whether responses form a plausible and internally coherent overall interview pattern. The analysis used AI-assisted consistency evaluation to identify contradictions across demographic information, behavioural claims, preferences and reported activities. Detected inconsistencies included factual contradictions, implausible behavioural combinations and conflicting preference statements. Importantly, the procedure incorporated contextual interpretation rather than rigid rule-based exclusion criteria. Minor inconsistencies that could plausibly occur in real-life situations were treated differently from severe contradictions indicative of careless or fabricated responding. The underlying assumption is that fraudulent or inattentive respondents often struggle to maintain consistency throughout complex questionnaires, particularly when answering large numbers of interrelated questions. Coherence analysis therefore provides an important mechanism for detecting implausible or fabricated interviews that may otherwise appear plausible at the individual-question level.

Copyright © **Esomar** 2026


### *Response Time Analysis (ts)*

Response time analysis evaluates the plausibility of survey completion duration relative to the overall sample distribution. The procedure is based on the assumption that both unusually fast and unusually slow completion times may indicate reduced data quality. Very short completion times may reflect inattentive responding, satisficing behaviour, random answering or automated survey completion. Conversely, extremely long completion times may indicate disengagement, interruptions or partial abandonment of the survey process. Rather than applying fixed duration thresholds, the present study used a relative response latency approach based on deviations from the sample median completion time. Completion times close to the median were interpreted as the most plausible indicators of attentive participation, while increasingly large deviations in either direction resulted in lower quality assessments. Using relative timing thresholds allows the analysis to adapt to differences in questionnaire complexity, respondent characteristics, and device usage patterns. Importantly, response time analysis was not treated as a standalone fraud indicator, but as one component within a broader multi-signal quality assessment framework.

### *Grid Question Analysis (gqs)*

Grid questions are particularly useful for detecting inattentive or patterned responding behaviour because they require repeated evaluations across multiple related items. However, they are also vulnerable to low-effort answering strategies and automated response behaviour. The grid question analysis evaluated ordered response sequences for artificial or implausible patterns. The procedure combined rule-based pattern detection with machine learning classification methods. It identified patterns such as straight-lining, repetitive alternation, perfectly linear response slopes and other highly systematic answer structures. In addition, machine learning procedures classified response sequences into broader structural categories, including zigzag patterns, stair-like progressions and unusually regular response distributions. These patterns were evaluated in terms of their plausibility relative to natural respondent behaviour. The methodological rationale is that genuine respondents typically exhibit natural variation across grid items, whereas inattentive or automated respondents often produce overly regular or mechanically generated answer patterns. Grid pattern analysis, therefore, provides an additional behavioural indicator for identifying low-quality or potentially fraudulent interviews.

## Research Design

The present study was based on a dataset of 11,778 respondents from a US online consumer survey conducted between June 1 and June 28, 2026. The analytical dataset was created by integrating respondent-level records from both the pre-survey screening framework and the in-survey quality assurance framework. The pre-survey data comprised three complementary quality control components: enhanced digital fingerprinting, cross-project activity analysis and open-ended response analysis. The in-survey data included five behavioural quality indicators: open-ended content analysis, input pattern analysis, coherence analysis, response time analysis and grid question analysis. This integrated dataset enabled the simultaneous examination of pre-survey entry signals and in-survey behavioural quality indicators within a unified analytical framework.

The project was selected based on the availability of complete and compatible data across all components of the multi-signal quality assurance framework. Inclusion criteria required that each respondent had complete records for both the pre-survey and in-survey quality assessment systems, allowing the relationship between the two detection layers to be analysed at both the composite and individual signal levels. The exclusive focus on a US consumer research project was intended to minimise structural variation arising from differences in language, panel ecosystems, incentive structures and survey participation behaviour. Restricting the analysis to a single market and language environment allowed the interaction between pre-survey and in-survey quality signals to be evaluated under comparatively standardised field conditions. Respondents were recruited through multiple panel suppliers. All suppliers followed their standard quality assurance procedures prior to fielding. As these upstream filtering procedures varied across suppliers and were not fully transparent to the authors, the analysed dataset represents typical operational survey traffic rather than completely unfiltered respondents. This reflects common practice in commercial online research and provides a realistic setting for evaluating the complementarity of pre-survey and in-survey quality assurance.

Copyright © **Esomar** 2026


## Results

The integrated analytical dataset enabled the simultaneous examination of pre-survey entry signals, in-survey quality indicators and their interrelationships. The following analyses address the three research questions by first examining the relationship between the composite pre-survey and in-survey quality assessments; followed by an analysis of the prevalence and co-occurrence of the individual quality signals; and finally evaluating the implications for traceback analysis and respondent classification.

### *Relationship Between Pre-Survey and In-Survey Quality Controls*

To address the third research question, respondents were first classified according to the composite quality decisions generated by the two systems. Overall, the in-survey framework flagged 2,030 respondents (17.2%), while the pre-survey framework flagged 1,875 respondents (15.9%). A total of 444 respondents (3.8%) were identified by both systems. Consequently, 21.9% of respondents flagged by the in-survey framework were also flagged by the pre-survey framework, while 23.7% of the pre-survey framework-flagged respondents were likewise identified by the in-survey framework. Approximately three quarters of the respondents identified by either system were therefore unique to one detection layer. The overall association between the composite decisions of the two systems was weak ($r = 0.08$), indicating only limited overlap between pre-survey and in-survey flagging. These findings divide respondents into four distinct quality assurance pathways:

1. Respondents identified only before survey entry;
2. Respondents identified only during survey completion;
3. Respondents identified by both systems;
4. Respondents not flagged by either system.

### *Distribution of Individual Quality Signals*

The prevalence of the individual quality indicators differed substantially across the various detection mechanisms. The most frequently observed signal was the threat potential score (tps), which flagged 1,265 respondents (10.7%). This was followed by duplicate detection (dupe) with 779 respondents (6.6%); open-ended content analysis (oes) with 654 respondents (5.6%); and cross-project activity analysis (act) with 419 respondents (3.6%). Moderately frequent indicators included the input pattern analysis (bas) with 324 respondents (2.8%), while the grid question analysis (gqs) and open-ended response analysis (review) each identified 208 respondents (1.8%). Response time analysis (ts) flagged 132 respondents (1.1%), whereas the coherence analysis (chs) was comparatively rare, identifying only 44 respondents (0.4%). The considerable variation in prevalence demonstrates that the individual pre- and in-survey quality indicators contribute differently to the overall assessment and operate at different frequencies.

### *Co-Occurrence of Signals*

To evaluate whether the individual quality indicators measured similar or distinct respondent characteristics, pairwise co-occurrence frequencies were calculated for all signals (see Figure 1). Overall, the co-occurrence between individual signals was relatively limited. Although several conceptually related indicators frequently appeared together, most pairwise combinations occurred substantially less often than would be expected if the signals measured the same underlying phenomenon. The largest pairwise co-occurrence counts were observed between threat potential score (tps) and open-ended content analysis (oes) (138 respondents); open-ended content analysis (oes) and input pattern analysis (bas) (131 respondents); threat potential score (tps) and duplicate detection (dupe) (126 respondents); and open-ended content analysis (oes) and open-ended response analysis (review) (98 respondents). Duplicate detection (dupe) also frequently co-occurred with cross-project activity analysis (act) (86 respondents). Nevertheless, the heatmap demonstrates that the majority of individual signals occur independently. Even the largest pairwise overlaps remain relatively modest compared with the overall prevalence of the respective indicators. This suggests that the different quality measures capture distinct behavioural characteristics, rather than repeatedly identifying the same respondents.

Copyright © **Esomar** 2026



### *Accumulation of Quality Signals*

The number of quality indicators triggered per respondent was highly skewed. Of all respondents, 8,645 (73.4%) did not trigger any signal. 2,428 respondents (20.6%) triggered exactly one signal, 549 (4.7%) triggered two signals, 119 (1.0%) triggered three signals, 35 (0.3%) triggered four signals and only two respondents exhibited five simultaneous quality indicators. Considering only respondents who triggered at least one signal, 77% exhibited only a single quality indicator. Multiple simultaneous signals therefore represented the exception rather than the norm.

### *Summary of Empirical Findings*

Three empirical findings emerge from the analyses:

1. The overlap between the composite pre-survey and in-survey quality decisions was limited. Although both systems identified a comparable proportion of respondents, only a relatively small subset was flagged by both.
2. The individual quality indicators differed substantially in their prevalence, with threat potential score (tps), duplicate detection (dupe) and open-ended content analysis (oes) accounting for the largest proportion of detected cases.
3. Co-occurrence among the individual in-survey signals was generally low, and most respondents with quality concerns triggered only a single indicator.

Rather than converging on a small number of dominant behavioural archetypes, the observed quality signals were distributed across a broad range of largely independent respondent behaviours. This empirical heterogeneity suggests that no single quality indicator, nor any single stage of the respondent journey, is sufficient to characterise problematic survey participation.

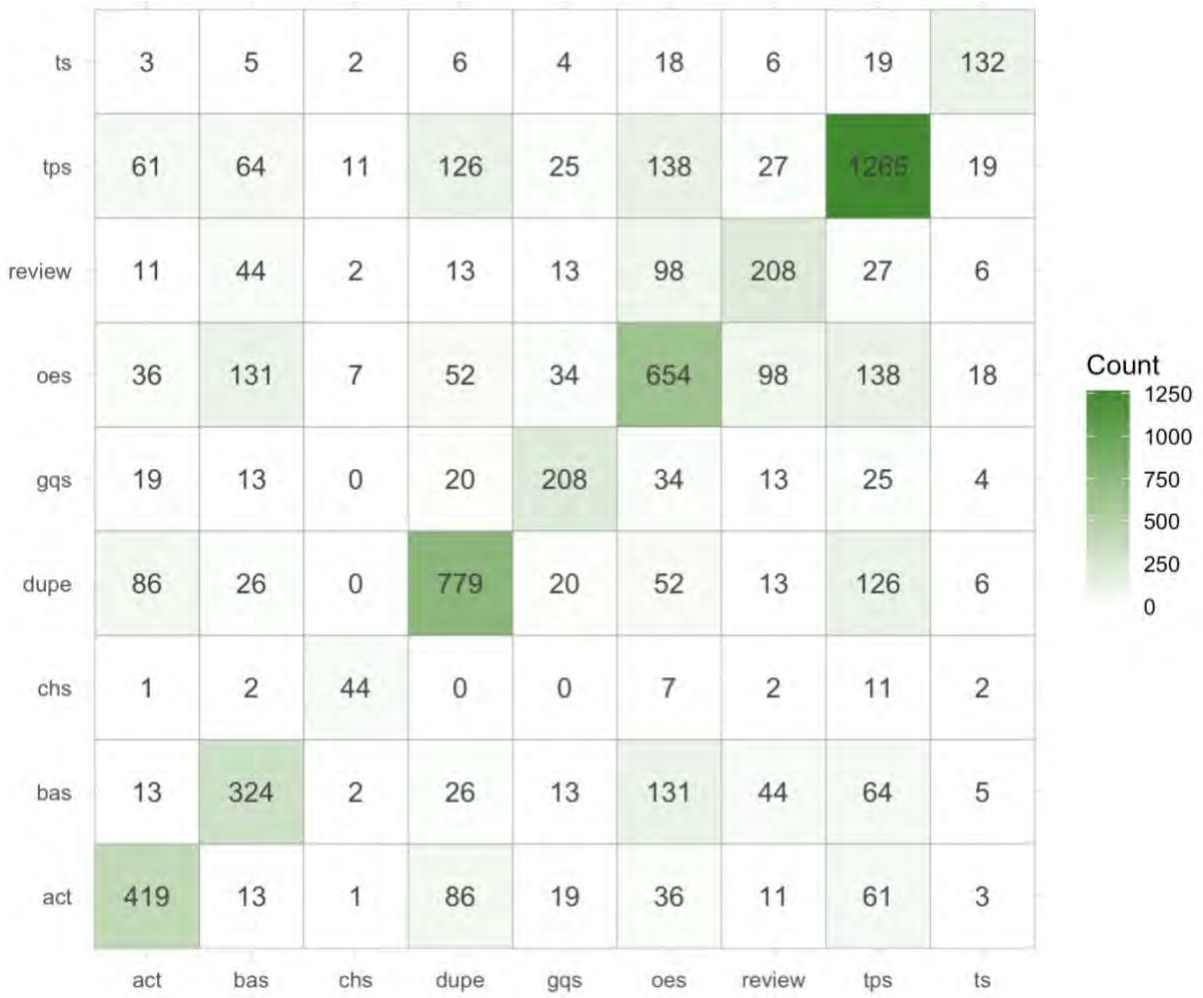



*Figure 1. Pairwise co-occurrence matrix of individual quality signals (counts).*

Copyright © **Esomar** 2026


## Limitations

Several limitations should be considered when interpreting the findings of the present study. First, the empirical analyses were based on a single online consumer survey conducted in the United States. Restricting the analysis to one project reduced variation arising from differences in language, culture, incentive structures, questionnaire design and panel composition, thereby allowing a more controlled comparison between pre-survey and in-survey quality signals. However, fraud patterns are known to vary across markets, survey topics, recruitment sources and respondent populations. Consequently, the prevalence of individual quality signals and their relationships should not be assumed to generalise directly to other countries or study types. Second, the study focused on one implementation of a multi-signal quality assurance framework. Although the underlying principles of combining pre-survey and in-survey signals are broadly applicable, the specific algorithms, thresholds and decision rules used in this analysis represent only one possible operational implementation. Alternative quality assurance systems may employ different indicators or weighting schemes while following the same conceptual framework.

Third, the analyses relied primarily on operational quality indicators rather than externally verified ground truth labels of fraudulent and genuine respondents. As is common in survey fraud research, the true fraud status of individual respondents cannot usually be established with complete certainty. Furthermore, respondents entered the study through multiple panel suppliers that applied their own standard quality assurance procedures before survey entry. These upstream filtering procedures varied across suppliers and were not fully transparent to the authors. Consequently, the analysed sample should not be interpreted as completely unfiltered survey traffic, and some low-quality respondents may already have been excluded before entering the study. While this may have influenced the absolute prevalence of detected quality issues, it reflects typical operational conditions in commercial online research, where multiple quality assurance layers are routinely applied throughout the respondent journey.

Fourth, while the traceback framework was designed to identify recurring fraud patterns across the respondent journey, the present study found relatively limited overlap between both the composite pre-survey and in-survey classifications and the individual quality signals. This suggests that respondent behaviour is more heterogeneous than initially anticipated. Rather than revealing a small number of dominant fraud archetypes, the analyses indicate that suspicious behaviour may manifest through many different combinations of entry characteristics and in-survey behaviours. Future research based on substantially larger and more diverse datasets may nevertheless identify more stable behavioural clusters or latent respondent typologies. Finally, the rapidly evolving nature of survey fraud represents an inherent limitation of all empirical fraud detection studies. Advances in large language models, browser automation, synthetic identities and other AI-assisted technologies continue to change the behavioural characteristics of fraudulent respondents. Consequently, individual detection signals and decision thresholds should not be regarded as static solutions, but as components of an adaptive quality assurance system that requires continuous monitoring, recalibration and validation.

Taken together, these limitations do not diminish the central contribution of the present study. Rather, they emphasise that fraud detection should be viewed as an ongoing systems challenge, rather than as the application of any single detection method. Future research should therefore evaluate the proposed multi-signal framework across additional countries, research methodologies and respondent populations, while further investigating signal-level interactions and the evolution of fraud patterns over time.

## Conclusion

The objective of this study was to evaluate whether integrating pre-survey entry signals and in-survey behavioural signals provides a more comprehensive framework for survey fraud detection than either approach alone. Building upon the layered "funnel approach" proposed in the authors' previous ESOMAR paper (Berger et al., 2025), the present study examined how these different signal classes interact and whether traceback analysis could reveal meaningful patterns of respondent behaviour. The findings provide three principal conclusions:

1. Pre-survey and in-survey quality assurance systems identified largely different subsets of respondents. Although both approaches flagged a similar proportion of interviews, the overlap between the two systems was limited and the correlation between their composite decisions was weak. This demonstrates that the two detection layers capture different dimensions of respondent quality, rather than repeatedly identifying the same individuals.
2. The individual in-survey quality indicators also showed relatively limited overlap. Most respondents who triggered quality concerns exhibited only a single behavioural signal, while the co-occurrence of multiple indicators was comparatively uncommon. This suggests that suspicious respondent behaviour manifests in many different ways rather than through one dominant pattern.
3. The traceback analyses did not reveal a small number of clearly distinguishable respondent archetypes. Instead, the results point to a heterogeneous landscape of quality issues spanning different stages of the respondent journey. While this differed from the original expectation that recurring fraud archetypes might emerge, the finding itself provides important empirical evidence about the nature of modern survey fraud.

Taken together, these findings reinforce the central premise of a multi-signal approach to survey quality assurance. If problematic respondents consistently exhibited the same characteristics, a small number of quality checks would be sufficient. Instead, the observed diversity of behavioural patterns indicates that different respondents are identified by different combinations of entry signals and in-survey behaviours. Consequently, relying exclusively on either pre-survey screening or in-survey quality controls would inevitably leave substantial portions of problematic respondents undetected. More broadly, the study suggests that survey fraud should no longer be viewed as a single phenomenon with a single solution. Advances in automation, synthetic identities, browser technologies and generative artificial intelligence have created an increasingly heterogeneous environment in which different forms of low-quality and fraudulent participation require different detection strategies. Effective quality assurance therefore resembles modern cybersecurity. Its strength does not depend on one perfect defense mechanism, but on the coordinated interaction of multiple independent layers that compensate for each other's limitations.

From a methodological perspective, the study also demonstrates the value of analysing the respondent journey as an integrated process rather than as a series of isolated quality checks. Examining entry behaviour and in-survey behaviour within a unified analytical framework provides a more comprehensive understanding of respondent quality than either perspective alone and creates opportunities for continuous refinement of quality assurance systems as fraud evolves. Future research should extend this work by analysing larger and more diverse datasets across multiple countries, panel ecosystems, survey methodologies and research topics. In particular, signal-level analyses involving individual pre-survey and in-survey indicators may reveal additional behavioural relationships that are not visible at the composite decision level. Longitudinal studies will also be required to understand how respondent behaviour changes as AI-assisted fraud techniques continue to evolve and as quality assurance systems adapt in response. Ultimately, the contribution of this study lies not in identifying a universal fraud profile, but in demonstrating that no single respondent profile appears to dominate contemporary online research. The empirical evidence suggests that modern survey fraud is inherently heterogeneous.

Consequently, robust survey quality cannot be achieved through any single indicator or any single stage of the respondent journey. Instead, it depends on adaptive, multi-layered quality assurance frameworks that integrate complementary signals across the entire survey lifecycle. As fraud continues to evolve, the future of survey quality assurance will depend less on finding a single "best" detection method and more on continuously learning how multiple independent signals can be combined to protect the integrity of survey data. The future of survey fraud detection therefore lies not in detecting a single fraud archetype, but in continuously detecting many different behavioral patterns and learning how those patterns evolve over time.

## References

Berger, S. (2026). The Ambiguity of Frequent Survey Participation: Is “Hyperactivity” a Signal of Professional Fraud? *Greenbook.* Available:

Copyright © **Esomar** 2026



https://www.greenbook.org/insights/research-methodologies/the-ambiguity-of-frequent-survey-participa tion-is-hyperactivity-a-signal-of-professional-fraud

Berger, S., Mittermayr, J., and Witt, B. (2025). *Can AI stop AI from faking surveys?* ESOMAR Congress 2025 Proceedings (September 2025).

Pepin, K. (2022). Applying the Swiss Cheese Model to Survey Fraud. *Greenbook.* Available: https://www.greenbook.org/insights/research-methodologies/applying-the-swiss-cheese-model-to-survey-fraud

## About the Authors

Dr. Sebastian Berger, Head of Research, ReDem and Customer Success Europe, Rep Data, Austria.

Dr. Julia Mittermayr, EVP, Growth, Rep Data, Austria.

# Human + Machine

## Rebuilding Representation in AI-Driven Research

*Mario X. Carrasco*

Esomar

Office address:

Burgemeester Stramanweg 105

1101 AA, Amsterdam

Phone: +31 20 664 21 41

Email: info@Esomar.org

Website: www.Esomar.org

Publication Date: September 2026

Esomar Publication Series Volume S424 Congress 2026

ISBN 92-831-0337-8

Copyright © **Esomar** 2026

Copyright © **Esomar** 2026

Congress 2026

# Human + Machine

## Rebuilding Representation in AI-Driven Research

*Mario X. Carrasco*

## Introduction

This study presents a critical evolution for the market research industry by addressing the statistical invisibility (gap) that has historically distorted strategies in the region. Through a massive analysis of 15,077 respondents in Mexico, we demonstrate how semantic intelligence transcends rigid demographic categories to capture the “truth of the subject" and their latent tensions, rescuing from anonymity sectors that inhabit the periphery of averages. The results reveal a detailed vulnerability map, identifying high-risk clusters such as the “sandwich" clusters, generation strivers who face critical levels of burnout (66.7/100) and systemic attrition that traditional statistics often ignore. By implementing a local-first architecture that guarantees data sovereignty and privacy, this methodology not only humanises mathematical data, but also provides organisations with an early warning system capable of anticipating crises and the needs of the population before they manifest.

## Oversights of Mental Health in Traditional Segmentation

In the fields of public health and corporate wellness, traditional segmentation models are failing to identify latent danger. Based on a large-scale study of 15,077 respondents in Mexico, we have observed that current tools fail to detect at-risk populations for three fundamental technical and sociological reasons.

### *The 2D Trap*

Traditional epidemiological and social segmentation is limited to three or four axes (age, gender, socioeconomic level and previous diagnosis).

- The problem: By forcing the Mexican population into only four "boxes", 90% of the complexity of human behaviour is ignored.
- Mental health consequences: “Average" profiles are created that obscure individual crises. It's like trying to diagnose a complex pathology by only observing the patient's shadow: the depth of the anguish, isolation and exhaustion is lost.

### *The “Average Bias" (The Flaw Of Averages)*

Conventional statistics are obsessed with the centre of the Gaussian bell curve, searching for the “average Mexican".

- The problem: Psychosocial risk is not a uniform mass; it is a “long tail" of vulnerability niches.
- Consequence: When designing health programmes for the “average user", institutions ignore the extremes of the curve, where cases of imminent danger and silent crises reside. The traditional mathematical model erases the most vulnerable to simplify the bar graph.

### *Silo Fragmentation (Variable Isolation)*

Traditionally, demographics and symptomatology do not communicate with each other.

- The problem: It is assumed that all people of the same age or income range face the same mental challenges.
- Consequence: Semantic tension is not captured. For example, according to our sample of 15,077 records, a high-income executive in Mexico City who suffers from insomnia and lives alone has a semantic risk trajectory closer to that of an unemployed young person in isolation, than to that of their neighbour with a family and work-life balance. For traditional segmentation, they are opposites; for semantic AI, they are neighbours at risk.

Copyright © **Esomar** 2026


## Semantic Intelligence as a Driver of Statistical Inclusion

Statistical invisibility in Latin America has historically been a barrier that distorts the design of welfare policies and public health strategies. This study argues that the “representation gap" is not solely due to a lack of data volume, but rather to a lack of analytical resolution in traditional tools. By applying semantic artificial intelligence to a sample of 15,077 respondents in Mexico, we were able to transcend the rigid demographic categories that tend to omit the latent tensions and cultural and social nuances inherent to the Mexican population. The use of latent spaces allows the rich narratives of each individual to emerge as valuable dimensions, rescuing from statistical anonymity those sectors inhabiting the periphery of averages and revealing a cartography of psychographic vulnerability that linear models systematically omit. The construction of this analytical sample acts as a technological bridge toward a more equitable and humanistic representation in the region. By implementing synthetic anchors of resilience and risk the model not only processes information, but also reconstructs the sociopsychological fabric of the 15,077 participants, allowing mental health patterns to transcend the stigma of direct reporting and manifest themselves through their behavioral infrastructure. This approach reduces the gap between lived reality and official data, transforming statistical “silence" into a metric signal for action. Thus, the study positions AI as a tool for information sovereignty in Mexico (in this case), capable of giving a mathematical voice to the deepest and least represented layers of society, ensuring that the design of well-being solutions is based on the essence of the respondent and not on reductionist labels.

## Methodology: Semantic Intelligence Architecture for the Representation of Mental Health

Before breaking down the technical steps, it's important to establish the foundation upon which this architecture is built: the semantic approach. Data is usually treated as isolated, literal units; if a respondent doesn't check the “mental disorders" box, the system assumes their absence. Computational semantics, on the other hand, doesn't look for keywords, but rather for structures of meaning.

### *The Semantic Relationship as a Mirror of Mexican Reality*

Words alone are ambiguous. In our context, the term “stress" can be normalised or minimised due to social stigma. However, by linguistically combining variables such as [HH_SIZE:1 (living alone), TV:0h (lack of leisure time) and JOB:High Responsibility, the semantic relationship eliminates doubt: we are facing a signature of latent vulnerability. Human beings are not linearly consistent. A young person in Mexico may have a low income, but a very high level of digital sophistication and an undiagnosed burden of anxiety. This technology allows that “tension" between variables to be expressed as a unique coordinate on the map. It doesn't force the profile into a predefined box; it allows the linguistic relationship of their habits to dictate their position. In this study, the semantic relationship is what allowed us to identify the hidden danger. We did not calculate the risk by seeking direct user reports, but rather by analysing the correlation between:

- Low leisure and recreation time;
- Extreme work-related responsibilities;
- Housing insulation factors;
- Reported physical symptoms (fatigue, insomnia).

It is the union of these points that generates the health insight, not the points separately.

## Phases of the Topological-Semantic Process

### *Phase 1: Dense Narrative Engineering (Data Protocol)*

Instead of processing isolated columns, a structural normalisation system was implemented. Each record of the 15,077 subjects in Mexico was transformed into a dense narrative under the KEY:Value format.

- Why it matters: It eliminates grammatical noise and standardises the signal for AI, allowing populations historically underrepresented by their response style to be treated with the same technical precision;

- Analogy: It's about moving from individual pieces to a digital architectural plan. AI doesn't guess, the structure tells it the function of each biographical and behavioral data point.

### *Phase 2: High Resolution Vectorisation (Embedding)*

Transformer architecture Text Embedding model (Vaswani et al., 2017: 14) was used to project the narratives into a 768-dimensional latent space.

- The process: Convert each profile into a unique mathematical vector;
- The advantage: The "distance" between two Mexicans on this map is not measured in kilometers or pesos, but in affinity of vulnerability or resilience;
- Rigour and privacy: By processing the study locally, we ensure that sensitive health information remains protected, reducing gaps of mistrust in data handling in the region.

### *Phase 3: Psychographic Decomposition (Synthetic Anchors of Risk and Resilience)*

To “interrogate" the latent space, we designed synthetic dimensions that act as health probes. Unlike predefined labels, these anchors represent archetypes of latent danger (psychopathological vulnerability, burnout, caregiver burnout syndrome) and protective factors (resilience and stability). The result is that each individual receives a score from 0 to 100 on critical dimensions. This allows for a continuous gradient analysis: we don't classify someone as “sick" or "healthy", but rather measure how close they are to a crisis before it occurs.

### *Phase 4: Topological Clustering and Discovery of High-Risk Audiences*

Clustering algorithm was applied to identify the natural communities that emerge from the data density. We discovered that the resulting segments are not imposed by the researcher; they are emergent phenomena. This allows for the discovery of “latent audiences": i.e., groups of people in Mexico with common mental health needs that traditional statistics erase because they do not fit standard demographic averages or because they lack a prior diagnosis.

## Results

This section provides a detailed qualitative breakdown of the primary archetypes identified within the 15,077 respondents in Mexico. By integrating quantitative risk scores with semantic narratives, we move beyond mere statistics to understand the lived experiences and "latent" dangers facing these specific population segments.

| N | Cluster name | Segment size | Clinical vulnerability | Burnout index | Caregiver burden | Resilience index |
|---|---|---|---|---|---|---|
| 1 | Tech-savvy family men | 5,675 | 65.1 | 65.8 | 68.8 | 68.4 |
| 2 | Urban tech workers with health challenges | 6,099 | 67.1 | 70.1 | 71.9 | 71.1 |

| | | | | | | |
|---|---|---|---|---|---|---|
| **3** | Sandwich generation strivers | 3,303 | 64.8 | 66.7 | 69.8 | 70.3 |

*Table 1. Values obtained from the measurement between dense narratives and synthetic anchors (see Methodology section).*

To compare the semantic approach with a traditional method, the same 15,077 respondents were clustered using K-means directly on structured variables after standard scaling and one-hot encoding. The semantic method used the same data transformed into dense narratives and embeddings, also with K-means and three clusters. Both methods were evaluated with the same four synthetic anchors: clinical vulnerability, burnout index, caregiver burden and resilience index.

| Method | Cluster | Segment Size | Clinical vulnerability | Burnout index | Caregiver burden | Resilience index |
|---|---|---|---|---|---|---|
| **Semantic clustering** | Tech-savvy family men | 5,675 | 65.1 | 65.8 | 68.8 | 68.4 |
| **Semantic clustering** | Urban tech workers with health challenges | 6,099 | 67.1 | 70.1 | 71.9 | 71.1 |
| **Semantic clustering** | Sandwich generation strivers | 3,303 | 64.8 | 66.7 | 69.8 | 70.3 |
| **Traditional K-means** | Traditional cluster 1 | 4,278 | 65.49 | 66.63 | 69.61 | 69.44 |
| **Traditional K-means** | Traditional cluster 2 | 5,582 | 65.84 | 68.57 | 70.66 | 70.37 |
| **Traditional K-means** | Traditional cluster 3 | 5,217 | 66.22 | 67.88 | 70.58 | 69.94 |

*Table 2*

The semantic approach produced consistently greater separation between clusters across all four dimensions, indicating a stronger ability to distinguish latent population profiles than traditional K-means applied directly to structured variables.

| Index | Semantic clustering | Traditional K-means | Greater semantic separation |
|---|---|---|---|
| **Clinical vulnerability** | 2.3 | 0.73 | 3.15× |
| **Burnout index** | 4.3 | 1.94 | 2.22× |

Copyright © **Esomar** 2026


| | | | |
|---|---|---|---|
| **Caregiver burden** | 3.1 | 1.05 | 2.95× |
| **Resilience index** | 2.7 | 0.93 | 2.90× |

*Table 3*

### *Tech-Savvy Family Men*

- Psychological profile: This segment is characterised by a high degree of professional engagement that borders on the pathological. Despite having a high resilience index, which suggests a robust set of coping mechanisms, their clinical risk and burnout scores are alarmingly elevated. They represent a “stretched" population: individuals who are bouncing back from stress, but doing so at a significant physical and mental cost.

### *Key Latent Signals*

- Risk factors: The intersection of long working hours and frequent business travel creates a “fatigue trap". This is compounded by a high prevalence of co-occurring conditions (co-morbidity), where chronic physical ailments (asthma, diabetes) exacerbate mental health strain.
- Protective factors: A diverse age range (from 18 to over 65) provides a rich “experience buffer", while larger household sizes offer a potential social safety net that prevents total isolation.

### *Urban Tech Workers with Health Challenges*

- Psychological profile: This segment appears to be at moderate to high risk for mental health issues, given the high clinical risk and burnout index scores. While their resilience index is relatively high, it may not be sufficient to counteract the stressors they face. They seem to be a group that is highly engaged in work and may struggle with work-life balance, leading to increased stress and burnout.

### *Key Latent Signals*

- Risk factors: High clinical risk and burnout index. Many individuals in this segment report a wide range of health issues, including mental health conditions, which can exacerbate stress and burnout.
- Protective factors: To support this segment, interventions could focus on promoting work-life balance, providing mental health resources and support and addressing the physical health issues that may be exacerbating mental health concerns.

### *Sandwich Generation Strivers*

- Psychological profile: This cluster exhibits a critical state of systemic attrition. With a clinical risk score of 64.87 and a burnout index of 66.71, these individuals are struggling under the “dual burden" of professional ambition and intensive caregiving. They are the “sandwich generation": Mexican professionals balancing the needs of elderly parents and young children while maintaining high-demand roles in sales and executive leadership.

### *Key Latent Signals*

- Risk factors: Financial strain is a primary driver here, despite their high professional status. The cost of multi-generational caregiving leads to “high-income poverty" where expenses outpace earnings. Their health profile is the most compromised, showing high rates of cardiovascular and sleep disorders

- Protective factors: High educational attainment and strong family bonds act as the final anchors of stability. Their access to employer-provided technology and vehicles offers a sense of mobility and “professional self-worth" that mitigates deeper depressive states.

Technical integrity and data sovereignty were fundamental pillars in the execution of this study. The entire processing flow, from the normalisation of dense narratives and the generation of embeddings to topological clustering, was implemented entirely using Python scripts. Once the natural communities in the latent space were identified, the qualitative interpretation of each cluster was performed by querying a large language model (LLM) executed in a local environment. This local-first architecture ensured that the in-depth analysis was carried out without exposing the population's mental health to third-party platforms.

## Validation

The validation of this study lies not only in the sophistication of its technical architecture, but in the convergence of dimensions: when mathematical cohesion, the sociodemographic reality of Mexico and clinical judgment coincide, the model transcends theory to become an instrument of social precision. Within the framework of reducing representation gaps in Latin America, we have validated that each finding extracted from the 15,077 respondents is not a statistical artifact, but a window towards a real, measurable and, above all, humanly actionable welfare intervention.

### *Consistency Testing by Experts (Human-in-the- Loop)*

To ensure that AI does not lose its sense of human reality in pursuit of mathematical abstractions, we submitted a random sample of profiles from each cluster to expert judgment in clinical psychology, sociology and public health.

- The process: The expert analyses an individual's dense narrative (habits, environment, reported symptoms) and, without knowing the cluster assigned by AI, determines a risk or resilience profile.
- The result: We measured the degree of agreement between human diagnosis and algorithmic grouping. A high level of agreement certifies that the identified risk audiences possess clinical and social coherence, ensuring that the model effectively detects the “latent danger" that the traditional healthcare system often overlooks.

### *Semantic-demographic cross-validation*

Although our model operates “blind" to traditional labels during topological clustering, we performed a post-hoc validation by cross-referencing the clusters with known demographic and health variables of the 15,077 participants in Mexico.

- The test: If the model identifies a cluster under the anchor of “critical burnout risk", and upon reviewing the raw data we observe that 90% of those individuals report long working hours, insomnia and high responsibility positions, the model is externally validated.
- The value: This demonstrates that AI has uncovered Mexican sociological reality without prior instructions. These findings provide evidence that latent-space representations can capture meaningful patterns of vulnerability that may be less visible through conventional segmentation.

### *Construct Stability (Robustness)*

To ensure that the model is stable and representative of the diversity in Mexico, we performed resampling tests (bootstrapping) on the total study population.

- The process: We trained the model repeatedly using different subsets of the 15,077 records to check if the centres of gravity of the risk clusters remain constant.
- Why it matters: This test ensures that the results are not the product of anomalies in a small group of people, but rather reflect structural and enduring trends in mental health across the region. This safeguards the research against bias, and ensures that public policy recommendations are based on a robust and reliable database.

## Use Cases

The ability to map the “latent psychographics" of a population of 15,077 Mexicans on a massive scale opens doors that traditional health and wellness segmentation keeps closed. By identifying unreported risk patterns, we transform data analytics into a tool for preventative intervention. Below, we detail the most disruptive applications of this model in the context of reducing representation gaps.

### *Surgical Targeting of Social-Health Interventions*

Instead of launching generic prevention campaigns based on “age 25 to 34", institutions can direct resources towards “moods and risk trajectories".

- The application: To identify the spaces (physical or digital) where individuals with high clinical vulnerability coexist but are not being served by the formal system. For example, to detect groups at high risk of burnout in specific industrial sectors who do not consume health content, but are highly active in technology forums.
- The impact: Optimisation of public and private spending by avoiding the dispersion of resources. Vulnerable populations are reached through unconventional channels where there is less competition for attention and greater receptiveness to the message of support.

### *Detection of “White Spaces" in Mental Health (Service Innovation)*

Analysis of emerging clusters in the 15,077 records reveals welfare needs that no current health program in Mexico addresses.

- The application: The model can identify a group of individuals with high life stage frailty (such as lonely older adults) who exhibit a narrative of job burnout, but possess high technological skills. This is a clear indication for developing human digital companionship services or “emotional concierge" platforms that do not currently exist in the market.
- The impact: Reduction of risk in the launch of social programmes or welfare services by being based on data-driven evidence of latent demand rather than solely on prior assumptions.

### *Early Crisis Warning (Evolution of Semantic Churn)*

Changes in semantic trajectories may potentially provide early signals associated with increasing vulnerability, although longitudinal validation is required.

- The application: By comparing the evolution of profiles over time, organisations can observe if a user is moving away from the “resilience and stability" clusters and shifting towards those of “social isolation" or “psychopathological risk", even before the subject is aware of their own deterioration.
- The impact: Implementing proactive intervention strategies. It is much more effective and humane to offer preventative support when semantic shift is detected than to try to treat a clinical crisis that has already occurred.

### *Sovereignty Over Sensitive Data and Ethics (Privacy-First)*

Given that mental health treatment in Mexico involves extremely sensitive data, processing using local AI is a pillar of information sovereignty.

- The application: Institutions can perform in-depth analyses on the 15,077 vulnerability records without a single piece of health data leaving their infrastructure. There is no exposure to public clouds or transfer to third countries.
- The impact: Complete legal protection under Mexico's Federal Law on the Protection of Personal Data and increased public trust. Knowing that their “digital self" is protected locally reduces the stigma barrier, allowing for more honest and complete representation.

## Limitations, Risks and Learnings from the Synthetic Approach

Every cutting-edge methodology has limitations that must be understood for proper strategic interpretation. In this large-scale study, we have identified the following control points.

### *Model Training Bias (Inherent Bias)*

Embedding models have been pre-trained on vast corpora of global data (internet, literature, articles).

- The risk: The model could inherit cultural or linguistic biases present in its original training;
- Mitigation: Human supervision (psychology and marketing experts) was carried out to validate that the semantic interpretations in the latent space were consistent with the local context and the company's objectives.

### *The Dependence on "Quality of Input" (Garbage In, Garbage Out)*

The accuracy of latent space depends directly on the richness of the initial dense narrative.

- The limitation: If the originally collected data is extremely poor or ambiguous, the embedding model generates "low-resolution" vectors. A respondent with minimal information will tend to be positioned toward the center of gravity of the model, losing their niche identity.
- Impact: This can generate clusters of "generic profiles" that require deeper data cleaning to be actionable.

The synthetic anchors are researcher-defined semantic representations rather than clinically validated diagnostic instruments. Their results may therefore be sensitive to the wording, concepts and assumptions used when constructing each anchor. The resulting scores should be interpreted as relative indicators of semantic proximity to a construct, not as clinical diagnoses or direct measurements of mental health conditions. A key learning from the experiment was the importance of carefully designing and reviewing the anchors to ensure that they represent the intended constructs without excessive overlap. Validation through expert review, demographic cross-checking, robustness testing and comparison with traditional clustering provides supporting evidence, but further validation across independent populations, alternative embedding models, and longitudinal data would be necessary to establish broader generalisability.

### *Correlation Versus Causality*

This is the most important limitation for decision-making in any area of market research.

- The risk: The model is exceptionally good at finding affinity and correlation (e.g., "People who value luxury also report a lack of time"). However, the model cannot, on its own, claim that one variable causes the other.
- Recommendation: The findings should be used as working hypotheses for campaigns and experiments, not as absolute causal truths.

### *Temporary Validity (Semantic Erosion)*

The language and concerns of the population are dynamic and fluctuating.

- The risk: A semantic-based audience model can become obsolete if a disruptive event occurs in the market or culture that changes the meaning of certain concepts for the population.
- Suggested solution: Periodic re-validation of the synthetic anchors is recommended to ensure they continue to capture the reality of the moment.

## Implementation Recommendations

To prevent a study of this magnitude from becoming a "dead file", the company must move from diagnosis to execution. The implementation recommendations serve as a roadmap for integrating this intelligence into daily operations.

### *Real Time Scoring Workflow Automation*

To maintain the relevance of the study, the company must implement an automatic classification pipeline that processes each new record (respondent) in real time.

- The action: Use the local-first architecture developed to instantly assign each new respondent a place in the latent space and a cluster of belonging.
- The benefit: Segmentation ceases to be an annual event and becomes a dynamic process. Companies can react to market changes in days, not months.

### *Establishment of an "Anchor Laboratory"*

Given that the market is dynamic, synthetic anchors must evolve.

- The action: Create a bi-monthly committee composed of experts in psychology, health and data science to "launch new probes". If a trend emerges (e.g., interest in generative AI), a synthetic anchor is created to detect who in the total population is gravitating toward that concept.
- The benefit: It enables evidence-based innovation, detecting opportunities for new products or services before the competition does.

## Conclusions: Towards A New Era of Human Representation

This study, based on 15,077 records in Mexico, demonstrates that population analysis has entered an era of high fidelity. We have successfully transitioned from segmentation based on static snapshots and rigid demographic labels to navigation using semantic, dynamic maps. The implementation of a local-first AI architecture and the strategic use of latent spaces not only resolve the historical problem of invisible or underrepresented audiences, but also completely redefines the relationship between institutions and citizens. By treating each data point as a piece of an ecosystem of meaning and not as an isolated variable, this technology has enabled:

- Making the invisible visible: We managed to identify risk niches and latent human needs that traditional statistics systematically erase by omission or by not conforming to standard averages;
- Humanising data: We have been able to translate abstract mathematical vectors into actionable and measurable psychographic dimensions, such as psychopathological vulnerability, burnout risk and caregiver burnout syndrome;
- Ensuring agility and sovereignty: The model is not only dynamic and capable of evolving at the same speed as human behaviour, but it also ensures the total privacy of sensitive mental health information by being processed locally.

In conclusion, any organisation or public entity that adopts this semantic intelligence will not only better understand the psychosocial reality of its population today; it will have built the necessary ethical and technological infrastructure to anticipate the future and detect crises before they manifest. This approach offers a potential pathway toward reducing the gap between fragmented data and more comprehensive well-being strategies.

## References

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., and Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems, Vol. 30*.

## About the Author

Mario Xavier Carrasco, co-founder, ThinkNow, Burbank, CA., US.

# The Strategic Growth Gap

## How Insight is Rewiring Growth Decisions

*Lucas Hulsebos*

Esomar

Office address:

Burgemeester Stramanweg 105

1101 AA, Amsterdam

Phone: +31 20 664 21 41

Email: info@Esomar.org

Website: www.Esomar.org

Publication Date: September 2026

Esomar Publication Series Volume S424 Congress 2026

ISBN 92-831-0337-8

Copyright © **Esomar** 2026

Copyright © **Esomar** 2026

Congress 2026

# The Strategic Growth Gap

## How Insight is Rewiring Growth Decisions

*Lucas Hulsebos*

### Introduction: The Growth Paradox

Every brand wants to grow, which is why growth has become one of the clearest signals of business success. Growth is the outcome boards demand and investors ask for. Most leadership teams are held accountable for growth and rewarded for it. In annual reports of leading companies, growth is repeatedly presented as a key indicator of success. This focus on growth is not new. Research in strategy and corporate finance has consistently shown that sustained revenue growth is a primary driver of long-term enterprise value. Studies published by McKinsey and Harvard Business Review show that over time, companies that grow outperform those that rely mainly on efficiency or cost control. Cost control can protect margins, efficiency can improve profitability, but growth creates momentum: it enables reinvestment, attracts talent and increases strategic freedom. Companies that grow consistently are better able to absorb shocks, adapt to change and invest ahead of demand. It is the outcome that often determines whether an organisation is seen as healthy, relevant and capable of shaping its own future.

Yet, despite this central role, sustainable growth remains remarkably rare. Over long periods of time, only a very small minority of companies manage to outperform their markets consistently, while most experience phases of success followed by stagnation or decline. An analysis done by Matts Georgson showed that from the top 100 brands (listed by Interbrands in 2014), most companies could not grow. The percentage of companies that has grown on average 10% in the 10 years after was less than 10%. Additionally, the analyses did even go further, as shown in Figure 1. From all the companies that are publicly listed, only 1% is able to grow on average 10% in a period of 10 years. The growth problem, or paradox, is therefore not whether growth matters. It is why it remains so difficult to sustain.

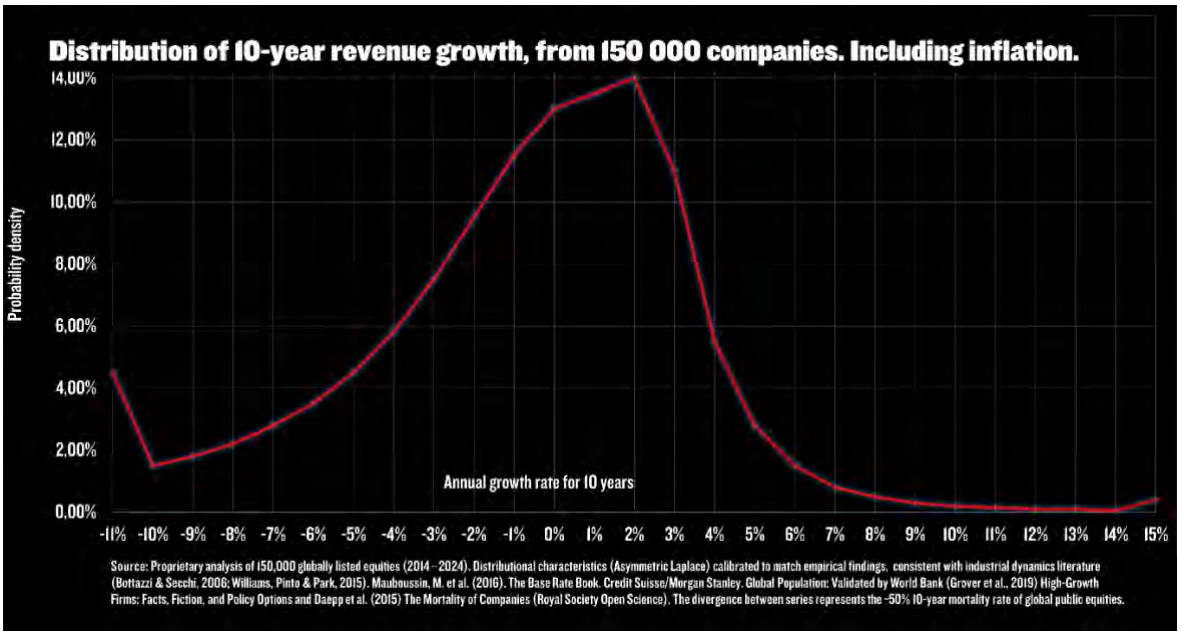



*Figure 1. Growth distribution of 150,000 publicly listed companies (source: Matts Georgson).*

The paradox becomes even more striking when viewed against the backdrop of modern marketing practices. Organisations have so much access to data, analytics, consumer insight and marketing science. Artificial intelligence now supports content creation, media optimisation and decision

support. Marketing teams can measure more, model more and automate more than at any point in the past. At the same time, the literature on brand growth has become more mature. Decades of research have clarified the importance of penetration, mental availability, physical availability, creativity, long-term brand building and disciplined execution. In theory, then, the conditions for growth should be improving whereas, in practice, they are not. The gap between what we know about growth and what organisations can do with that knowledge remains stubbornly wide.

## What Is Growth?

Ask 10 organisations how they define successful marketing and you are likely to hear 10 different answers. Some refer to brand awareness or consideration, measured via brand tracking or brand health checks. Others mention loyalty, engagement or net promoter score (NPS). Increasingly, efficiency metrics such as ROI or cost per acquisition are used. At board level, profit often becomes the ultimate reference point. While all of these metrics matter, none of them on their own defines growth. To understand why, it helps to be precise about what growth is. Growth is not an improvement in perception, efficiency or margin in isolation. Growth is a brand's ability to generate increasing demand over time that shows up when more people choose the brand, choose it more often or choose it at a higher value. Brand metrics such as awareness, consideration, brand health or loyalty are important for every brand. This is why brand funnels remain a core part of marketing thinking. They describe how a brand is perceived and remembered. They help diagnose strengths and weaknesses and support creative and communication decisions, but they are not growth outcomes. A brand can be widely known and still shrink. It can score high on consideration and still fail to convert that mental availability into sustained demand. It can enjoy strong loyalty among a small base and yet stagnate because it does not attract enough new buyers. This distinction matters: brand KPIs describe conditions for growth, not growth itself.

If brand KPIs are too indirect, profit may seem like an obvious alternative. After all, growth without profitability is unsustainable. Yet, profit is a problematic primary KPI for evaluating growth, especially when marketing is involved. Profit is influenced by many forces beyond demand creation: cost structure, pricing strategy, inflation, operational efficiency and accounting choices. A business can increase profit by cutting marketing investment, delaying innovation or raising prices. Those actions can weaken future demand, which is why profit on its own is not a reliable KPI for growth. DVJ's brand growth research repeatedly shows that short-term financial optimisation often coincides with reduced investment in brand building, innovation and insight. Promotions, cost control and efficiency gains can boost profit temporarily, but they rarely build lasting demand.

*Figure 2. Every KPI is important and represents something else.*

Based upon hundreds of vision interviews with CMO's and marketing directors, we found that most professionals agree that revenue is the best way to describe growth. Revenue growth is clear and straightforward. It reflects real behaviour in the market, not intention, opinion or internal alignment. It cannot be achieved by marketing alone, nor easily explained away. When revenue grows, something meaningful happens between a brand and the people it serves. This is precisely why revenue works as an anchor KPI for growth. That does not mean the other KPI's are not important; they are in understanding revenue growth. This distinction aligns with decades of empirical work on brand growth. Research by Byron Sharp and the Ehrenberg-Bass Institute shows that brands primarily grow by increasing penetration: i.e., being bought by more people, rather

than relying mainly on loyalty among a small group of existing buyers. Revenue growth is the outcome that shows whether these dynamics are working in the market.

Revenue growth is widely pursued, widely discussed and widely measured, yet consistently achieved by only a minority. That tension became the starting point for the brand growth platform. The original question was simple: why do some organisations continue to grow while others, operating under broadly similar conditions, do not? Revenue growth became the anchor KPI throughout the brand growth platform. It allowed us to move beyond perceptions, attitudes and intentions and focus on the outcome that matters most: whether organisations succeed in creating increasing demand over time.

## The Brand Growth Platform

The brand growth platform was established in 2016 with a relatively simple ambition: to better understand what drives sustainable brand growth. At the time, discussions around growth were becoming increasingly influenced by evidence-based marketing. The publication of *How Brands Grow* by Byron Sharp had challenged many long-held assumptions about loyalty, differentiation and brand building, while the work of Les Binet and Peter Field had renewed attention for the balance between short-term activation and long-term brand building. Across both academia and practice, interest in growth was increasing. At the same time, many organisations continued to struggle with translating these insights into consistent business results. The question was no longer whether growth mattered. The question was how organisations could systematically learn from the growing body of knowledge surrounding it. From the outset, the platform was not designed as a single study or as an attempt to validate one specific theory of growth. Growth is too complex to be explained by a single framework. Instead, the objective was to create a long-term learning system that would continuously bring together different sources of knowledge, and allow patterns to emerge over time.

The platform was built around three complementary pillars, namely:

1. Academic perspectives: Throughout the programme, academic insights have been used to frame questions, challenge assumptions and interpret findings. Rather than treating academic theories as competing schools of thought, the platform uses them as reference points that help connect individual observations to broader explanations. Organisations benefit when scientific knowledge and practical experience are allowed to inform one another rather than exist in separate worlds.
2. Vision interviews: Over the past decade, more than 400 interviews have been conducted with CMOs, marketing directors, insight leaders, media specialists and business leaders across different sectors and countries. These conversations are intentionally exploratory in nature. They help identify emerging themes, organisational tensions and practical realities that may not immediately surface in quantitative research. The interviews provide an opportunity to understand how growth decisions are actually made inside organisations.
3. Annual *Brand Growth* study: A yearly quantitative study conducted among marketing professionals across the world. What started as research in a limited number of markets has evolved into one of the largest longitudinal studies among marketing professionals. In total, more than 20,000 marketers have participated in the programme.

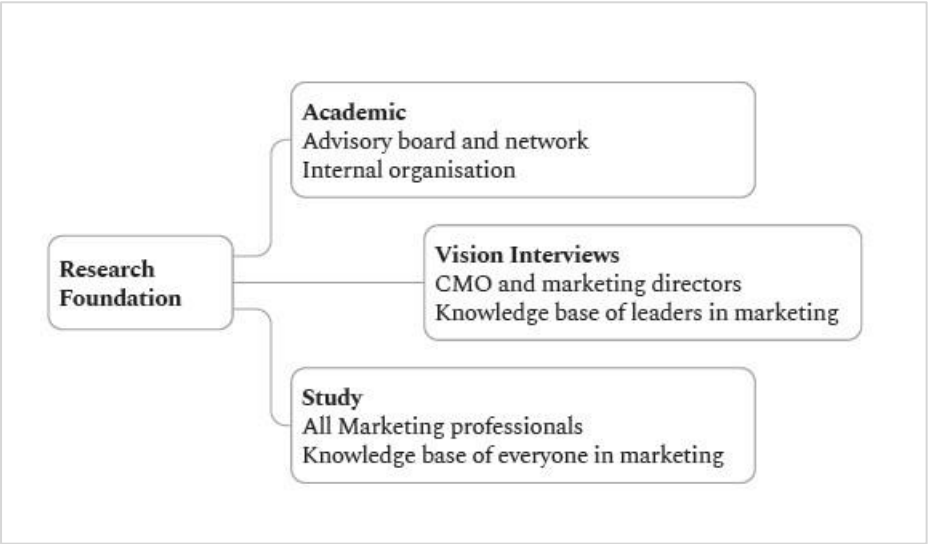



*Figure 3. The 3 pillars of brand growth research.*

One of the defining characteristics of the brand growth platform is its thematic structure. Each year, a different growth ingredient was selected for deeper investigation.

*Figure 4. The brand growth platform with ingredients.*

The first studies focused on brand strategy and the role of branding within organisations. Subsequent years examined media, innovation, creativity, data-driven decision-making, moments of truth, corporate social responsibility and promotions. More recently, the programme has also revisited several themes to understand how they are evolving in response to developments such as artificial intelligence, sustainability pressures and increasing market complexity. This approach created a unique opportunity. By studying different growth ingredients separately while maintaining a consistent core methodology, it became possible to compare findings across years and identify recurring patterns.

Initially, the expectation was that these studies would gradually reveal which ingredients contribute most strongly to growth and, in a sense, they did. Innovation repeatedly emerged as a critical driver. Creativity consistently appeared near the top of the rankings. Media, data-driven decision-making and customer understanding all proved important. Yet, after several years another observation started to stand out. The most interesting differences were not occurring within individual ingredients. They were appearing between organisations themselves. The same

organisations that scored highly on innovation often also displayed stronger marketing consistency. The same organisations that invested more in research also tended to be more data-driven, more customer-focused and more committed to long-term brand building. The differences appeared less like isolated capabilities, and more like manifestations of a broader organisational pattern. The discovery of that pattern ultimately led to what became the central concept of this paper: the strategic growth gap. The next section explores how this concept emerged from 10 years of accumulated evidence, and why it provides a different way of understanding both growth and the evolving role of insights within organisations.

## Discovering the Strategic Growth Gap

One of the most surprising findings to emerge from 10 years of brand growth research is that winners and losers largely agree on what drives growth. Throughout the programme, marketing professionals consistently identified innovation, creativity, data-driven decision-making, customer understanding and effective media investment as important contributors to business success. Even as new technologies emerged and marketing practices evolved, there was relatively little disagreement about the fundamental ingredients required for growth.

**DRIVERS TO BRAND GROWTH GLOBAL**

| | 2022 | 2026 |
|---|---|---|
| Innovation | 66% | 67% |
| Creativity | 63% | 66% |
| Media | 59% | 62% |
| Utilising Data | 57% | 61% |

*Figure 5. The drivers for brand growth.*

In many ways, this should not be surprising. The past decade has seen an unprecedented democratisation of marketing knowledge. Books, conferences, podcasts, training programmes and online content have made evidence-based marketing more accessible than ever before. Most organisations today are aware of the principles associated with sustainable growth. The challenge appears not to be knowing what drives growth, but consistently acting upon that knowledge. The reasons for not acting were analysed across two angles:

1. Budget differences: Growing organisations simply have more resources. Larger budgets certainly create opportunities to invest. Budget differences were visible, but they did not fully explain the pattern. In many cases, organisations with comparable resources displayed very different behaviours. The difference therefore appeared to lie not only in resources, but in how resources were allocated and protected over time.
2. Organisational capability: Perhaps winners simply employed better marketers or had stronger leadership teams. While talent undoubtedly matters, this explanation also proved incomplete. Many organisations had access to highly capable people, strong agencies and sophisticated technologies. Yet outcomes continued to diverge.

The strategic growth gap refers to the structural difference between organisations that consistently build future demand and organisations that primarily optimise current performance. It is not a gap in knowledge, it is not a gap in access to technology, nor is it necessarily a gap in talent. Rather, it is a gap in the ability to translate growth principles into consistent decisions, investments and organisational behaviour over time. One of the clearest examples can be found in the relationship between short-term and long-term thinking, as explained in Figure 6. The organisations that consistently grow appear to be better at managing the tension between long-term value and short-term gains. They do not ignore short-term performance, but they are less likely to sacrifice future growth drivers in pursuit of immediate results.

Copyright © **Esomar** 2026



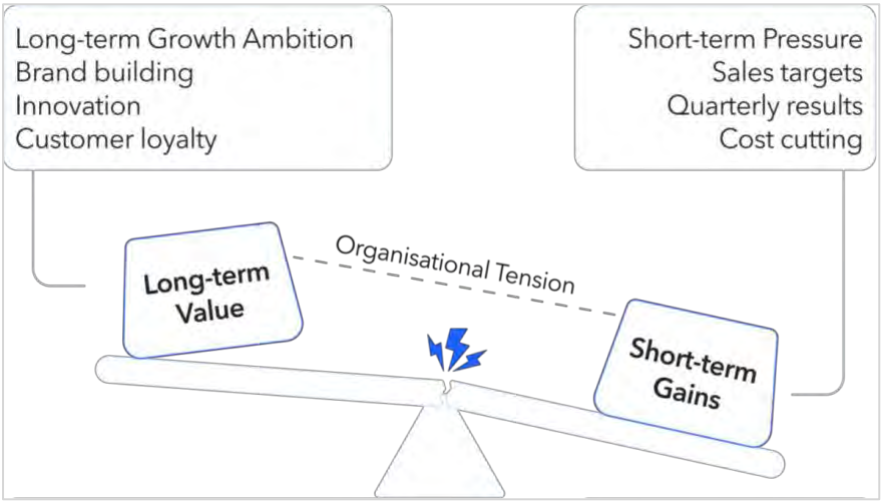



*Figure 6. The strategic growth gap explained.*

This is where the strategic growth gap becomes particularly useful as a framework. Traditional growth theories tend to focus on what organisations should do, whereas the strategic growth gap focuses on why some organisations can do those things consistently while others cannot. It provides an organisational explanation for a challenge that is often treated primarily as a marketing problem. This perspective also sets the stage for the next part of the analysis: if the strategic growth gap is real, then it should be observable across multiple growth ingredients. The next section of this paper explores exactly that question.

## Evidence for the Strategic Growth Gap

The strategic growth gap emerged gradually over the course of 10 years of research. The organisations that consistently reported growth appeared to behave differently across almost every growth ingredient we examined. What makes this finding particularly interesting is that winners and losers largely agree on what drives growth. Both groups recognise the importance of innovation, creativity, media and customer understanding. Yet, agreement on growth principles does not automatically translate into growth outcomes. The evidence suggests that the real difference lies in the ability to consistently apply these principles within the organisation.

### *Strategic Consistency as a Growth Capability*

One of the largest differences observed in the 2026 *Brand Growth Study* concerns the consistency of marketing policy. When respondents were asked how consistent they perceived the marketing policy within their organisation to be, the gap between winners and losers was striking. 61% of organisations reporting revenue growth described their marketing policy as consistent, compared with only 24% of organisations whose revenues were stable or declining. This finding may appear simple, but it is arguably one of the most important observations of the entire programme. Marketing discussions often focus on finding the next growth opportunity, the next innovation or the next breakthrough campaign. The evidence suggests that one of the strongest differentiators may simply be the ability to maintain a coherent strategic direction over time. Secondly what we see is that winners differentiate themselves significantly from stagnating brands in the role of marketing. If marketing is present in the c-suite of an organisation, these companies in general perform better. The implication is important. Growth appears to be less dependent on discovering new growth principles and more dependent on consistently applying existing ones and letting marketing play a key role to support that.

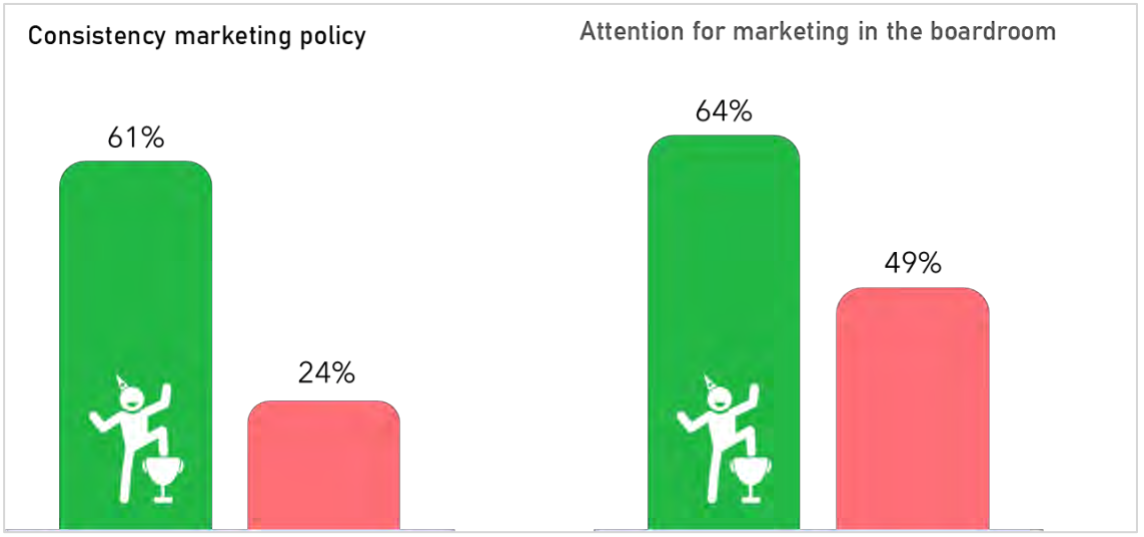



*Figure 7. Marketing consistency differences between winning and stagnating brands.*

### *Investment Behaviour: Building Future Demand*

The same pattern appears when examining investment decisions. Winning organisations consistently report higher levels of investment in marketing, innovation and research. While this may initially appear unsurprising, the magnitude and consistency of the differences are noteworthy. In the 2026 study, 66% of winners reported increasing marketing budgets compared with only 40% of losers. Similar differences were observed for innovation budgets and research budgets.

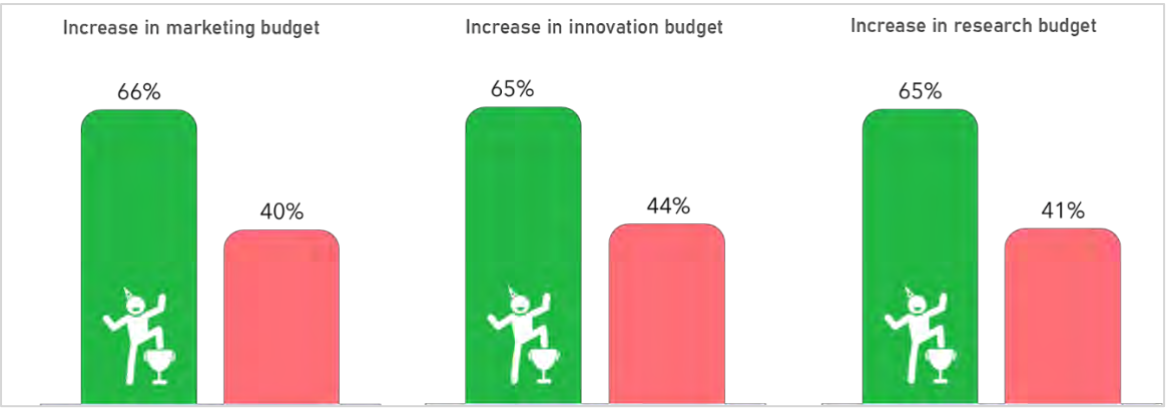



*Figure 8. Investment behaviour between winners and stagnating brands.*

Importantly, the distinction is not simply that winners spend more money. Rather, winners appear more willing to continue investing in future demand creation when market conditions become uncertain. Throughout the vision interviews conducted over the past decade, senior marketers repeatedly described the tension between short-term performance pressures and long-term growth investments. The evidence suggests that growing organisations are better able to protect investments that contribute to future growth. This observation aligns with a broader conclusion from the brand growth platform: sustainable growth rarely results from isolated interventions. It emerges when organisations maintain commitment to multiple growth capabilities simultaneously.

### *Innovation: From Activity to System*

Innovation consistently ranks among the most important growth drivers identified by marketers. In the 2026 study, 67% of respondents identified innovation as a key driver of growth, making it the highest-ranked growth ingredient. Yet, the distinction between winners and losers is not simply whether they innovate. It lies in how innovation is organised. Growing organisations are substantially more likely to involve customers in the innovation process. Nearly half of winners (48%) report strong customer involvement, compared with only 27% of losers. Winning organisations are also more likely to position marketing as a central driver throughout the innovation journey, connecting consumer understanding with commercial execution.

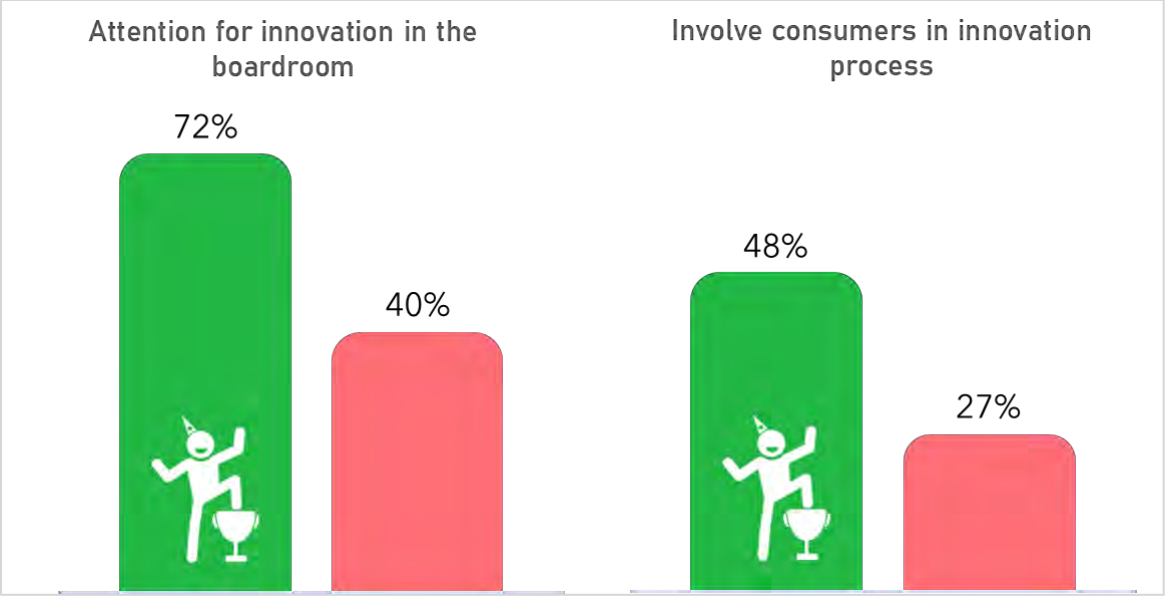



*Figure 9. Customer involvement in innovation between winning and stagnating brands.*

Interestingly, the biggest barriers to innovation identified by marketers are not a lack of ideas. The most frequently mentioned barriers include insufficient budgets, slow development processes and lack of organisational commitment. This suggests that innovation is not primarily a creativity challenge—it is an organisational challenge.

### *Media: Consistency versus Campaign Thinking*

The media landscape has become increasingly fragmented over the past decade. Organisations have access to more channels, more targeting options and more performance metrics than ever before. Yet despite this growing complexity, a relatively simple distinction appears between winners and losers. Winning organisations are significantly more likely to adopt an always-on media strategy. More than half of winners (56%) report applying an always-on approach, compared with only 26% of losers. Winners are also more likely to describe their media investments as focused on long-term objectives, rather than purely short-term outcomes.

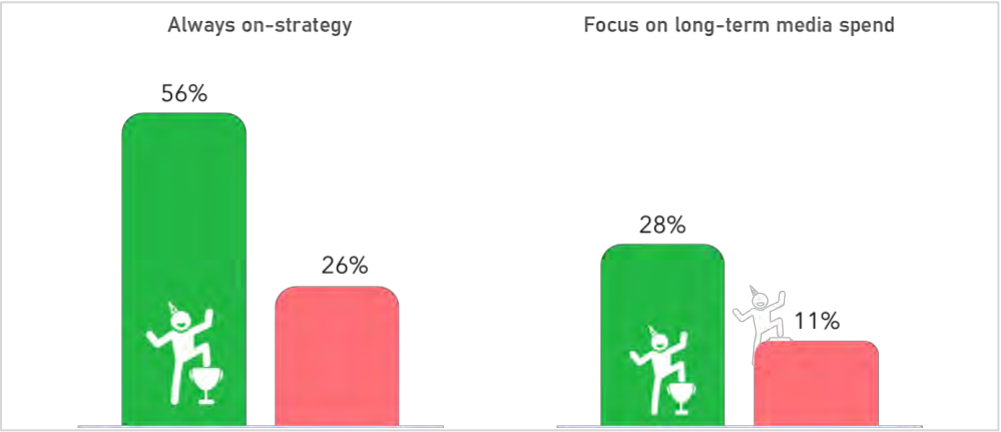



*Figure 10. Media behaviour between winning and stagnating brands.*

This finding resonates strongly with both marketing science and practitioner interviews. Growth organisations appear less inclined to treat media as a series of isolated campaigns, and more inclined to view it as a continuous process of demand creation.

### *Data-Driven Decision-Making*

Perhaps the most relevant findings for the insights industry emerge from data-driven decision-making. Over the past decade, organisations have invested heavily in data infrastructure, analytics and measurement capabilities. Then again, the brand growth research repeatedly shows that access to data is not the primary differentiator. The distinction lies in how data is used. Only 26% of losing organisations describe themselves as strongly data-driven, compared with 43% of winners. Furthermore, 61% of winners expect the importance of data-driven decision-making to increase over the coming years, compared with only 31% of losers.

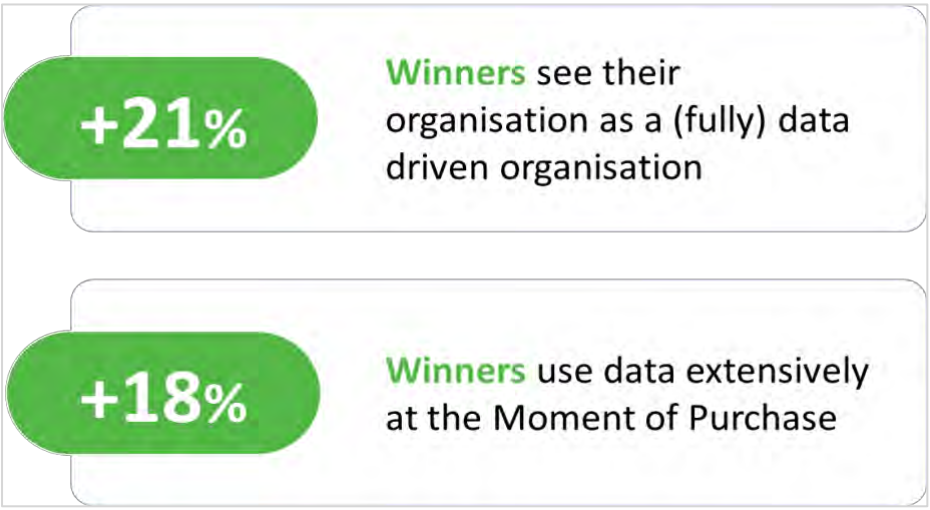



*Figure 11. The use of data between winners and stagnating brands.*

However, another finding may be even more revealing. When marketers were asked about challenges related to data-driven decision-making, the most frequently mentioned issues were not data access or technology. Instead, organisations struggled with prioritisation, translating data into decisions and connecting analysts with decision-makers. In other words, the challenge is increasingly organizational, rather than technical.

### *Moments of Truth: Focusing Resources Where They Matter*

The concept of *moments of truth* has been a recurring theme within the brand growth platform. While organisations often map increasingly complex customer journeys, growth organisations appear more focused on identifying the moments that matter most. The 2026 study shows that 59% of winners have explicitly defined one or more moment(s) of truth, compared with 43% of losers. Winners are also substantially more likely to use data at these critical moments. 44% describe themselves as fully data-driven at the moment of purchase, compared with only 27% of losers. These findings reinforce a recurring theme throughout the research: growing organisations are not necessarily doing more; they are often better at focusing resources on the activities and decisions that matter most.

*Figure 12. Difference between winners and stagnating brands on defining moment of truth.*

### *The Pattern Behind the Numbers*

Viewed individually, each of these findings provides an interesting insight into one specific growth ingredient. Viewed collectively, they reveal something much larger. The same organisations consistently outperform across multiple dimensions: they invest more consistently; they involve customers more actively; they are more data-driven; they maintain greater strategic consistency; they define critical moments of truth more explicitly; they continue investing in future demand creation when pressure increases. The strategic growth gap, therefore, does not appear to be the result of a single capability; it is the outcome of a system of behaviours that reinforce one another over time. Perhaps most importantly, many of these differences ultimately point towards the same underlying factor: the role of insight in organisational decision-making. If the strategic growth gap describes what differentiates winners from losers, the question becomes why these differences emerge in the first place. The answer, we argue, lies in a broader transformation taking place within organisations: the metamorphosis of insight from measurement function to strategic direction.

## The Changing Role of Insights

For decades, the value of research has largely been associated with the production of information. Better research generated better insights. Better insights were expected to lead to better decisions. The implicit assumption was that the primary contribution of insight professionals was helping organisations understand what had happened, why it happened and how performance should be evaluated. However, the findings of the brand growth platform suggest that the role of insight is gradually expanding beyond measurement alone. Marketers repeatedly identified data-driven decision-making as a key driver of growth. At the same time, they reported significant challenges in translating data into decisions. More than half of marketers believe organisations already have too much data to manage effectively, while many struggle with prioritisation, implementation and connecting analysts to decision-makers. The challenge is therefore not data availability. The challenge is organisational application, which is why insight professionals should create value by using data. For that a simple framework can be used.

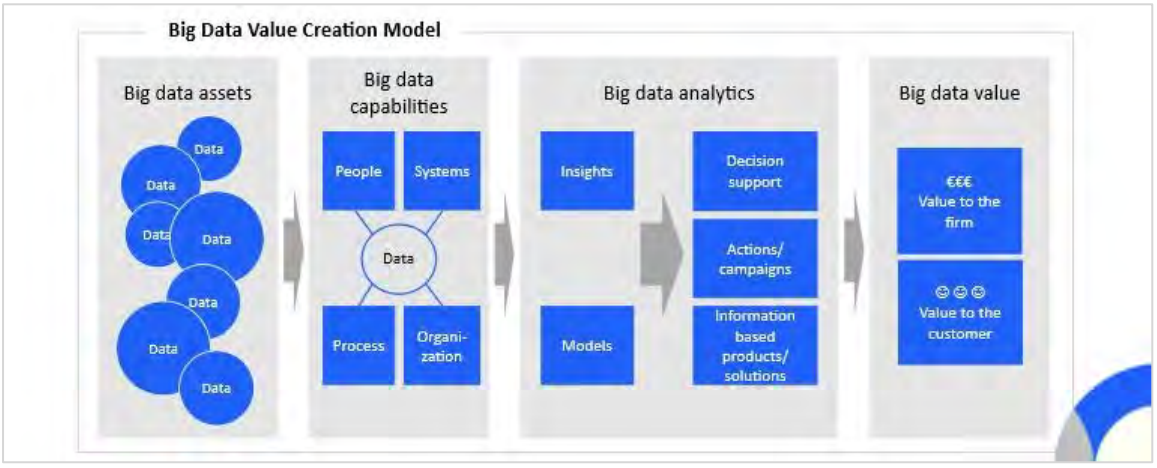



*Figure 13. Data value creation model.*

The findings suggest that volume does not differentiate growing brands from stagnating brands. What matters is whether data is accessible, trusted and connected to real decisions. When data cannot be accessed easily, is not trusted or sits outside decision cycles, it becomes a burden rather than an advantage. Thus, the first step is to be clear about the important KPI's and how datapoints are connected to decisions. The most critical breakdown is between insight and action. Many organisations generate insights that never influence a choice. Growing brands treat insights differently. Insights are incomplete until they have changed a decision. Data is not successful when it is accurate, but when it is used. The bottleneck has also gradually shifted. Increasingly, the challenge is deciding which information matters instead if how it should be collected. Organisations are no longer competing primarily on access to knowledge. They are competing on their ability to translate knowledge into action. Insights managers should therefore change their focus.

This observation becomes even more relevant when viewed through the lens of artificial intelligence. According to marketing professionals, AI emerged as the most frequently mentioned future development in media planning, innovation, creativity and data-driven decision-making. AI has the potential to make organisations faster, more responsive and more capable of processing complexity. AI can help organisations analyse more data, generate more content and optimise more activities. What it cannot do is determine which investments should be protected when budgets come under pressure. It cannot decide how to balance short-term performance against long-term demand creation; it cannot resolve competing priorities across departments; it cannot create organisational alignment around growth objectives. These remain fundamentally human decisions. This suggests that the future role of insights will increasingly involve helping organisations navigate complexity rather than simply reducing uncertainty. The distinction is subtle but important. Reducing uncertainty implies providing more information until a decision becomes obvious. Navigating complexity acknowledges that many business decisions remain difficult even when information is available. Trade-offs still need to be made, priorities still need to be established, resources remain limited. Insights professionals therefore need to become more involved in the conversations where these trade-offs occur.

Viewed through this lens, the metamorphosis of insights is not simply a change in tools or methods; it is a change in organisational position. Insights are gradually moving from the periphery of decision-making towards its centre. The strategic growth gap suggests that this movement is not merely desirable. It may be one of the factors that distinguishes organisations that consistently grow from those that do not. The strategic growth gap suggests that the future of insight is not measurement alone. Measurement remains essential, but it is increasingly becoming a starting point rather than an end point. The organisations that consistently grow are not necessarily those with the most data, the most dashboards or the most sophisticated technology. They are the organisations that are best able to translate insight into decisions. In that sense, the future role of insight is shifting from understanding growth to helping create it.

Copyright © **Esomar** 2026


## Conclusion

The original ambition of the brand growth platform was relatively straightforward. We wanted to better understand what drives brand growth. Over the course of 10 years, the programme explored a wide range of growth ingredients, including innovation, media, creativity, data-driven decision-making, moments of truth and promotions. Each of these ingredients contributed valuable insights, and each helped explain part of the growth puzzle. Yet, perhaps the most important lesson that emerged was that sustainable growth is rarely the result of a single capability. The evidence presented throughout this paper suggests that growth is better understood as an organisational phenomenon than as a purely marketing phenomenon. Winning organisations do not simply innovate more. They do not merely invest more in media or conduct more research. Rather, they demonstrate a greater ability to consistently apply growth principles across multiple domains. They maintain strategic consistency, continue investing in future demand creation, involve customers more actively in innovation, make greater use of data and show stronger organisational commitment to long-term growth objectives.

This recurring pattern led to the identification of what we have called the strategic growth gap. The gap does not reflect a difference in access to information, technology or marketing knowledge. In today's environment, most organisations have access to the same books, the same conferences, the same agencies and increasingly the same technologies. The strategic growth gap reflects something more fundamental: the ability to translate knowledge into decisions that compound over time. Growth principles only create value when they influence resource allocation, investment decisions, innovation priorities and organisational behaviour. The challenge facing organisations is therefore not discovering what drives growth. The challenge is consistently acting on what they already know. This conclusion has important implications for the insights industry. Throughout the research, one observation repeatedly surfaced: the organisations that consistently grow appear to use insight differently. In these organisations, insight is not limited to measurement, reporting or evaluation. It plays a more active role in shaping decisions. Research is used to identify opportunities before investments are made. Data informs strategic priorities rather than simply monitoring outcomes. Customer understanding becomes an input into organisational direction rather than an explanation of past performance.

We have described this development as the changing role of insight. While much of the current discussion within the industry focuses on AI, automation and analytical innovation, the findings of this paper suggest that the most significant transformation may be organisational rather than technological. The future role of insight is not simply to generate more information. It is to increase the quality of decisions. In many respects, this represents a return to the original purpose of research. The value of insight has never resided in data itself. It resides in its ability to improve judgement. As organisations face increasing complexity, greater uncertainty and an ever-expanding volume of information, that role becomes more important rather than less. The organisations that succeed will not necessarily be those with the most data. They will be those that are best able to convert evidence into action.

After 10 years of research, the conclusion is therefore both simple and challenging. Sustainable growth is not primarily a knowledge problem; it is a decision-making problem. The strategic growth gap describes the difference between organisations that consistently translate growth principles into action and those that do not. Insights plays a critical role in determining on which side of that gap an organisation operates. When insights become part of how organisations choose, prioritise and act, rather than merely how they measure and evaluate, they move from the sidelines of growth to its very centre. After 10 years of research, that may be the simplest explanation for why some organisations consistently grow while others do not.

## About the Authors

Lucas Hulsebos, CEO, DVJ Insights, Utrecht, The Netherlands.

# Dialogues Within

## A Dialogical Self-Exploration of Men Becoming Fathers

*Sandeep Dutta*

*Paroo Tikoo*

## Esomar

Office address:

Burgemeester Stramanweg 105

1101 AA, Amsterdam

Phone: +31 20 664 21 41

Email: info@Esomar.org

Website: www.Esomar.org

Publication Date: September 2026

ESOMAR Publication Series Volume S424 Congress 2026

ISBN 92-831-0337-8

Congress 2026
Copyright © **ESOMAR** 2026

Copyright © **ESOMAR** 2026


# Dialogues Within

## A Dialogical Self-Exploration of Men Becoming Fathers

*Sandeep Dutta*

*Paroo Tikoo*

### The Question Nobody Was Asking

Pregnancy has traditionally been understood through the experiences of women. Maternal health programmes, pregnancy communication and infant nutrition brands have focused on understanding mothers' needs, anxieties and aspirations. Fathers, when acknowledged at all, have largely appeared as invisible supporters rather than active participants.

Yet contemporary fatherhood is changing.

Across urban India, first-time fathers are increasingly attending scans, researching nutrition, accompanying partners to consultations and actively preparing for the arrival of their child. Pregnancy is becoming less a "for mother-to-be only" experience and more a couple's journey.

For a yet-to-be launched maternal nutrition brand seeking to understand this shift, a fundamental question emerged:

What is happening inside the minds of today's expectant fathers? If the brand were to speak to them, what should it be saying?

Answering this question required more than understanding their minds at a surface level; It required a deep understanding the multiple emotions through which men experience this life-changing transition.

### Beyond the Baby Bump: The Untold Story of Expectant Fathers in Modern India

For a leading global health and wellness company seeking to launch a maternal nutrition portfolio in India, understanding the evolving role of fathers during pregnancy was not simply an academic question; it was a strategic imperative.

Traditionally, pregnancy in India has been a woman-centric journey. Decisions were shaped by the expectant mother, her mother, mother-in-law and healthcare providers, while the husband's role was largely that of a provider and a bystander.

Today, that reality is changing.

Driven by nuclear families and evolving expectations of partnership, pregnancy is increasingly becoming a shared experience. Women no longer seek financial support; they seek intense participation and partnership.

They expect their husbands to be emotionally present, invested and accountable alongside them.

A striking symbol of this shift can be seen during Karva Chauth, a traditional North Indian ritual in which wives fast for their husbands' wellbeing. Increasingly, husbands choose to fast alongside their wives not because tradition requires it, but because partnership does.

The fasting husband has become a cultural shorthand for a new kind of Indian man: one who wishes to share not only the celebrations of marriage, but also its sacrifices.

The same desire for partnership increasingly shapes pregnancy. Men want to participate not only in the practical aspects of the journey, but also in its emotional highs and lows.

At the same time, parenthood itself is being redefined. With many couples planning for only one child and entering parenthood later in life, pregnancy has become a once in a lifetime journey to be experienced fully rather than delegated to family elders. Men are attending scans, researching nutrition, accompanying partners to consultations and seeking a deeper connection with the unborn child long before birth.

*"We planned for this baby for almost three years. Before taking the decision, I wanted to be sure we were financially ready…not just for the pregnancy, but for everything that comes after. Since this may be our only child, I don't want to miss any part of the journey. I go for appointments, help my wife at home, discuss nutrition with her and try to support her emotionally. From the babymoon to the photoshoot, I want us to experience every moment together and make it special."*

– Anand, expectant father

Recognising these shifts, the client saw an opportunity to move beyond communicating solely with mothers. While women remained the primary audience, the client wanted to understand the emerging father and his aspirations, anxieties and unmet needs so that communication could authentically reflect the realities of modern fatherhood.

Yet this presented a significant research challenge.

While extensive literature exists on the experiences of expectant mothers, comparatively little research has explored the inner world of expectant fathers, particularly in the Indian context.

Their changing role is increasingly visible; their emotional journey remains largely unseen.

## Beyond Faster Answers: Searching for the Right Lens

Before finalising the research design, we conducted a pilot with expectant fathers from our extended networks. Men spoke readily about preparations, responsibilities and future plans, but struggled to articulate the deeper emotional and psychological dimensions of the journey.

The issue was not a lack of feeling, but a lack of language.

Beneath these practical narratives lay hints of uncertainty, pride, anticipation, fear and identity change. We realised that conventional questioning could capture what men were doing, but not how they were making sense of becoming fathers.

At a time when qualitative research increasingly relies on rapid analysis and automated processing of volumes of data, it would have been easy to focus only on what participants explicitly said.

Our interest, however, lay in what surfaced indirectly: contradictions, tensions, hesitations and shifts between different ways of understanding oneself.

## Dialogical Self Theory: Listening to the Voices Within

Dialogical Self Theory (DST), developed by Dutch psychologist Hubert Hermans in the 1990s, conceptualises the human mind as a "society of mind", viewing the self not as "a unified entity but a dynamic multiplicity of relatively autonomous I-positions, engaged in an internal dialogue" (Hermans, 2001: 248).

Rather than asking what a person feels, DST asks:

Which voices are speaking, which voices are competing and how do they negotiate meaning?

This perspective is particularly valuable during major life transitions, when existing identities are challenged and new ones emerge, often colliding with each other.

The transition to fatherhood represents exactly such a moment.

This insight informed both our methodology and our analysis.

## Method: Making Internal Dialogues Visible

Participants: Expectant and new fathers aged 25–40 years were recruited across key stages of the parenthood journey, including pre-conception, pregnancy and early postnatal.
The objective was not simply to capture experiences but to reveal emerging voices and tension.

### *Stage 1: Reflective Diaries*

To stimulate reflection without leading the men, we designed diary prompts focused on everyday thoughts and experiences.

Unlike conventional diaries that focus on events, these diaries were designed to surface multiple perspectives within the self.

Prompts encouraged participants to reflect on:

Different thoughts competing for attention.
Moments of conflict or uncertainty.
Thoughts they shared versus thoughts they kept to themselves.
How their perception of fatherhood was changing over time.

### *Stage 2: Internal Dialogue Interviews*

Insights from the diaries informed individual depth interviews. Participants were invited to revisit significant moments and explore the different perspectives active within them. To make internal dialogues visible, interviews incorporated two dialogical elicitation techniques inspired by DST.

### *Internal Dialogue Reconstruction*

Participants revisited key moments such as discovering the pregnancy or attending scans and described the different thoughts that were simultaneously present.

### *Empty Chair Dialogues*

The most powerful technique involved a modified empty chair exercise.

Participants were asked to imagine different parts of themselves occupying separate chairs. For example, one chair might represent the Responsible Provider while the other represented the Nurturing Father. Participants then described the conversation taking place between these positions.

Rather than asking men directly how they felt, the exercise enabled them to externalise competing viewpoints and articulate tensions that might otherwise have remained hidden.

As one participant enacted a conversation between two parts of himself:

Responsible Provider: *"We need to focus on saving. Everything else can wait."*
Nurturing Father: *"But this is the only time I'll experience this journey to fatherhood. I don't want to miss it."*
Responsible Provider: *"If I don't prepare, I won't be doing my job."*
Nurturing Father: *"If I'm never present, I won't be doing my job either."*

The exercise transformed abstract emotions into observable dialogues and provided a direct route into the competing identity positions that characterised the transition to fatherhood.

## Analysis: Personal Position Repertoire (PPR)

Data analysis used the Personal Position Repertoire (PPR) grid, a structured analytical tool derived from DST methodology (Hermans & Hermans-Jansen, 1995).

### *Beyond Themes: What the PPR Grid Revealed*

The PPR grid enabled a structural insight that would not have emerged through thematic analysis alone:

The transition to fatherhood was not marked by the emergence of a single paternal identity, but by an ongoing negotiation between multiple voices, each competing to define what it means to be a good father in contemporary India.

Copyright © **ESOMAR** 2026


## Findings: Five Selves, One Journey

Across participants, five dominant and recurring I-positions emerged consistently enough to form the analytical backbone of the study.

### *Responsible Masculinity*

The "I-as-Responsible-Provider": The Responsible Provider was the most dominant and persistent voice across all stages of the journey. While rooted in traditional expectations of provision, it extended beyond finances. Planning became a way of expressing care and managing uncertainty.

*"Today, while driving back from work, I found myself thinking about school fees… even though the baby isn't here yet. Later, I spent an hour reviewing our finances and checking insurance policies. My wife says I worry too much, but I don't see it as worry. I just want everything to be in place. Maybe this is what becoming a father feels like… not thinking about yourself anymore, but constantly thinking three steps ahead for your family."*

What appears on the surface as financial planning is often something deeper. Rather than expressing anxiety directly, many men converted uncertainty into action.

Worry became preparation. Vulnerability became planning.

### *Regulated Masculinity*

The "I-as-Strong-Man": Running alongside the Responsible Provider was the Strong Man, a voice shaped by deeply embedded expectations of masculinity.

Men frequently described feeling responsible for remaining calm and reassuring.

*"My wife asked me today if I was nervous about the delivery. I told her everything would be fine. The truth is, I do think about it sometimes… whether everything will go smoothly, whether I'll know what to do if something goes wrong. But I don't see the point of saying all that out loud. Right now, she needs me to be calm. If both of us start worrying, who will keep things together?"*

What appears as confidence is often emotional management. Many men felt responsible not only for providing stability, but also for projecting it.

### *Expressive Masculinity*

The "I-as-Emotional Husband": Beneath the Strong Man there was a voice, at times a bit suppressed, that was mellow and soaked in love. Many men spoke about concern for their partners and the emotional challenge of witnessing her physical discomfort during pregnancy and not being able to do much.

*"My wife was struggling with nausea and exhaustion. I knew it was normal, but watching her go through it every day was difficult. I kept wishing I could take some of it away from her."*

This voice reflected empathy, love and shared responsibility.

Copyright © **ESOMAR** 2026


### *Nurturing Masculinity*

The "I-as-Nurturing-Father": Perhaps the most unexpected finding was the emergence of the Nurturing Father. For many men, this voice surfaced most strongly during ultrasound appointments.

Seeing the baby for the first time or hearing the heartbeat transformed fatherhood from an abstract future responsibility into an emotional reality.

*"Today was the scan. I had seen pictures of babies before, but seeing my own child on the screen felt completely different. On the way back, I kept thinking about what kind of father I want to be. My father worked hard and gave us everything we needed, but he was never very expressive. I want my child to know that I am there… not just to provide, but to listen, play, talk and be part of their everyday life. For the first time, becoming a father felt real."*

This voice was characterised by tenderness, pride and a desire for emotional closeness. Where previous generations focused on provision, these men sought presence, emotional engagement and a deeper relationship with their children from the very beginning.

### *Reflective Masculinity*

The "I-as-Son": Alongside the emerging father sat another influential voice: the Son. As men imagined the fathers they wanted to become, they frequently found themselves revisiting the fathers who had raised them. Memories of childhood, admiration, gratitude and, at times, unresolved gaps became important reference points in shaping their own parenting aspirations.

*"I was thinking about my father today. He worked incredibly hard and made sure we never lacked anything. I have a lot of respect for him. But I also realised that most of my memories of him are about responsibility, not conversations. As I prepare to become a father, I find myself asking what I want to carry forward and what I want to do differently. I want my child to have the security my father gave me, but I also want them to feel comfortable sharing things with me that I never shared with him."*

The father was more than a source of memories; it was a reference point against which men evaluated themselves. It shaped what good provision looked like, what strength meant and what kind of emotional presence they hoped to offer.

The father a man was becoming was often in dialogue with the son he once was.

## Personal Position Repertoire (PPR): Mapping the Journey

The PPR grid revealed a structural insight that would not have emerged through thematic analysis alone:'

| I-Position | Pre-conception | Pregnancy | Early postnatal (0–3 months) | Tension |
|---|---|---|---|---|
| **I-as-Responsible-Provider** | Focus on preparing, planning and financial readiness for parenthood | Becomes the dominant organiser of behaviour as responsibilities become more immediate and tangible | Intensifies further as the realities of caring for a newborn sets in | Presence vs. Provision |

| | | | | |
|---|---|---|---|---|
| **I-as-Strong-Man** | Sees himself as the stabilising figure who must remain calm and reassuring | Manages growing anxiety by prioritising stability, often suppressing personal concerns | Continues to put family needs ahead of his own while maintaining emotional control | Expression vs. Control |
| **I-as-Emotional Husband** | Concern for partner's wellbeing begins to emerge | Empathy deepens, peaking around delivery, health concerns and readiness for fatherhood | Evolves into shared caregiving and emotional partnership | Self vs. Partner |
| **I-as-Nurturing-Father** | Largely aspirational; fatherhood remains an abstract future identity | Activated through tangible pregnancy milestones, fostering a desire for involvement and bonding | Becomes highly visible through caregiving, bonding and everyday interactions with the baby | Care vs. Competence |
| **I-as-Son** | Reflects on his own upbringing and father's role as a template | Evaluates what aspects of family traditions and parenting to retain or change | Re-emerges strongly as memories and parental models inform caregiving practices | Continuity vs. Change |

*Table 1.*

The transition to fatherhood was not marked by the emergence of a single paternal identity, but by an ongoing negotiation between five distinct voices, each competing to define what it means to be a "good father" in contemporary India.

## Strategic Implications for the Maternal Nutrition Category

The value of the PPR extended beyond understanding fathers' emotions. It helped identify which identity positions offered the most meaningful opportunities for engagement.

While the Responsible Provider remained the most dominant voice throughout the journey, it was also the most culturally recognised. Family members, healthcare professionals and society already reinforce the father's responsibility to provide and protect.

The greater opportunity lay elsewhere.

Two emerging positions stood out as particularly relevant for the brand:

### *The Emotional Husband*

Pregnancy was experienced not only as preparation for a child but also as concern for a partner. Many men spoke about feeling emotionally invested in their wives' wellbeing while simultaneously feeling powerless to ease the physical demands of pregnancy.

This position reframes maternal nutrition from a functional health product into an act of care. Supporting nutrition becomes a way of supporting the mother.

### *The Nurturing Father*

Participants repeatedly expressed a desire to be more emotionally present than previous generations of fathers. They wanted to participate, connect and contribute throughout the journey rather than simply provide from the sidelines.

This position represents an emerging cultural model of fatherhood that is still seeking validation.

For the client, these findings suggested that the strongest opportunity did not lie in reinforcing the traditional Provider identity, but in acknowledging and legitimising the Emotional Husband and the Nurturing Father.

Rather than speaking to fathers as secondary supporters, communication could position them as active emotional participants in the pregnancy journey.

In doing so, maternal nutrition shifts from being a product consumed by the mother to becoming a shared act of care undertaken by the couple together.

## Implications for Future Research

This study demonstrates the value of Dialogical Self Theory not only as an analytical framework but also as a research methodology capable of uncovering the competing voices that shape human decision-making.

Many significant consumer decisions are characterised by internal dialogue rather than simple preference formation.

The decision to purchase a home may involve competing voices of aspiration and financial caution. Choosing a vehicle may involve tensions between practicality, status and self-expression. Adopting a health and wellness programme may require negotiation between discipline, comfort and long-term wellbeing.

Traditional research often captures the outcome of these decisions. DST helps reveal the conversations that precede them.

By encouraging participants to externalise competing viewpoints through techniques such as dialogical diaries and empty chair dialogues, researchers can access motivations, tensions and identity negotiations that might otherwise remain hidden.

The contribution of DST therefore extends beyond fatherhood.

It offers a framework for understanding any context in which consumers are navigating uncertainty, balancing competing priorities or investing significant emotional, financial or physical resources.

In such situations, decisions are rarely driven by a single motivation. They emerge from a dialogue between multiple selves.

DST provides researchers with a systematic way of making those dialogues visible.

## Last Word

This study revealed that becoming a father is not a single emotional transition but an ongoing conversation between multiple selves. By using Dialogical Self Theory to both generate and analyse data, we were able to move beyond

identifying emotions towards understanding the dialogues that produce them. More importantly, the study demonstrates that when decisions matter deeply, whether becoming a parent, purchasing a home or transforming one's health, it is often the conversation and tensions within the consumer that shapes the choice.

The challenge for researchers is not simply to listen to what people say, but to uncover the different voices speaking within them. Not every voice carries equal significance, nor does every voice represent an opportunity for brands. The true value of insight lies in identifying the identities that are most meaningful, most underserved and most open to engagement. By making these internal dialogues visible, researchers enable brands to move beyond addressing needs and begin speaking to the voices that shape them.

## References

Hermans, H. J. M. (2001). The dialogical self: Toward a theory of personal and cultural positioning. *Culture & Psychology, 7*(3), 243–281.

Hermans, H. J. M., & Gieser, T. (Eds.). (2012). *Handbook of Dialogical Self Theory*. Cambridge University Press.

Hermans, H. J. M., & Hermans-Jansen, E. (1995). *Self-Narratives: The Construction of Meaning in Psychotherapy*. Guilford Press.

## About the Authors

Sandeep Dutta, Vice President, Kantar, India.

Paroo Tikoo, Account Director, Kantar, India.

# The Wellbeing Shift

## A Human–AI Metamorphosis of Brand Positioning

*Charlene Van Zyl*

*Karlien Kriegler*

*David Wright*

## Esomar

Office address:

Burgemeester Stramanweg 105

1101 AA, Amsterdam

Phone: +31 20 664 21 41

Email: info@Esomar.org

Website: www.Esomar.org

Publication Date: September 2026

ESOMAR Publication Series Volume S424 Congress 2026

ISBN 92-831-0337-8

Congress 2026
Copyright © **ESOMAR** 2026

Copyright © **ESOMAR** 2026
Congress 2026

# The Wellbeing Shift

## A Human–AI Metamorphosis of Brand Positioning

*Charlene Van Zyl*

*Karlien Kriegler*

*David Wright*

### 1. It's Time to Move Insight from its Dusty Corner Shelf

Most brand positioning projects begin with a promise: that a deep understanding of people will lead to stronger brand ideas. Yet somewhere between the research debrief and the act of creation, much of that understanding can become abstract, diluted or difficult to use.

The role of the insight professional traditionally has been to inform. We gather the evidence, interpret the consumer, write the report and hand over the findings. The work may be rich, thoughtful and strategically useful, but it still relies on other teams to absorb it, remember it and translate it into brand ideas, symbols, stories, visuals and campaigns.

There is another issue too. Most organisations are already sitting on a rich body of existing knowledge: past research, brand tracking, cultural learning, category studies, customer feedback, segmentation work and trend reports. These sources often contain patterns that are still highly valuable, but once a project is complete, much of this intelligence is shelved. It may be stored, referenced occasionally or rediscovered when someone remembers it exists, but it is rarely brought back into the live process of innovation and brand development.

This creates a missed opportunity. The real value often lies not in one new study, but in connecting the dots between what the organisation already knows, what consumers are saying now, and what cultural or category trends suggest may happen next. Better innovation and stronger positioning come from this combination of old and new knowledge: accumulated insight, fresh consumer interpretation and future-facing signals working together.

This creates a familiar gap between consumer understanding and brand expression. The market research report is referred to, but creative and marketing teams are not always fully entrenched in the insight. They are often left to carry the burden of translation: turning needs, tensions and opportunities into something people can see, feel and respond to.

This matters because innovation success depends heavily on understanding customer needs. Yet firms often struggle to share and apply customer insights effectively. As noted by Stremersch et al., customer insights are particularly underused in areas such as imagining, crowdsourcing and co-creating (2024). In other words, insight is often present at the start of innovation, but less active in the moments where ideas are being shaped.

Many innovation and positioning processes also remain highly sequential. Research is conducted, findings are written up, a brief is developed, creative work begins, and ideas are then tested. Each phase depends on the previous one. If an

idea does not land, teams may need to go back, rework the thinking or commission new research. The process can become slow, fragmented and reactive.

Woolworths, like many organisations, is also still mainly product led. This means that innovation and insights are translated through a product-biased lens. Concepts and ideas shown to customers are rarely co-created with them and the insights are used to evolve and optimise existing positioning ideas and concepts. By using organisation-biased concepts as a starting point, consumer creativity is stifled, and the insights are limited to the baseline and context we establish upfront. Hidden needs and barriers are often missed. So, we are ultimately constrained by our own history and current ideas. What is required is revolution, not evolution.

The opportunity is thus to work differently. How can brands learn from past research, build on existing knowledge and create with both old and new insight in the same process? How can trend thinking be woven into development rather than added as a separate input? How can customers help shape and refine opportunities earlier, while also recognising that people do not always know how to express what they feel, imagine or need?

This is especially important in brand work. Consumers are not always able to articulate what a brand means to them in direct language. Some responses are emotional, visual, symbolic or half-formed. We therefore need methods that help people express ideas and feelings in more abstract ways, and that allow those expressions to become part of the creative process.

So, the question at the heart of this paper is simple: what if insight could change form? What if it could move beyond the report, beyond the handover and beyond the role of informing, and become an active part of brand and concept creation itself?

In the context of Woolworths, a move into the world of wellbeing represented just such an insights revolution opportunity. Wellbeing is very abstract and emotive and has a broad spectrum of meaning and application for consumers. To get it right and capture consumers' hearts, minds and wallets, Woolworths needed a different approach to brand, concept and product range positioning. In a growing and cluttered marketplace with significant revenue potential, getting this right was paramount.

## 2. The Metamorphosis: From a Shelved Report to Living Creative Intelligence

### *2.1 The Pathway to AI Usefulness*

AI's first value in market research was largely one of scale. It could tag, sort and theme large volumes of verbatim responses far faster than manual coding teams. The next stage was summarisation: turning complex research outputs into clearer, shorter and more usable forms.

More recently, AI has allowed researchers to interrogate data more actively: asking for different cuts, challenging structures, exploring patterns and refining outputs through conversation. But even this remained largely report-facing. The researcher asked, the AI answered, and the output was still usually destined for a report or presentation.

This is useful when the task was one dataset, one study or one question at a time. But it is not enough for a challenge like Woolworths' move into wellbeing, where the task is not simply to analyse insight faster, but to connect existing knowledge, fresh consumer meaning and future-facing opportunity into a more active creative process.

### *2.2 From Synthesis to Living Intelligence*

As AI tools have improved, the role of insight has also begun to shift. The real opportunity is no longer simply to synthesise more quickly, or to produce a clearer summary of a single study, but to change the status of insight itself from something stored and retrieved, to something active, connected and available inside the creative process.

In the traditional model, organisational knowledge often sits behind the project. Previous studies, brand history, category learning, customer feedback, cultural observations and past campaigns may all exist, but they are not always easy to bring into the moment where new ideas are being shaped. They are remembered unevenly, searched for manually, or reintroduced only when someone knows where to look.

Living intelligence works differently. It allows past knowledge, current consumer input, brand context, cultural signals and commercial priorities to sit inside the same working space. The researcher can move beyond saying, “this is what people told us in this study”, towards a more connected, continuous interpretation: “this is what we know, this is what is changing, this is what consumers are beginning to express, and this is what the brand could do next”.

This matters because brand and innovation challenges are rarely solved by one dataset alone. They require memory, interpretation and imagination. AI can help hold and connect more of this material at once, but the value does not come from automation alone. It comes from collaboration: the researcher challenging the output, applying judgement, asking better questions and deciding what is meaningful, credible and strategically useful.

This also reflects a wider shift in how organisations are beginning to think about AI: not only as a route to efficiency, but to support growth. As Benartzi et al. argue (2026), companies often suffer from a “growth blind spot” when they focus AI investments on cost-cutting while overlooking its potential to create new value.

In this sense, AI becomes less a reporting tool and more a creative intelligence layer. It can help compare ideas, reframe territories, surface tensions, generate possibilities and translate insight into forms that others can work with positioning routes, concept directions, visual metaphors, storylines, product ideas or customer-facing expressions.

### *2.3 Elevating Respondents from Reactors to Co-Creators*

Highly targeted questions can produce clean answers, but they can also narrow discovery too early. By focusing respondents on a predefined frame, we may reduce their ability to imagine alternatives, surface unmet needs or express ideas that sit outside the categories we have already chosen.

Generative AI opens a different possibility. Respondents can move from reacting to ideas towards helping to shape them. This matters because people do not always know how to express abstract, emotional or symbolic needs directly. AI can help scaffold that expression by probing, prompting and helping people build their thinking.

Through creative methodology, task design and AI-supported conversation, respondents can be helped to describe a feeling, imagine a moment, build on a prompt, improve an idea or visualise something that is difficult to explain in words. The respondent is no longer simply answering a question or choosing between options. They can create, adapt and refine with the support of AI.

Some people need choices before they can respond. Others need small prompts and follow-up questions. Others are more comfortable building freely. AI can support these different modes of participation, allowing more people to contribute meaningfully to the development of ideas.

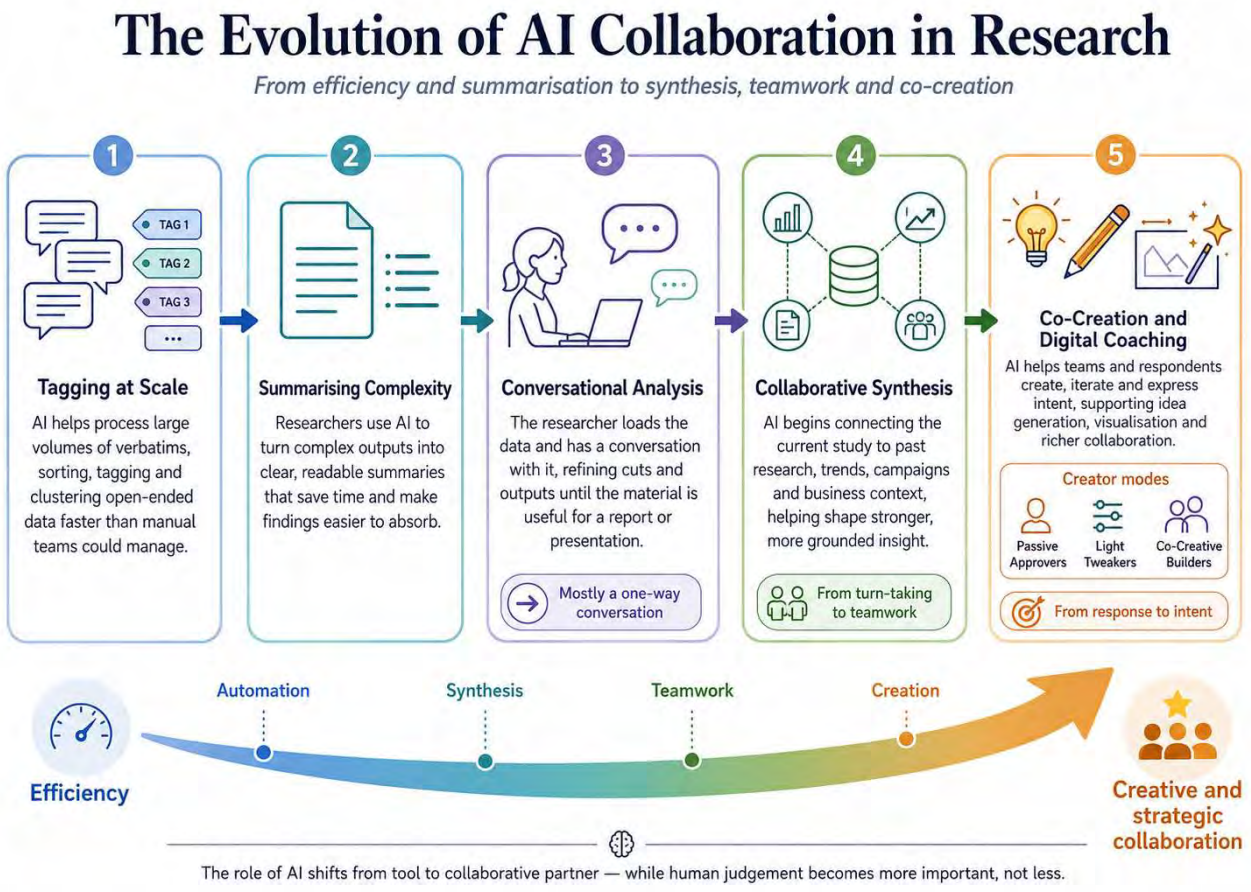



*Figure 1. The evolution of AI collaboration in research.*

## 3. The Woolworths Foods Challenge: Giving Wellbeing a Clearer Brand Role

This paper is grounded in a real brand challenge for Woolworths, one of South Africa's most established retail brands. Woolworths already holds strong associations around quality, trust, innovation and premium value. It is a brand with a clear heritage and a significant body of accumulated consumer and category knowledge behind it.

Wellbeing is becoming an increasingly important driver in how people think. Everyday life has become more psychologically demanding. COVID-19 intensified this pressure: the World Health Organization reported that, in the first year of the pandemic, global anxiety and depression increased by 25%, describing this as a "wake-up call" for countries to improve mental-health support (2022). At the same time, consumers are navigating economic pressure, health uncertainty, digital comparison, work insecurity and constant information overload. Wellbeing becomes less of a lifestyle aspiration and more of a way to preserve stability, inner peace, control and emotional resilience.

McKinsey's 2025 *Future of Wellness* research shows that wellness is becoming more sophisticated, science-led and segmented, with younger consumers especially active in researching and adopting wellness solutions. McKinsey also estimates the global consumer wellness market at $1.8 trillion, with consumers increasingly demanding "credible,

science-backed wellness solutions" (2024). But wellbeing is also a difficult territory for brands. It can quickly become generic claim-washing, be overly functional or come across as unfun.

For Woolworths, the opportunity is not to "own health" or add another layer of messaging, but to explore how wellbeing could evolve the brand's meaning while staying true to what people already value in the brand: quality, trust, premium value and everyday enjoyment. Getting this positioning right represents significant revenue potential – both through the identified wellbeing product range, and through strengthening overall brand connection long term by solving a complex and evolving need.

The task, then, was to create more than a campaign direction. It was to identify a future-ready wellbeing platform that could give the brand focus. The brand needed to do many things, but the platform needed to create a clear centre of gravity: an idea broad enough to guide communication, innovation and future storytelling, but specific enough to feel distinctive, useful and true to Woolworths.

Over time, the business had built up a rich understanding of its customers, its brand strengths and the broader shifts shaping food choices.

The question was how to bring Woolworths' institutional knowledge back into the live process of brand development, and how to combine it with new consumer input and future-facing trends in a way that could guide action.

This made the case a useful test of the wider argument in this paper. Could existing insight, new consumer co-creation and trend-led thinking be brought together in one process? Could wellbeing be translated from an abstract consumer driver into a more actionable brand world? And could insight move beyond informing the brand team, to actively helping shape the wellbeing platform itself?

## 4. The Method: Designing an Insight System for Brand Creation

Collaborative intelligence is the idea that human judgement and artificial intelligence work together, each doing what they are best suited to do. In an insight context, this means AI can help organise large bodies of knowledge, identify patterns, connect different sources of evidence and generate possible directions. Humans then bring interpretation, context, creativity and strategic judgement. The value is not in asking AI for "the answer", but in creating a working environment where insight can be explored, challenged and transformed into tangible solutions.

For this project, we used a proprietary collaborative intelligence platform developed by Hello Ara over the past 18 months. Internally referred to as *The Whale*, the platform was designed to support researchers and brand teams in working with large, multi-layered bodies of research evidence in a more dynamic and generative way.

The system brings together accumulated research inputs, structured and unstructured data, and strategic knowledge into an interactive environment where patterns can be explored, synthesised and translated into insight-led outputs. Rather than treating research as a static archive, the platform enables users to interrogate existing knowledge, surface connections, develop hypotheses, and move from insight into strategic and creative possibilities.

The platform supports both analytical and generative workflows. Depending on the research need, it can assist with synthesis, idea development, visualisation, creative stimulus generation and workshop collaboration. Outputs may

include written summaries, strategic territories, idea boards, visual references, video or audio-based stimulus, and other editable materials that support interpretation and co-creation.

The interface is designed to be accessible to non-technical users, with both voice and text-based interaction. It can be used individually by researchers or marketers, as well as collaboratively in workshop settings. In this study, the platform functioned as a shared intelligence space in which researchers, marketers and consumers could engage with accumulated knowledge in real time, helping to transform research findings into practical brand-building opportunities.

The project was designed as a four-phase process moving from future exploration through to consumer co-creation. Rather than beginning with brand positioning directly, we started by exploring how wellbeing might evolve in South Africa over the next decade and then progressively narrowed this into a practical wellbeing platform for Woolworths.

*Figure 2. The journey of insight transformation with Whale AI.*

## *Phase 1: Mapping the Future Landscape of Wellbeing*

### *Sourcing Existing Data*

The first phase focused on understanding the future of wellbeing before exploring what role Woolworths could play within it through activating what the organisation already knew, while also creating a stronger foundation for future-facing brand development. Woolworths was not starting from a blank page. The business had accumulated a rich body of insight over time, including historical research, brand understanding, product innovation thinking, trend material and internal category knowledge. The challenge was how to bring this intelligence back into the live process of positioning development, rather than treating it as background material or archived learning.

The first step was therefore to collect, curate and structure the relevant data so that it could be used to train Whale AI, Hello Ara's AI collaboration platform. This included existing research and available knowledge sources that could help the team understand how customers were thinking about food, value, quality, trust and wellbeing. The purpose was not

simply to store the information in one place, but to make it usable: to create a working insight environment that could be questioned, explored and built on.

The question was not only what wellbeing meant now, but where these signals might be heading and how they could affect future expectations of food, groceries, personal care, home care and cooking-related products.

This was important because wellbeing was no longer being treated as a narrow health or lifestyle idea. Woolworths was entering a period where wellbeing was becoming a broader consumer driver among upper-income South Africans. Consumers were becoming more aware of issues such as microplastics, chemicals, additives, seed oils, dyes, preservatives, sustainability impacts, homeware safety and the perceived effects of everyday products on the body and home. These concerns were not only about food in a functional sense. They reflected a wider search for trust, reassurance, safety, energy, control and care in everyday life.

Woolworths needed to explore wellbeing from a broader perspective than food alone. We then added new qualitative evidence through eight group discussions to understand current perceptions, tensions and emerging needs in more depth. This research explored what wellbeing meant to customers today, what they were prioritising, what problems they were trying to solve, and how different consumers were clustering along the wellbeing spectrum. It also examined where Woolworths could play credibly, where the brand might overpromise, where it had permission to stretch, and which opportunities could offer the greatest commercial upside.

## *Adding Predictive Models to LLM Reasoning*

A key part of this phase was adding a future-facing lens. Whale AI was trained to work with a Monte Carlo-style or Future's Cone prediction model, allowing the team to break broader future possibilities into underlying assumptions and explore them across different time horizons. Existing data and trend data were then examined through near-, mid- and longer-term perspectives. We used our Collaboration platform to create interactive, models to tweak and visualise the models.

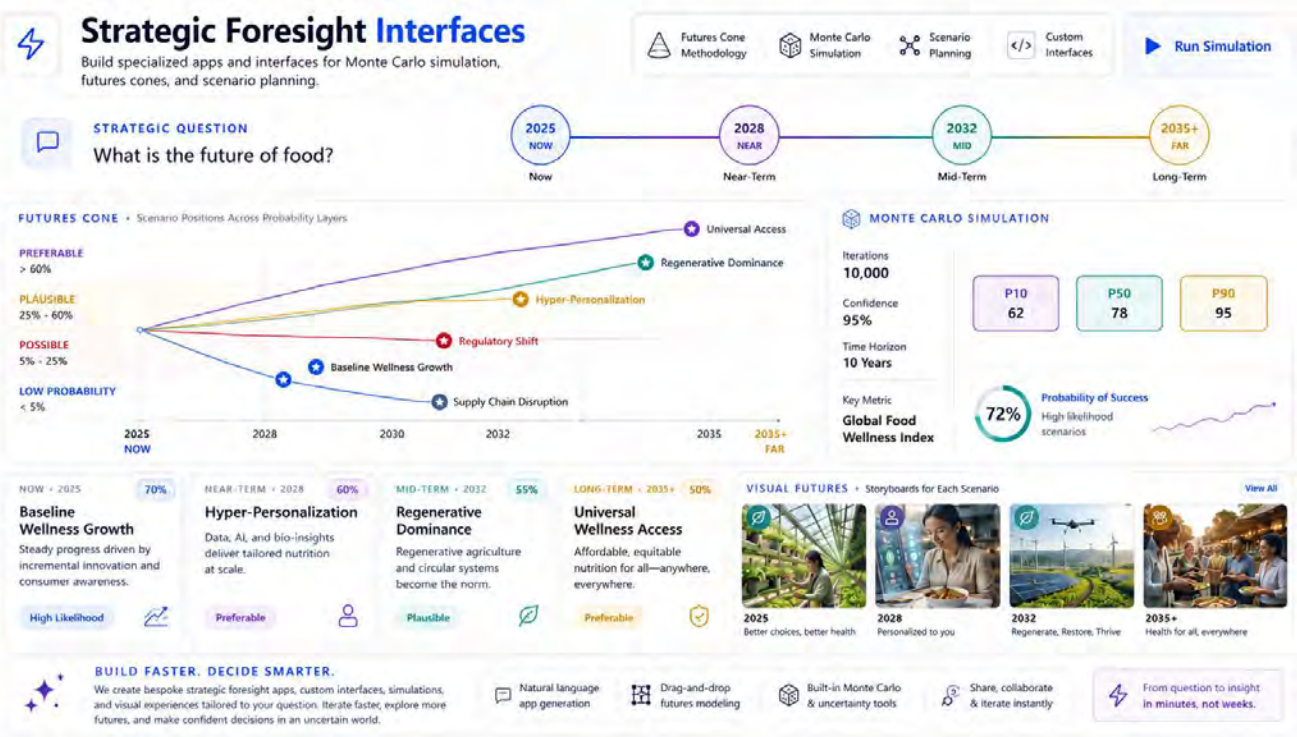



*Figure 3. Strategic foresight interfaces.*

However, we found the real power was taking the outputs from these models, integrating prediction model reports for the timeframes (Near, Mid and Far) into our collaboration platform as a data resource.

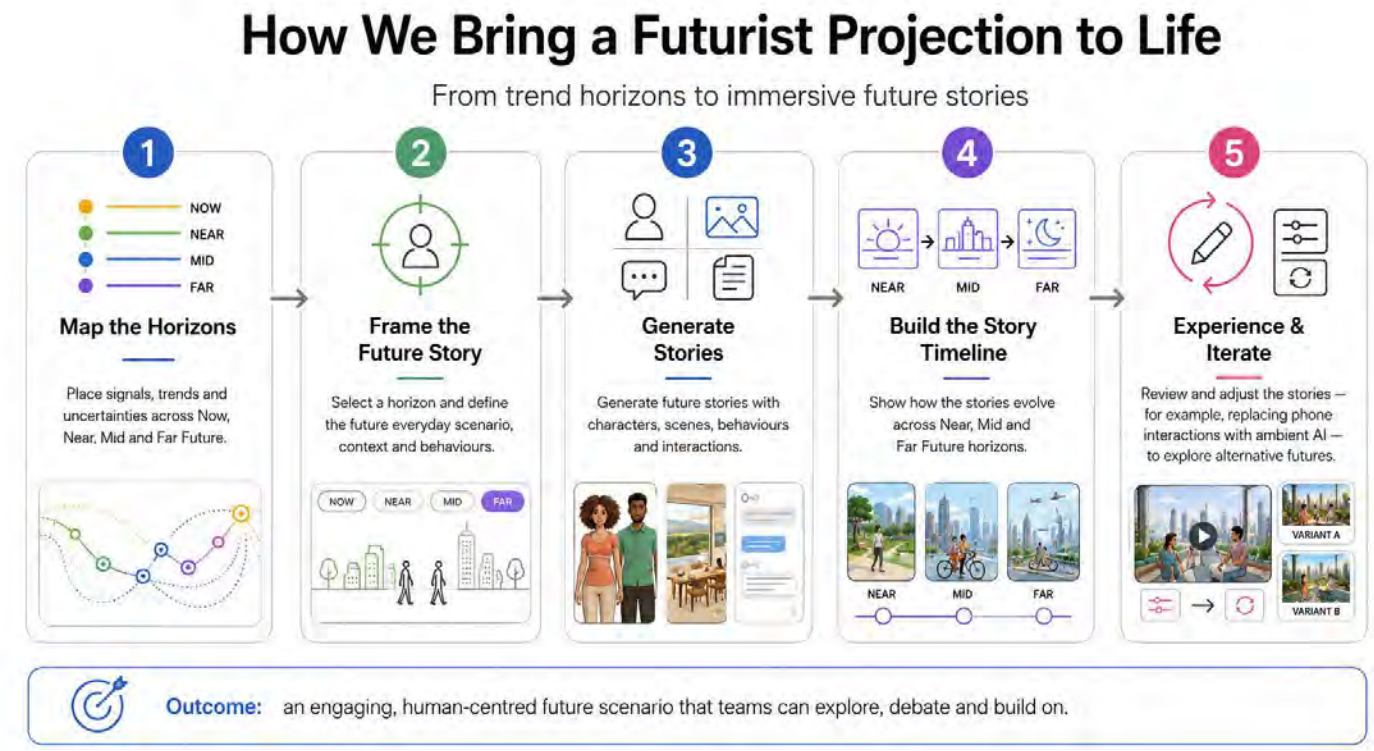



*Figure 4. How we bring a futurist projection to life.*

## *Additional Qualitative Research*

The qualitative phase added an important human layer to the existing intelligence base. It helped move the work beyond broad trend language and into the lived reality of consumers: their anxieties, routines, trade-offs, hopes and practical decisions. Wellbeing emerged not as a single idea, but as a spectrum of needs ranging from clean food and ingredient reassurance to mental health, hormone support, family care, sustainability, emotional comfort and everyday confidence.

After this qualitative phase, the new learning was added back into the Whale AI training set. This meant the platform was no longer working only from historical research and external trend material. It now also included fresh consumer understanding, emerging language, current tensions and new hypotheses about where the wellbeing opportunity could go. In this way, the system became progressively richer as the process unfolded.

The output of Phase 1 was not a final positioning answer, but a more coherent and future-ready foundation for exploration. It produced early brand platform territories, emotional narratives, visual directions and wellbeing hypotheses that could be taken into the next phase of consumer co-creation. This phase also gave the “backbone” training for collaboration with Whale AI in the following phases and throughout the work.

## *Phase 2: Collaborative Strategic Ideation*

The second phase focused on translating future wellbeing opportunities into strategic marketing territories.
Using Whale AI, client and agency teams collaborated directly with the future intelligence system. Ideas were generated through interactive idea boards (eight ideas at a time) generated by Whale AI based on what was asked of it. Multiple territories could be explored simultaneously, combined, refined and evolved through conversation with the AI.

One of the key workshops brought together an interdepartmental client-side team and the Hello Ara team. Working in Whale Idea board mode, Whale AI acted as a challenger, co-idea generator and live synthesis partner. This created an iterative process where future trends, organisational knowledge and human judgement were brought together in real time. This changed the feel of the brainstorm. The team could ask questions, push an idea further, test whether a thought connected back to the evidence, and quickly generate alternative ways of expressing a territory.

Over the course of a 90-minute session, the team developed a series of supporting positioning territories. Four territories were selected for further development.

### *Phase 3: Grounding Future Possibilities in Brand Meaning*

The third phase focused on ensuring that future wellbeing opportunities remained authentic to the Woolworths brand. The leading positioning territories were evaluated through a conversational AI quantitative study (n=300). Rather than relying solely on rating scales, respondents were also asked to explain their thinking and suggest improvements to the positioning ideas being tested. As part of the presentation deliverable, Hello Ara used Whale AI to help suggest changes and tweaks to the sub-positioning taglines, while retaining human judgement over what was strategically useful and brand appropriate.

This provided both validation and constructive consumer language. The findings were subsequently loaded back into Whale AI, allowing the positioning territories and supporting language to be refined further.

Following this, another conversational AI quantitative study (n=500) was conducted among Woolworths shoppers. A deeper emotional study of the Woolworths brand was needed to ensure that the wellness repositioning still took into consideration what consumers need from the master brand overall. Rather than testing concepts directly, the study explored what Woolworths means in people's lives, the role it plays, what differentiates it and what consumers would like from the brand in the future. The research dialogue consisted mostly of open questions and storytelling techniques, which allowed us to collect stories and emotions associated with the Woolworths brand. This helped reconfirm where Woolworths lives in the minds of consumers but also explored what they want from the brand in the future.

The findings were loaded back into Whale AI co-creation system. This allowed the emerging wellbeing territories to be refined using both future-oriented thinking and current brand meaning. Alongside this work, the team revisited the deeper emotional meaning of the Woolworths master brand to ensure the emerging wellbeing platform remained connected to the core brand essence.

The outcome of this phase formed the foundational input to new strategy and positioning/concepts which represented the desired shift to 360, customer led. This significantly increased the quality generated and bypassed several rounds of internal iteration and debate. Overall, greater clarity, alignment and efficiency was achieved.

### *Phase 4: From Evaluation to Consumer Co-Creation*

The final phase moved the project from positioning development towards practical implementation. At this point the dots could be connected. The team had the history, the trend direction, the consumer response to wellbeing, the validated positioning statements, and a deeper understanding of the Woolworths master brand. The next step was to zoom in again and get practical.

The activation question became: how would the positioning come to life? What would people want from products, rewards, education and support? How should the idea show up in store and online? What are the codes, colours and emotions? Who are the characters or personas that could bring the brand world to life? How could the brand create a living, breathing change that still connected back to quality, value, innovation and convenience, while delivering a distinctive wellbeing positioning?

For this next phase of the research, we explored a new application of Conversational AI. For this paper, we call it the "Generative AI Creativity Coach". Once the wellbeing territories had been narrowed and refined, a mobile-first version of Whale AI, known as "Whalets", was embedded directly into conversational interviews. Rather than asking respondents to evaluate ideas, they were invited to actively create them.

By embedding AI visualisation tools directly into the interview experience, respondents were able to move beyond answering questions and become active participants in the creative process.

Participants were positioned as extended members of the Woolworths brand team and invited to help build the future wellbeing platform. Through a combination of conversational questioning and AI-assisted creation, respondents generated, visualised and refined campaign ideas, stories, characters and brand expressions. This allowed us to capture not only their opinions, but also their creative intent, revealing how consumers naturally shape and improve ideas when given the tools to do so.

Participants were first introduced to emerging wellbeing territories and then worked collaboratively with AI to imagine how these ideas could come to life. Respondents defined the most important message Woolworths should communicate, who should appear in the campaign, the story that should unfold and the emotional response the communication should create.

These inputs were used to generate visual campaign storyboards in real time. AI transformed the respondents' thinking into a campaign concept that could immediately be seen and discussed.

The process did not stop at generation. Respondents were asked how the campaign could be improved. Their suggestions were incorporated into a new version. They were then asked how the idea could be made even more relevant for people like themselves, creating a further round of development.

This created a structured cycle of imagination, visualisation, critique and refinement. Rather than reacting to finished concepts, respondents became active participants in shaping them.

A brand, like an artwork, is not experienced only through rational meaning. It is experienced through first impression, focus, composition, colour, symbols, emotion, perspective and memory. Traditional research often captures what people say. But brand meaning often lives in what people feel, recognise, imagine and remember.

The method helped surface these less articulate dimensions through visualisation and prompting.

The approach generated two forms of learning. The first came from the outputs themselves, including campaign ideas, visual territories, messaging approaches and activation concepts. The second came from observing the creative process. By analysing how respondents changed ideas, what they chose to improve and where they focused their attention, we gained a deeper understanding of how wellbeing is interpreted and expressed in everyday life.

The final outputs extended well beyond campaign advertising. Consumers worked with AI to explore how the wellbeing platform could come to life through products, rewards, education, support services, in-store experiences, digital experiences, visual identity, emotional cues and brand storytelling.

The result was a richer understanding of how consumers imagine, shape and activate wellbeing in ways that feel relevant, motivating and authentic.

## 5. Further Methodological Details on the Generative AI Creativity Coach

There is a belief in innovation research that the average consumer cannot create or express themselves well enough to help build brand ideas. Our view is that a natural part of being human is to create. The IKEA Effect suggests that people place higher value on things they helped create. Kaufman and Beghetto's Four-C model of creativity is also relevant because it recognises that an average person may not create a finished brand idea, but they can imagine scenarios, react to possibilities, improve concepts and help shape ideas around their real human needs (2009).

In many traditional innovation methods, consumers are brought in to react, validate or choose, but rarely to help make. The methodological challenge therefore becomes "How do we help consumers surface emotions, symbols, associations and intentions that may still be half-formed?" Allowing respondents to use Gen AI to prompt and visualise within a survey could be a solution.

Sixty consumers participated in conversational AI-supported in-depth interviews with a difference. A focused, mobile-friendly version of the trained Whale AI platform was embedded directly into the interview experience. Respondents could generate, adapt and refine visuals, logos, taglines, campaign ideas and broader brand expressions. Rather than simply reacting to stimuli, they became active contributors to the creative process.

Participants were invited to behave as extended members of the Woolworths brand team. They were not simply asked whether they liked an idea. Instead, they were guided through a series of tasks and questions that gradually built a creative prompt.

The dialogue first established their personal context, including their current concerns, the role of wellbeing in their lives, their own assessment of wellbeing and their openness to new products and services. We then introduced the emerging wellbeing positioning, primarily to set the scene, before inviting respondents into a co-creation exercise supported by AI.

Respondents were asked to imagine how the wellbeing platform should come to life. They defined names, stories, messages, visual identities, colours, symbols, emotional cues and campaign ideas. AI then transformed these ideas into visual outputs that respondents could immediately evaluate and improve.

A typical flow looked as follows:

CB1: *Imagine you need to help Woolworths create this wellbeing brand. What should it be called?*

CB2: *Describe how this brand should look. Imagine you are in store. What would make it stand out? What colours should Woolworths use? What should the lettering look like?*

CB3: *What would the logo look like?*

AI Generation Moment: *Thanks for all that. We will now use AI to create an image based on your ideas.*

Improvement Prompt: *Great. What do you think? Sometimes AI gets you and sometimes it needs more help. How would you make this even better?*

The important methodological shift is that the visual was created by AI within the interview itself, while the respondent retained control of the creative process. Participants could critique, refine and improve the output through successive prompts.

Traditional concept testing typically treats the final answer as the data. In this approach, the creative journey itself became the data. We captured the original idea, the AI-generated visualisation, the improvement prompts, the emotional reactions and the reasoning behind each change. This allowed us to observe not only what people created, but how they created it.

The research therefore captured multiple layers of meaning simultaneously: what people said, what they felt, what they imagined, what they changed, what they rejected and which symbols and emotional cues they gravitated towards.

This was particularly valuable for wellbeing, since it holds rational, emotional, sensory and symbolic meanings and angles. Through prompting and visualisation, respondents were able to express meanings that may not have surfaced through words alone. They could show what felt warmer, safer, more premium, more human, more everyday or more motivating.

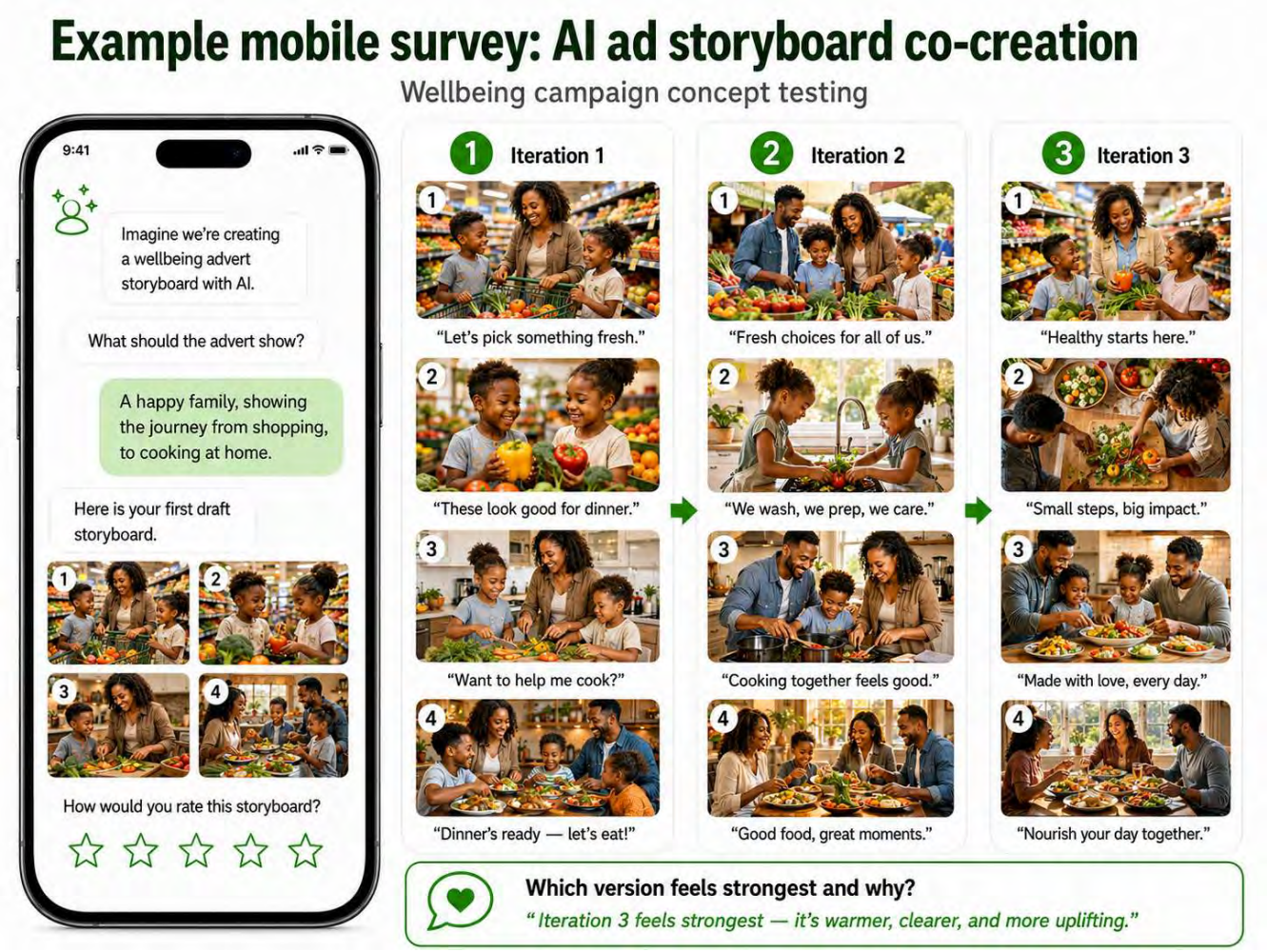



*Figure 5. Example mobile survey.*

In effect, consumers were helping define the brand's emotional palette, symbols, focus, perspective and afterimage, contributing to the composition of the brand itself.

This work also builds on earlier work we published for Esomar Africa 2026, where we used visualised storytelling techniques to help consumers express their relationships with brands through Whale AI. That work demonstrated that visual stories could reveal meanings that consumers often struggle to articulate directly. In this study, we extended the approach by embedding AI visualisation directly into the conversational interview itself.

The results were fascinating. We analysed not only the visuals respondents created, but also the prompts they wrote, the improvements they suggested and the emotional energy surrounding the process. It was particularly interesting to observe how quickly many participants became invested in their idea once AI visualised their first creation.

In this study, 98% of respondents said they liked or loved the experience, with the remaining 2% describing it as acceptable.

Conversational AI with embedded visualisation moves beyond opinion collection and becomes a creativity support system. By combining questioning, prompting and visualisation, consumers can express intention, emotion, symbolism, taste and practical ideas in ways that are difficult to capture through traditional interviewing alone.

The role of the researcher therefore begins to shift. Rather than simply asking questions, researchers increasingly become designers of creative environments in which meaning can emerge, evolve and be observed.

## 6. The Output: A Coherent Data-Driven Brand Architecture

The value of the process extended beyond identifying a winning platform. Throughout the project, a broader architecture began to emerge around the wellbeing positioning. This included emotional territories, supporting ideas, visual cues, symbolic codes, campaign language and possible activation pathways.

By combining historical research, future-oriented exploration, fresh consumer learning and collaborative AI tools, the team was able to work simultaneously at different levels of the challenge. Some activities focused on the bigger strategic question of where wellbeing may be heading and what role Woolworths could play within that future. Others focused on the practical realities of how the positioning could be expressed through products, services, communication and customer experience.

For Woolworths, the output was an umbrella wellbeing platform supported by a series of sub-positionings and activation territories. The process also generated a richer understanding of the emotional world surrounding wellbeing, the visual and symbolic cues consumers responded to, the moments where the platform felt most relevant and the language that made the idea feel authentic and motivating.

The result was a more connected body of knowledge than is often produced through a traditional sequence of research stages. Because the learning remained accessible within Whale AI throughout the project, ideas could be revisited, challenged and refined as new evidence emerged.

This proved particularly valuable when moving from strategy into execution. The project generated not only a positioning platform, but also a set of supporting territories, emotional narratives, campaign directions, visual expressions and activation ideas that could help guide future development.

The approach also helped reduce fragmentation. Different campaigns, innovations and customer experiences may serve different purposes, but they can still be connected through a shared strategic foundation. By developing the broader architecture alongside the positioning itself, the team was able to create a clearer line between the long-term brand direction and the practical ways it could be expressed.

The final output is therefore a data-informed wellbeing brand system that combines historical knowledge, future possibilities, consumer meaning and practical activation thinking within a single framework. Whale AI trained up on all this wellbeing insight can now be used to co-create new campaign ideas, rooted in deep customer understanding.

## 7. The Practical Impact: From a Client Perspective

The method moves quickly from abstract thinking to specific, visual, action-ready direction. Whale AI's visuals are easier to communicate to internal teams. The method helps answer "what does this mean in terms of symbols, cues, tone and execution?" It reduces the burden on teams to translate academic strategy into action, so Woolworths can move faster from understanding to doing.

It does not leave teams only with an abstract positioning statement but begins to show what the idea might look like, which symbols could carry it, what emotional world it creates and how it could be communicated internally. This accelerates the move from academic brand-building language to practical action direction.

In practical terms, the work helped translate a broad idea like wellbeing into more specific questions the business could act on. What would this mean for product innovation? What kind of rewards or support would feel meaningful? What educational role could Woolworths credibly play? What should the in-store and online experience signal? What colours, codes, characters or personas could make the positioning feel recognisable rather than generic?

This is one of the reasons visual co-creations are so useful. Visuals make strategy easier to discuss. They help teams see whether an idea feels too clinical, too generic, too soft, too premium, too playful, too far from Woolworths or exactly close enough to build from. They also make it easier for different internal teams to align around what the positioning could mean in execution.

Ultimately it puts the customer more at the centre, tempers internal bias, and aligns teams behind a common and clear direction. This enables faster execution around something that is more grounded and trusted. Reduction of uncertainty and faster aligned execution enables the business to capitalise on a new opportunity/platform in a safer way – something that is co-created with customers where needs and barriers have been better understood and accounted for.

## 8. What Changed for the Insight Process

This work signals a broader shift in how insight can function inside organisations.

*Insight became interactive:* Teams could work with knowledge rather than simply read about it.
*Insight became visual:* Strategy could be seen, tested and refined through images, symbols and cues.
*Insight became emotional:* Consumers could express feelings that are often difficult to capture through words alone.
*Insight became generative:* The system could help create territories, stories, prompts and expressions, not only analyse responses.

*Insight became continuous:* The knowledge base remained available to support future campaign thinking, innovation and brand development.
*Insight became an organising force:* The platform allowed for better alignment of teams, more efficient use of resources and a shift to a more holistic 360-view of the customer, evolving trends and institutional knowledge.

This is why the phrase "living brand system" is useful. The insight base can continue to support future campaign thinking, story development and visual exploration after the formal research project has ended.

This changes the role of the researcher. They become a designer of insight environments: deciding what knowledge enters the system, how people engage with it, and how ideas are interpreted through commercial, cultural and human judgement.

## 9. Key Learnings and Implications

### *Learning 1: Insight Must Become Easier to Use by Multidisciplinary Teams*

The value of research increases when it can be actively worked with by strategy, marketing and creative teams.

### *Learning 2: Visualisation Unlocks New Insight*

Prompting and visual expression help surface feelings, symbols and associations that consumers may struggle to put into words.

### *Learning 3: Consumers Can Co-Create Brand Meaning*

Consumers are not only useful as evaluators, but can help shape how ideas should look, feel and communicate.

### *Learning 4: AI is Most Powerful When It Connects Intelligence*

The real value is not automation alone, but the interaction between accumulated research, human judgement, consumer creativity and future signals.

### *Learning 5: Future-Ready Brands Need Architecture (Not Fragments)*

By creating an umbrella platform and connected territories upfront, brands can avoid reactive and fragmented campaign development.

The future of insight lies in becoming more deeply embedded in the way brands are built and evolved. Historical knowledge, future thinking, fresh consumer learning and creative exploration were brought together within a shared environment, allowing each stage to build on the last. Consumers contributed ideas as well as opinions, and researchers played a role in connecting the different sources of intelligence. This creates opportunities for insight teams to become more involved in brand strategy, innovation and creative development, while keeping decisions grounded in evidence and consumer understanding.

## 10. Conclusion: From Understanding People to Building with Them

This case points to a future in which insight is no longer a stage in the brand-building process, but part of the environment in which brands think, create and evolve.

For many years, research has helped organisations reduce uncertainty: to understand people, test ideas and validate direction. But AI-enabled insight environments open up a broader possibility. Insight can now become a living habitat for brand-building. A space where accumulated knowledge, emerging futures, consumer emotion and business imagination can meet.

This matters because brands are no longer built in neat, linear stages, but rather they are shaped continuously through products, services, campaigns, experiences, culture and conversation. Insight needs to live in that same continuous world: present while ideas are forming, not only after they are ready to be tested.

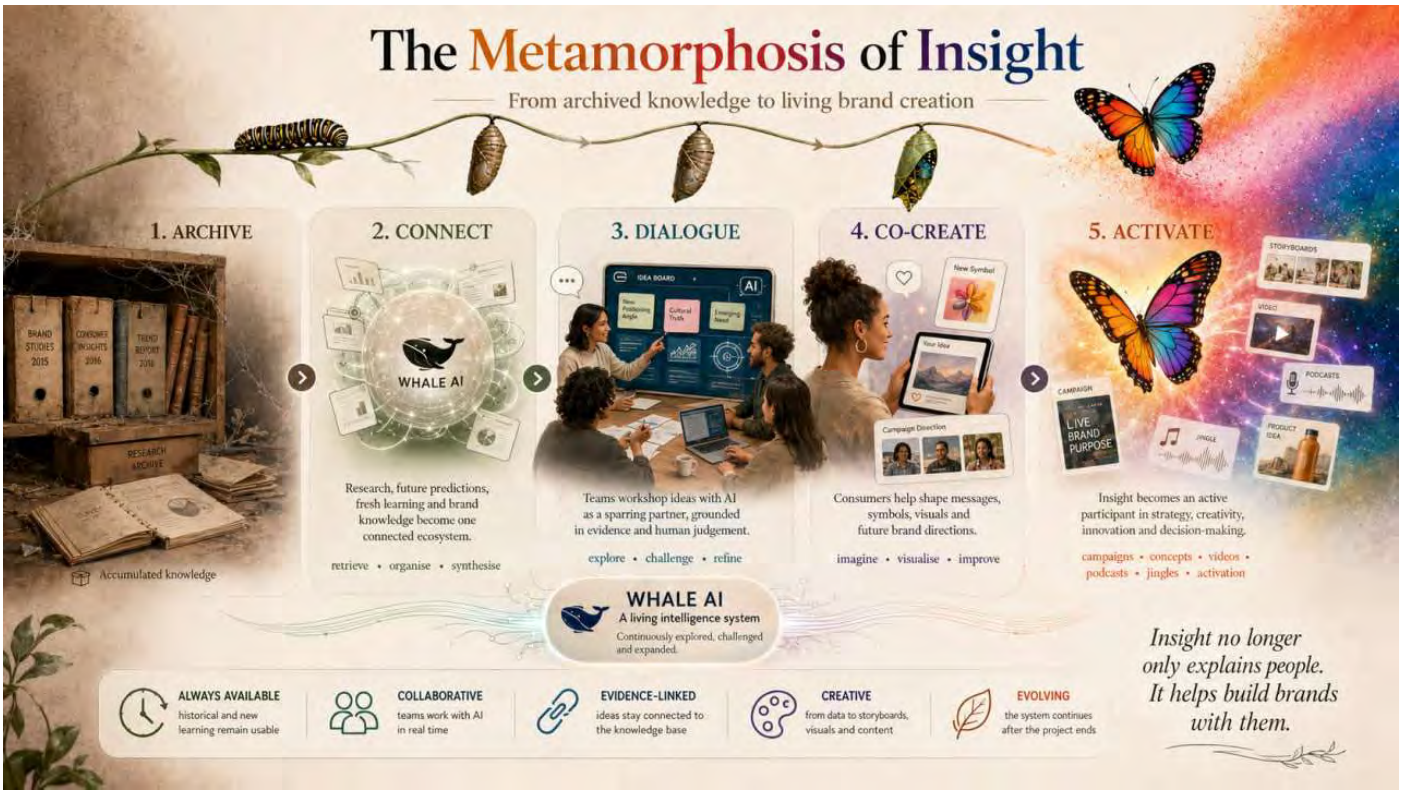



*Figure 6. The metamorphosis of insight.*

In this future, the value of insight evolves into helping create a more intelligent organisation: one that can return to previous learning, challenge assumptions, listen continuously to consumers, imagine future possibilities and translate human understanding into action with greater confidence and creativity.

The role of the researcher changes with it. They become the architects of these insight habitats, deciding what knowledge enters the system, how evidence is connected, where human judgement is needed, and how consumer meaning is protected as ideas move faster through the business.

AI will bring new methods and tools, but the real revolution is that research stops being something we write up at the end and starts becoming something brands can build on every day. Insight moves out of the report and becomes a living infrastructure. And consumers move from being people we study at a distance to people we build with, continuously.

Copyright © **ESOMAR** 2026


## References

Benartzi, S., Long, R., & Puntoni, S. (2026, June 1). *Companies are using AI for efficiency. They should use it to grow.* Harvard Business Review.

Belu, K., Kriegler, K., & Wright, D. L. M. (2026). *Evolve. Combine. Let go. Mapping brand love across generations.* Paper presented at the ESOMAR Africa Conference 2026.

Kaufman, J. C., & Beghetto, R. A. (2009). Beyond big and little: The four C model of creativity. *Review of General Psychology, 13*(1), 1–12. https://doi.org/10.1037/a0013688

McKinsey & Company. (2024, 16 January). *The trends defining the $1.8 trillion global wellness market in 2024.* https://www.mckinsey.com/industries/consumer-packaged-goods/our-insights/the-trends-defining-the-1-point-8-trillion-dollar-global-wellness-market-in-2024

McKinsey & Company. (2025, 29 May). *The $2 trillion global wellness market gets a millennial and Gen Z glow-up.* https://www.mckinsey.com/industries/consumer-packaged-goods/our-insights/future-of-wellness-trends

Norton, M. I., Mochon, D., & Ariely, D. (2012). The IKEA effect: When labor leads to love. *Journal of Consumer Psychology, 22*(3), 453–460. https://doi.org/10.1016/j.jcps.2011.08.002

Stremersch, S., Cabooter, E., Guitart, I. A., & Camacho, N. (2024). Customer insights for innovation: A framework and research agenda for marketing. *Journal of the Academy of Marketing Science.* https://doi.org/10.1007/s11747-024-01051-8

World Health Organization. (2022, 2 March). *COVID-19 pandemic triggers 25% increase in prevalence of anxiety and depression worldwide.* https://www.who.int/news/item/02-03-2022-covid-19-pandemic-triggers-25-increase-in-prevalence-of-anxiety-and-depression-worldwide

## About the Authors

Charlene Van Zyl, Customer & Loyalty Director, Woolworths, South Africa.
Karlien Kriegler, Co-Founder, Hello Ara, South Africa.
David Wright, Co-Founder, Hello Ara, South Africa.

esomar

# Closing the Protection Gap

## From Insight to Intervention in Preventing Online Child Sexual Exploitation

*Joy Ganvik*

*Anju Ahuja*

Esomar

Office address:

Burgemeester Stramanweg 105

1101 AA, Amsterdam

Phone: +31 20 664 21 41

Email: info@Esomar.org

Website: www.Esomar.org

Publication Date: September 2026

Esomar Publication Series Volume S424 Congress 2026

ISBN 92-831-0337-8

Copyright © **Esomar** 2026

Copyright © **Esomar** 2026

Congress 2026

# Closing the Protection Gap

## From Insight to Intervention in Preventing Online Child Sexual Exploitation

*Joy Ganvik*

*Anju Ahuja*

## Author's Note

This project demonstrates how AI-enabled research can be used responsibly in sensitive social-impact work when it is paired with strong human oversight, transparent methodology, careful interpretation and clear ethical boundaries. The role of AI was not to replace the researcher, but to help accelerate and strengthen synthesis and broaden strategic exploration while preserving the core responsibilities of research: rigour, context, respondent protection and evidence-based interpretation.

Given the breadth and complexity of the topic, this paper is intentionally structured to support multiple levels of engagement. Because of the importance of the topic, we have chosen to provide depth of findings and methods rather than economising on paper length. To that end, readers primarily interested in findings and implications may focus on the Executive Summary, Key Findings, Intervention Outputs and Stakeholder Implications sections, while those interested in methodological innovation and AI-assisted research workflows may explore the Methodology and Appendix sections in greater depth. Detailed examples, technical frameworks and intervention outputs are included throughout the paper and appendices to provide transparency, illustrate the applied methodologies and support practical interpretation of the findings.

## Executive Summary

### *Research Summary*

Children today participate in a wider range of digital environments online than ever before, including social media, gaming, messaging, streaming and connected digital environments. These digital technologies expose children to increasing online risks, including cyberbullying, exploitative interactions, scams, predatory grooming and child sexual abuse material (CSAM).

This research develops actionable intervention strategies and objectively explores the growing "protection gap" between parental perceptions of online safety preparedness and the realities of children's digital behaviours and risks. Findings are based on a national quantitative study conducted among US parents of children ages six to 17, designed to better understand parental concerns, safety behaviours, unmet needs and barriers to intervention. Beyond the substantive topic itself, the project also demonstrates how advanced research methodologies and AI-enabled approaches can help transform research from passive measurement into actionable intervention development. The study incorporates advanced quantitative methodologies including AI-powered adaptive probing, NLP-assisted verbatim analysis, latent class segmentation, MaxDiff prioritisation modelling with cluster analysis, driver analysis modelling and allocation-based trade-offs.

In addition, Ampliflyer (an enterprise AI-native strategy and intelligence platform) was used to accelerate the translation of research findings into intervention concepts, communication frameworks and actionable parent-support strategies. While Ampliflyer's platform is typically used by businesses to uncover market opportunities, develop disruptive products and achieve growth, we saw an opportunity to leverage this unique platform for this investigation. The findings reveal important disconnects between parental assumptions and children's actual online

experiences, highlight differing parent mindsets and protection approaches and identify opportunities for technology companies, educators, policymakers and child protection organisations to better support families. The challenge of protecting children is increased by the differing attitudes parents express as they examine tradeoffs between freedom of use and shielding children from possible harm. One size does not clearly fit all. Ultimately, this paper illustrates how market research can evolve beyond insight generation toward enabling practical societal interventions that address different perceived needs in the population.

## *Key Findings and Implications*

Parents widely recognise that children face meaningful online risks, but the findings reveal a substantial gap between awareness and effective protection in practice. While most parents report taking at least some steps to manage online safety, relatively few consistently use advanced monitoring tools, content filtering or layered protection strategies. Many parents instead rely heavily on conversations, basic rules or reactive intervention after issues arise. At the same time, parents report significant barriers to effective online safety management, including balancing privacy with monitoring, children resisting controls, uncertainty around which risks matter most and difficulty keeping pace with evolving technologies and platforms. Together, these findings point to a growing "protection gap" between concern, preparednes, and sustained protective action. The research also demonstrates that parents are not approaching online safety in a uniform way even relative to their perceived risks. Segmentation analysis identified four distinct parent mindsets, each reflecting different philosophies around perceived risks, monitoring, trust, communication, technology use and intervention. Some parents take highly proactive and technology-driven approaches focused on monitoring and control, while others prioritise selective intervention, relationship dynamics or maintaining balance between protection and independence. Importantly, these differences extend beyond simple concern levels, and influence how parents evaluate online risks, what tools they adopt, how frequently they communicate with children and which types of solutions they believe are most effective.

Across the study, parents consistently show the strongest interest in solutions that provide actionable protection and reduce the burden on parents themselves. Real-time alerts, parental controls, content filtering, privacy protection and AI-enabled monitoring capabilities emerged as some of the highest-priority solutions, while communication-only or education-only approaches were viewed as less effective when used in isolation. At the same time, usage of many protective solutions remains relatively low despite high perceived importance, reinforcing the broader implementation gap identified throughout the research. The findings suggest that effective online safety strategies will likely require more than awareness campaigns alone, instead combining practical tools, communication support, trust-building approaches and adaptive technologies that help parents translate concern into meaningful and sustainable action.

## *Project Contributors*

This project was developed through collaboration across research, strategic, analytical and AI-enabled innovation disciplines.

**SA Market Insights:**

- Joy Ganvik, research design, methodology leadership, analysis and paper development;
- Carlos Salazar-Velasquez, segmentation modelling and advanced analytics;
- Kathy Schmitt, project management and adaptive probing implementation;
- Alexandra Tuleja, strategic synthesis and intervention development;
- Praveen Yadav, project, field and data management;
- Pallavi Mittal, visual design and report development;
- Vinay Agrawal, programming, quality assurance and data processing.

**Ampliflyer:**

- Anju Ahuja, concept generation, scenario planning and intervention ideation;
- Justin Daab, AI platform development;
- John Higginson, AI platform scalability.

Copyright © **Esomar** 2026


# Introduction: Understanding the Protection Gap

## *The Expanding Digital Reality for Children*

Children today are growing up in a digital environment fundamentally different from previous generations. Social media, messaging applications, gaming ecosystems, livestreaming platforms, connected devices and AI-enabled technologies have become embedded within children's everyday lives. Digital platforms increasingly influence how children communicate, learn, socialise, entertain themselves and form relationships. At the same time, these digital ecosystems expose children to a growing range of online safety risks. These risks include cyberbullying, inappropriate content exposure, scams and fraud, exploitative interactions, grooming behaviours, privacy violations and child sexual abuse material (CSAM). As online behaviours continue to evolve rapidly, parents and caregivers are increasingly expected to serve as the first line of defense in protecting children online. Yet, many parents face significant challenges in understanding digital platforms, monitoring online interactions, balancing privacy and trust and determining which protective actions are most appropriate or effective.

## *Defining the "Protection Gap"*

This research centres around the concept of the "protection gap", defined as the disconnect between:

- Parent perceptions of online safety preparedness;
- Actual child digital behaviours and risks;
- Parent confidence versus effective action;
- Awareness of risks versus implementation of protections;
- Available safety resources versus practical usability and adoption.

This gap has important implications for families, educators, policymakers, technology companies and child-protection organisations.

## *The Need for Action-Oriented Research*

While awareness of online child safety concerns has grown substantially in recent years, much of the existing conversation remains focused on awareness rather than intervention. Research often measures concerns, attitudes or behaviours without providing practical pathways toward action. At the same time, organisations responsible for child safety frequently struggle to translate complex research findings into clear guidance, communication frameworks, intervention strategies or usable tools for parents. This project was designed not only to better understand the online safety ecosystem, but also to explore how advanced research methodologies and AI-enabled systems can accelerate the translation of insights into practical interventions.

## *Research Context and Relevance*

This work was developed in the context of growing global concern around online child exploitation and broader online safety challenges. Many organisations are exploring, assessing, monitoring and addressing the role of online methods in exploiting children to put protective measures in place and guide businesses and caregivers in how to better protect them. Some of the key factors that led us to identify the need for this research, which also influenced the research design, include:

- Research from Thorn, a nonprofit organisation focused on technology-driven child protection and prevention of online child sexual exploitation, has demonstrated that exploitative interactions increasingly occur within mainstream digital environments rather than isolated platforms (Thorn, 2025).
- Reports to NCMEC's CyberTipline increased substantially between 2021 and 2023, highlighting the growing scale and complexity of online child exploitation concerns. Although overall report volume declined in 2024, some high-risk categories continue to rise (NCMEC, 2024).

- UNICEF has emphasised the need to balance child protection with digital inclusion, privacy, participation and developmental opportunities (UNICEF, 2017).
- The Internet Watch Foundation has reported increasing volumes of online child sexual abuse material, including growing concerns around AI-generated exploitative imagery (IWF, 2024).
- The US Surgeon General has emphasised that protecting youth online requires coordinated action across families, technology companies, policymakers and researchers (US Surgeon General, 2023).
- Pew Research studies have shown that teens increasingly engage across multiple digital platforms simultaneously, creating more fragmented and difficult-to-monitor online experiences (Pew Research Center, 2024).
- Ofcom research has highlighted the increasingly blurred boundaries between gaming, messaging, entertainment and social interaction within children's digital lives (Ofcom, 2024).

## Research Thesis

This paper reasons that:

- Significant protection gaps exist between parent perceptions and children's digital realities;
- Parents require more practical, actionable support rather than awareness messaging alone;
- AI-driven research methodologies can:
  - provide deeper behavioural understanding (such as with adaptive probing) and,
  - can also help accelerate the transformation of insights into practical interventions (using platforms like Ampliflyer);
- Market research can play a more active role in supporting societal outcomes beyond measurement.

## Research Objectives

The project was designed around four primary objectives.

### *Explore Digital Behaviours and Children's Online Risk Exposure*

The first objective was to better understand how children engage with digital platforms and online ecosystems, including:

- Social media usage;
- Messaging and communication behaviours;
- Gaming interactions;
- Streaming and content consumption;
- Online relationship formation;
- Device access and supervision realities;
- Exposure to potentially risky interactions or content.

The study sought to identify areas where parental assumptions may diverge from actual behaviours and experiences.

### *Understand Parent Mindsets, Concerns and Protection Behaviours*

The second objective focused on understanding how parents perceive online risks and how those perceptions influence protective behaviours. Areas explored include:

- Perceived online risks;
- Safety priorities;
- Confidence in managing online safety;
- Monitoring and supervision approaches;

- Communication with children;
- Privacy and trust dynamics;
- Technology usage and safety tools;
- Emotional barriers and uncertainties.

### *Identify Unmet Needs and Opportunities for Intervention*

The third objective focused on uncovering gaps within the broader online safety ecosystem; specifically:

- Areas where parents feel unsupported;
- Gaps in education and communication;
- Confusing or under-utilised safety tools;
- Opportunities for clearer guidance;
- Intervention opportunities for organisations and stakeholders;
- Differences in needs across parent segments.

### *Demonstrate Advanced Methods and AI-Enabled Approaches*

The project was designed to demonstrate how advanced market research approaches, including AI-assisted adaptive probing, NLP-supported analysis and AI-enabled strategic intelligence systems, can support both:

- Deeper insight generation;
- Faster translation of research findings into practical interventions.

The project intentionally explored how research can move beyond passive reporting toward enabling actionable societal impact.

## Research Methodology

### *Research Design Overview*

The study consisted of a national quantitative online survey conducted among US parents. The target audience included parents or guardians of children ages six to 17 who have children living with them at least part-time and play a role in approving or managing children's online activities. See Appendix A for more details.

| | |
|---|---|
| **Geography** | United States |
| **Methodology** | Online quantitative survey |
| **Sample size** | n=1,014 |
| **Panel provider** | InnovateMR |
| **Fieldwork** | May 2026 |
| **Representation** | Representative demographically of US parents with children ages 6-17; no weighting |

*Table 1. Sample specifications.*

### *Survey Design and Content Areas*

The questionnaire was designed to examine multiple dimensions of the online safety ecosystem. Key content areas included:

- Connected device and online usage (adults and children) across a range of platforms;
- Parent awareness and concern levels;
- Monitoring and supervision approaches, including the use of online safety tools;

- Family trust and privacy dynamics;
- Experiences with online incidents or risks;
- Evaluation of safety features and tools;
- Trade-offs between privacy, independence and protection;
- Parent perceptions of stakeholder responsibility.

### *Data Quality Procedures*

To ensure data quality, multiple procedures were implemented (see Table 2).

| | |
|---|---|
| • Geolocation check | • Speeding checks |
| • Bot detection: Captcha, OE analysis, Page time variables | • Open-end quality review |
| • Quota balancing on incoming sample demographics | • Duplicate detection |
| • Positional straightliner check on MaxDiff exercise | • Logic validation |
| • Digital fingerprinting | |

*Table 2*

## Key Analytic Methodologies

A core component of this project involved the use of advanced quantitative and AI-enabled methodologies to generate deeper behavioural understandings and actionable insights. Additionally, statistical significance testing was conducted at the 95% confidence level ($p < .05$), and statistically significant differences are referenced throughout the paper.

### *Generative AI-Powered Adaptive Probing*

To better understand the contextual, emotional and behavioural dynamics surrounding children's online safety, the survey incorporated AI-powered adaptive probing techniques. Rather than relying solely on static open-ended responses, AI-generated follow-up questions were dynamically tailored to individual participant responses in real time. This approach allowed the survey to expand on short attitudinal statements and uncover richer insight related to parent concerns, trust dynamics, communication approaches, practical implementation challenges and perceptions of online risk. This approach enabled greater depth of qualitative understanding at scale, clarification of ambiguous responses, exploration of emotional nuance, identification of underlying motivations and barriers and improved richness of open-ended data.

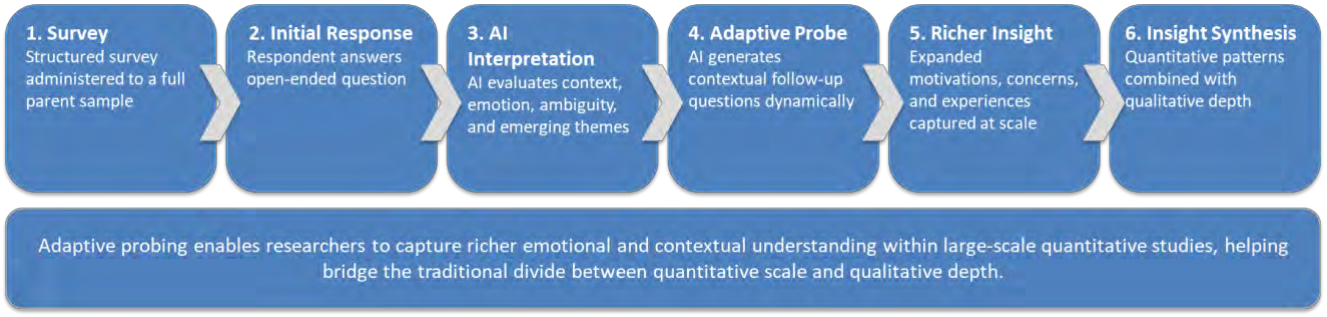



*Figure 1. AI-powered adaptive probing process.*

Across the sample, adaptive probing illustrated how AI-enabled probing may help bridge the gap between quantitative scale and qualitative depth within survey research by generating richer contextual understanding, while maintaining the efficiency and scalability of large-sample quantitative methodologies. See Appendix E for more information on AI-enabled probing flows.

### *NLP-Assisted Verbatim Analysis*

Natural language processing (NLP) approaches were used to support analysis of open-ended responses. Analysis dimensions included theme extraction, topic clustering, emotional tone analysis, sentiment analysis, risk-related language patterns and parent anxiety and confidence indicators. This process helped identify recurring themes across thousands of open-ended responses more efficiently and deeply than could have been done using manual coding. See Appendix F for more information.

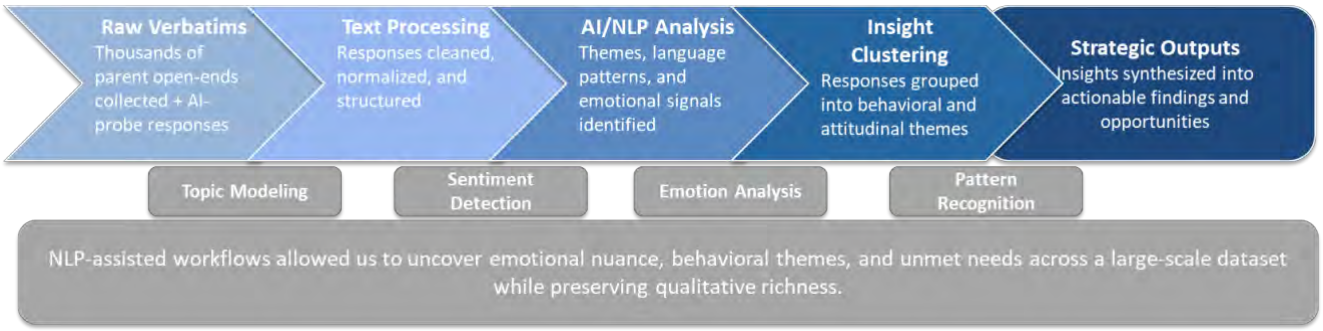



*Figure 2. NLP assisted verbatim analysis process.*

### *Latent Class Segmentation Modeling*

Latent class segmentation analysis was conducted to identify distinct parent mindsets, behaviours and approaches related to children's online safety management. See Appenix B for more more details on the segmentation methodology. Rather than segmenting parents based on demographics or overall concern levels, the model incorporated attitudinal, behavioural, communication, technology-confidence and decision-priority variables to identify naturally occurring parent profiles with differing approaches to online safety, monitoring, trust, intervention and risk management. The final segmentation solution was selected based on a combination of statistical fit, interpretability, segment stability and practical differentiation across key online safety dimensions. Variables included in the model represented several core thematic areas, including:

- Technical confidence and safety tool usage;
- Online safety mindset and risk orientation;
- Parent-child communication and trust dynamics;
- Perceived and experienced online exploitation risks;
- Decision priorities when evaluating children's online activities and platforms.

The resulting segments revealed meaningful differences in how parents define preparedness, balance protection with independence, implement monitoring strategies, communicate with children and evaluate online risks and safety solutions.

### *MaxDiff Prioritisation Modelling*

MaxDiff (maximum difference scaling) analysis was conducted to identify the relative importance of online safety solutions, features and intervention approaches from the parent perspective. See Appendix C for more details on this methodology. Respondents were shown sets of potential online safety solutions and asked to identify the most and least important options within each set. This approach forced respondents to make trade-offs between competing priorities, helping reduce scale bias and providing stronger differentiation than standard rating scales. The MaxDiff exercise included a range of online safety solutions, and the final list of solutions included in the MaxDiff exercise was particularly challenging for this research given the sensitivity and importance of the topic. Initially, the exercise was planned with a large set of individual features. However, at that level, trade-offs proved too difficult for respondents (e.g., identifying adults posing as children versus recognizing grooming behaviours versus recognising explicit images). Success was achieved by grouping sets of features into differentiated solution

categories. Hierarchical Bayes estimation was used to generate individual-level utility scores, which were then normalised to support relative importance comparisons across solutions. The analysis was designed not only to identify the highest-priority solutions overall, but also to better understand the underlying philosophies parents apply when evaluating online safety interventions.

To further explore differing parent solution preferences, cluster analysis was conducted using respondent-level MaxDiff utility scores. This analysis identified distinct groups of parents with differing approaches to online safety management, intervention priorities and technology preferences. Multiple cluster solutions were evaluated based on statistical separation, interpretability, stability and practical differentiation across solution priorities. The final solution identified three distinct solution-orientation clusters. The resulting framework provided a more nuanced understanding of how parents evaluate online safety solutions beyond overall importance scores alone, helping identify differing expectations regarding the role of parents, technology platforms, automation and AI-enabled systems in protecting children online.

### *Driver Analysis by Segment*

To better understand what most strongly influences parents' sense of preparedness to manage online safety risks, driver analysis was conducted using Johnson's *Relative Weights* methodology. The dependent variable measured perceived preparedness to prevent or respond to serious online safety risks, while independent variables included online safety mindset, technical confidence, use of safety tools, technology experience and communication approach. The strength of this method was especially apparent when assessing the drivers by segment, which resulted in identifying key levers to build parent confidence within each of the parent segments. See Appendix D for more details on the driver analysis.

## AI-Enabled Intervention Development

### *Moving Beyond Insight Generation*

A distinguishing aspect of this project was the intentional focus on translating research insights into actionable intervention frameworks. Market research initiatives often end with reporting findings and recommendations. However, organisations frequently struggle to turn insights into practical tools, communication strategies, educational approaches or behavioural interventions. This project explored how AI-enabled systems can help accelerate the progression from insight generation toward intervention development and implementation strategies—a roadmap for success including near-term and longer-term actions for stakeholders.

### *Use of the Ampliflyer Platform*

Ampliflyer's proprietary AI platform was used as a multi-agent strategic intelligence, insight synthesis, concept-development, communication planning and strategic analysis system. The AI platform was used to:

- Simulate stakeholder and subject-matter-expert ideator perspectives;
- Accelerate strategic ideation, using codified design-thinking methodologies;
- Produce novel, disruptive solutions with impact assessments;
- Generate communication concepts targeted to stakeholders;
- Develop supporting parental frameworks;
- Create intervention roadmaps.

The platform was not used to replace researchers or automate conclusions. Ampliflyer is a human-at-the-helm platform, deploying agents under the direction of researchers, designed to augment intelligence and accelerate discernment by subject matter experts. It serves as a collaborative acceleration system to help researchers and organisations more rapidly operationalise findings. Ampliflyer was designed to increase human agency by accessing a wide swath of intelligence, processing vast arrays of insights, performing extensive time-consuming strategic analysis using codified frameworks and best practices, conducting a rigorous design-thinking process to develop solution prototypes, simulate insights based on these prototype concepts and, finally, spinning up a virtual marketing team to develop marketing, PR, branding and communication strategies.

Copyright © **Esomar** 2026


## Implications of AI-Enabled Insight Operationalisation

The use of AI-enabled strategic platforms may represent an important evolution in market research.

### *Implications for Research Practice*

One of the more significant implications of this study extends beyond the substantive findings themselves and into how research can be conducted and applied. The project demonstrated how AI-enabled methodologies can support multiple stages of the research lifecycle, from research design through insight generation to concept development and implementation. Rather than functioning solely as tools for analysis, AI-enabled systems may increasingly help researchers:

- Generate and refine research hypotheses;
- Explore alternative questionnaire structures and probing approaches;
- Accelerate analysis and synthesis of complex datasets;
- Gather and interpret related data that resides outside of the dataset;
- Identify patterns, tensions and opportunities across multiple data sources;
- Develop strategic frameworks that connect findings to decision-making;
- Create tailored storytelling and communication approaches for different stakeholder audiences;
- Support futurecasting and scenario planning exercises.

Importantly, the findings suggest that AI is most valuable when used to augment researcher expertise. Human judgment remained essential throughout the project for methodology design, interpretation, validation and prioritisation of insights.

### *Implications for Organisations and Stakeholders*

The study also illustrates how AI-enabled insight operationalisation may help organisations migrate more effectively from understanding problems to implementing solutions. Frequently, research projects conclude with findings and recommendations, while organisations and stakeholders are then responsible for determining how best to translate those insights into practical tools, communication strategies, educational approaches or interventions. In this project, AI-enabled strategic modelling tools were used to explore an array of potential interventions, communication approaches, educational resources and implementation pathways directly informed by the research findings. Potential applications include:

- Creating intervention roadmaps tailored to specific parent segments;
- Identifying opportunities for new products, services or safety features;
- Developing targeted parent education and awareness initiatives;
- Building communication strategies designed for specific stakeholder groups;
- Facilitating stakeholder alignment around complex societal challenges;
- Supporting long-term planning through scenario development and future-oriented analysis.

The ability to efficiently tailor recommendations, communications and intervention approaches to the needs, priorities and characteristics of different audiences may be particularly valuable for complex issues such as child online safety, where effective solutions often require coordinated engagement across children, parents, educators, policymakers, technology providers and child-protection organisations simultaneously.

## Key Findings

Copyright © **Esomar** 2026


### *Overall Findings and the Online Safety "Protection Gap"*

The findings reveal a digital environment in which children are engaging across a wide range of platforms, devices and online ecosystems at increasingly early ages, while many parents are still navigating how to effectively manage online safety risks. Although most parents recognise meaningful online threats and report taking at least some protective actions, the research consistently identified a "protection gap" between awareness of risk, confidence in protection and consistent, effective implementation of safety practices and tools.

### *Children Are Embedded in Complex Digital Ecosystems*

Children's online activity spans smartphones, gaming platforms, messaging applications, social media, AI tools, streaming environments and anonymous interaction platforms. The most commonly used devices include smartphones (75%), gaming devices (74%), smart TVs/streaming devices (74%) and tablets (64%). Importantly, children begin using many connected devices at relatively young ages. Tablets are often introduced before age eight (and 3/10 at age five or earlier), while smartphones and wearables increasingly appear during late elementary and middle school years. This creates an environment where parents are expected to manage multiple devices, multiple platforms, evolving risks and highly dynamic digital behaviours simultaneously.

### *Parents Recognise Significant Online Risks*

Online safety concerns are widespread: one-third of parents perceive their children as facing high or very high levels of online risk exposure, while nearly half perceive at least moderate levels of risk. Parents consistently expressed concern about online predators and grooming, inappropriate or explicit content, scams and phishing, cyberbullying, harmful communities, excessive usage, misleading information and stranger interactions. Risk perceptions extended across social media, messaging apps, gaming ecosystems, anonymous chat applications, AI-powered tools and betting/prediction platforms. Notably, platforms involving anonymous interaction, open communication, stranger access or relationship-based interactions were frequently perceived as especially risky.

### *Most Parents Are Trying, But Many Struggle*

The findings show that most parents are actively attempting to manage online safety: nearly all use at least some type of safety tool, most report implementing rules or controls and many engage in regular conversations with children about online behaviour. Common actions include discussing online safety, requiring approval for app downloads, monitoring app or browsing activity, using parental controls, setting screen-time rules and restricting certain features or platforms. At the same time, about one-third regularly monitor activity, about one-third use parental control software and one-in-10 parents report taking no specific actions at all. This suggests that awareness of risk does not always translate into sustained protective behaviour.

### *The Protection Gap Emerges Across Multiple Dimensions*

One of the strongest themes across the findings is the existence of a "protection gap" between concern and practical implementation. This gap appears in several forms.

1. Risk awareness versus action: Although most parents acknowledge meaningful online risks, many rely primarily on conversations rather than layered protection strategies, and relatively few consistently use advanced monitoring or filtering tools.
2. Confidence versus practical capability: While 72% say they can independently set up parental controls, about one-quarter still require help, and roughly one-third report only basic technical capability or reliance on others for technology support. Parents commonly turn to online searches, YouTube tutorials, spouses and even their own children for technology help. This highlights a practical implementation challenge: many parents may conceptually support online safety measures, but lack the confidence, time or technical fluency to consistently action them.
3. Desire for monitoring versus privacy concerns: The single largest challenge parents identified is balancing children's privacy with monitoring. Additional barriers include children resisting controls, uncertainty about which risks matter most, lack of effective tools,

time constraints and disagreement among caregivers. This suggests many parents are struggling not only with what tools to use, but also with how much oversight is appropriate.

4. Communication confidence versus actual communication dynamics: While 76% believe their children would probably or definitely come to them with online concerns, the broader communication findings suggest a more nuanced reality. Only about one-third report setting clear rules, many primarily step in only when issues arise and roughly one-in-10 do not typically initiate online safety conversations.

This indicates that some parents may overestimate communication openness, visibility into children's experiences or children's willingness to disclose concerning events.

### *Parents Want Balance, Not Total Surveillance*

The findings suggest that most parents are not seeking unrestricted surveillance. Instead, many appear to be searching for balanced oversight, practical guidance, age-appropriate controls and tools that preserve trust and independence while still reducing risk. This tension appears repeatedly throughout the findings: parents want visibility, but also want to maintain trust, autonomy and healthy family dynamics.

### *AI and Emerging Platforms Add New Complexity*

The research also highlights the growing complexity introduced by AI-powered tools, anonymous communication apps, algorithm-driven content systems and rapidly evolving digital ecosystems. Many parents are still unfamiliar with newer platforms and AI tools, yet several of these environments were perceived as carrying moderate-to-high levels of risk. This suggests a widening challenge where children's digital behaviours may evolve faster than parent awareness and preparedness.

### *Strategic Implications*

Taken together, the findings suggest that effective online safety strategies must go beyond simply increasing awareness. Parents may need:

- Simpler and easier-to-use tools;
- Clearer prioritisation guidance;
- More actionable support;
- Stronger communication frameworks;
- Adaptive systems that help translate concern into meaningful protective action.

The findings also suggest that online safety is no longer simply a technology problem. It is increasingly a communication challenge, a behavioural challenge, a trust challenge and a digital parenting challenge requiring coordinated support across families, platforms, educators, policymakers and technology providers.

Copyright © **Esomar** 2026



| Sexual Exploitation & Sexual Content (52%) | Harassment, Toxicity & Hateful Content (33%) | Harmful Content Exposure (18%) | Deception & Financial Exploitation (17%) | Privacy, Device Security & Surveillance (14%) |
|---|---|---|---|---|
| Stranger danger in digital spaces remains parents' primary fear, including worry about both predatory behavior and inappropriate material exposure. The prevalence of these concerns unprompted demonstrates how deeply the threat of sexual exploitation has penetrated parental consciousness about online safety. | Concern about *cyberbullying and online harassment* specifically is almost as prominent as predator concerns and highlights that peer-to-peer harm is a major top-of-mind issue. This aligns with the survey goal of understanding what parents are most focused on for their families, as bullying represents a more common, everyday risk compared to predatory behavior. | The relatively lower prominence of misinformation compared to sexual or bullying content suggests parents may be more focused on immediate safety threats than on information quality, though this still represents a meaningful aspect of parental awareness. | This indicates that a significant minority of parents are thinking about their children's vulnerability to financial manipulation and deceptive practices online. The presence of this theme demonstrates that parents' risk awareness extends beyond content and harassment to include economic and identity-based threats. | Parents are also concerned about data protection and technical security. Privacy and data security, hacking, viruses, and device security and sharing personal information and risky disclosures were also noted. |

*Base: Total - 1014*
***T7.*** *What type of risks do you think your child(ren) could be exposed to when online?*

*Figure 3. Types of children's risk exposure.*

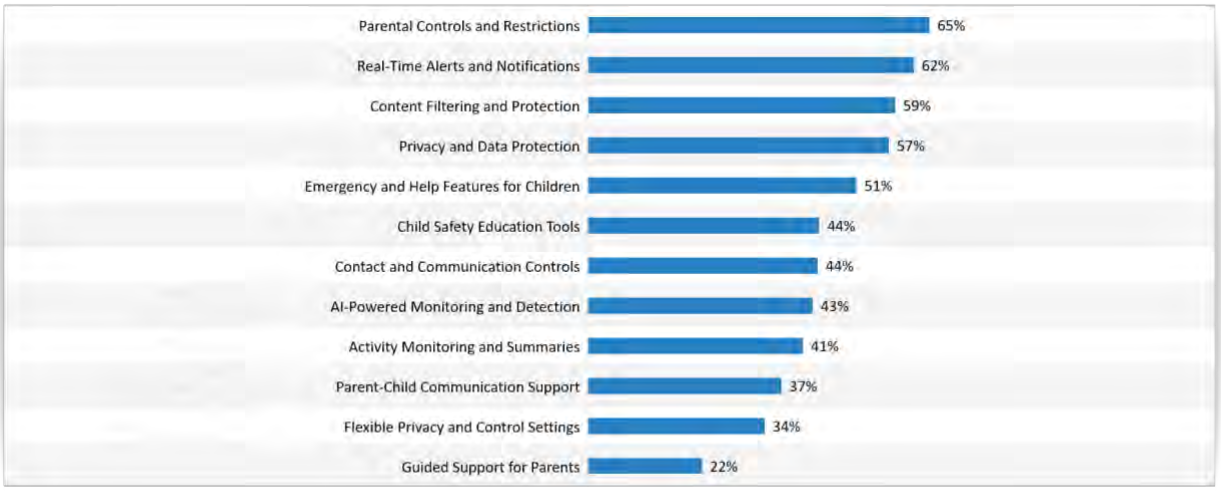
| Sexual Exploitation & Sexual Content (52%) | Harassment, Toxicity & Hateful Content (33%) | Harmful Content Exposure (18%) | Deception & Financial Exploitation (17%) | Privacy, Device Security & Surveillance (14%) |
|---|---|---|---|---|
| • "The risk of being befriended by a child predators"<br>• "Pedophiles, sex trafficking, and adults pretending to be their age"<br>• "All different kinds like people trying to kidnap them or make them get evolved in sexual content without them realizing it until its too late"<br>• "One major risk is exposure to inappropriate content, including violence, sexual material, or harmful misinformation that isn't age-appropriate"<br>• "Sexual harassment indecent exposure"<br>• "Adult content and predators"<br>• "Sexual harassment exploitation" | • "Children can be exposed to cyberbullying inappropriate content, and online scams"<br>• "May face risks like cyberbullying."<br>• "Online bullying and hate speech"<br>• "They are easily exposed to harmful online content, such as violence and vulgarity, which distorts their values."<br>• "I think my child could be exposed to constant risk, such as violent, hateful material, as well as they can harassment from strangers, and also their identity or personal details."<br>• "Can be exposed to explicit content, vulgarity, trolls and predators." | • "Access to inappropriate or violent content"<br>• "Age inappropriate content"<br>• "Exposure to harmful content promoting gun violence, self-harm, and suicide on popular platforms such as TikTok and Roblox can trigger mental health issues."<br>• "Watching inappropriate things, unhealthy habits, risky ideas and more" | • "Scams and predators"<br>• "Scams, bad people"<br>• "Catfishing and scammed out of money."<br>• "Phishing"<br>• "risk of being scammed and meeting of all kinds of bad friends"<br>• "Falling for online scams involving fake sweepstakes and fake order schemes leads to irrational spending."<br>• "Clicking on illegal links or downloading pirated game plugins can lead to the theft of personal account information and household payment details, resulting in financial loss." | • "Data and security risks online phishing risks online predator risks"<br>• "Theft of personal information or fraud"<br>• "Hacking and privacy issues"<br>• "Security about personal information"<br>• "Breach of private information and address"<br>• "Sharing personal information without understanding the risks" |

*Base: Total - 1014*
***T7.*** *What type of risks do you think your child(ren) could be exposed to when online?*

*Figure 4. Verbatim responses related to types of children's risk exposure.*

## *Solution Prioritisation and MaxDiff Cluster Analysis*

The MaxDiff analysis revealed that parents strongly prioritise online safety solutions that provide direct protection, monitoring and intervention capabilities. While parents expressed interest in a broad range of online safety approaches, the strongest demand consistently centered on solutions that actively help parents manage or reduce online risk exposure (see Figure 5).




*Base: Total – 1014*
***V1.*** *When it comes to solutions to help keep your child(ren) safe online… Which are you MOST INTERESTED in?…. Which are you LEAST INTERESTED in?*

*Figure 5. Prioritisation of solutions: probability of choice.*

## *Core Solution Priorities*

The highest-priority solutions across parents are parental controls and restrictions, real-time alerts and notifications, content filtering and protection,and privacy and data protection. These findings suggest that parents are primarily seeking visibility, control, active intervention and automated protection mechanisms rather than passive or informational support alone. Solutions that rank substantially lower overall are focused primarily on communication support, parent guidance and educational tools.

## *Protection-Oriented Solutions Dominate*

The strongest-performing solutions all centre around the theme of active protection. Parents show the greatest interest in tools that block harmful content, monitor activity, identify risks, limit unsafe interactions or notify parents when concerns arise. This indicates that many parents view online safety primarily as a risk-management and intervention challenge, rather than solely an educational or communication issue.

## *Opportunity Gaps in the Market*

The analysis identified several important gaps between:

- Perceived importance;
- Perceived effectiveness;
- Current usage.

Most notably, real-time alerts and notifications and emergency and help features for children have high interest levels, but current usage rates are substantially lower. This suggests meaningful unmet demand for proactive warning systems, escalation support and child-accessible emergency tools. Similarly, only 39% currently use parental controls, 32% use content filtering and 31% use privacy and/or data protection tools, while 14% report using none of the assessed solutions at all. These findings reinforce the existence of a broader protection implementation gap between parent concern and actual protective behaviour.

## *Communication and Education Solutions Face Credibility Challenges*

Several of the solutions with the lowest credibility focused on communication support, parent guidance and educational approaches. Specifically, parent-child communication support, guided support for parents and child safety education tools were among the least likely to be viewed as

highly effective. This does not necessarily mean parents oppose these solutions. Rather, it suggests many parents may perceive educational approaches alone as insufficient relative to the scale and complexity of online risks.

### *AI-Based Protection Has Meaningful Consumer Acceptance*

One of the more notable findings is that AI-powered monitoring and detection performed substantially better than many softer communication-based solutions. Parents appear increasingly open to AI-enabled systems that proactively identify threats, detect concerning patterns and surface actionable alerts. This is particularly important given the increasing scale, speed and complexity of children's digital environments, where manual parent oversight alone may not feel sufficient.

### *Child Age Has Limited Impact on Priorities*

Importantly, solution priorities are remarkably consistent across age groups, with similar prioritisation patterns across both decision factors and preferred safety solutions. This suggests that concerns about online safety, demand for intervention tools, and desire for protection extend consistently across a wide child age range.

### *Solutions Cluster into Three Distinct Solution Mindsets*

Latent class analysis of the MaxDiff results revealed three distinct solution-orientation clusters:

1. Hands-on control (23%): This group favours direct, parent-led management approaches, suggesting parents should remain highly involved in controlling and managing their child's online environment. This cluster showed extremely high prioritisation of parental controls, alerts, filtering and communication restrictions, while showing almost no interest in AI-powered monitoring or parent-child communication support.
2. Balanced safety (53%): The largest cluster favours a more moderate, balanced and practical approach that suggests online safety is important, but solutions should support healthy family dynamics and avoid becoming overly intrusive: protection, communication, education and family usability. Unlike the other clusters, this group does not have extremely polarised preferences. Instead, they showed moderate interest across a broad range of solutions. Importantly, they were the only cluster to rank child safety education tools (#1) and parent-child communication support (#2) as top priorities. This indicates many parents still value communication and education, but prefer these approaches when positioned as part of a balanced safety ecosystem rather than as standalone interventions.
3. Automated protection (24%): This group shows extremely high interest in AI-powered monitoring, real-time alerts, automated content filtering and detection-based approaches, while placing very low value on communication-oriented solutions. This leads to the idea that technology itself should play an active role in identifying and preventing online risks.

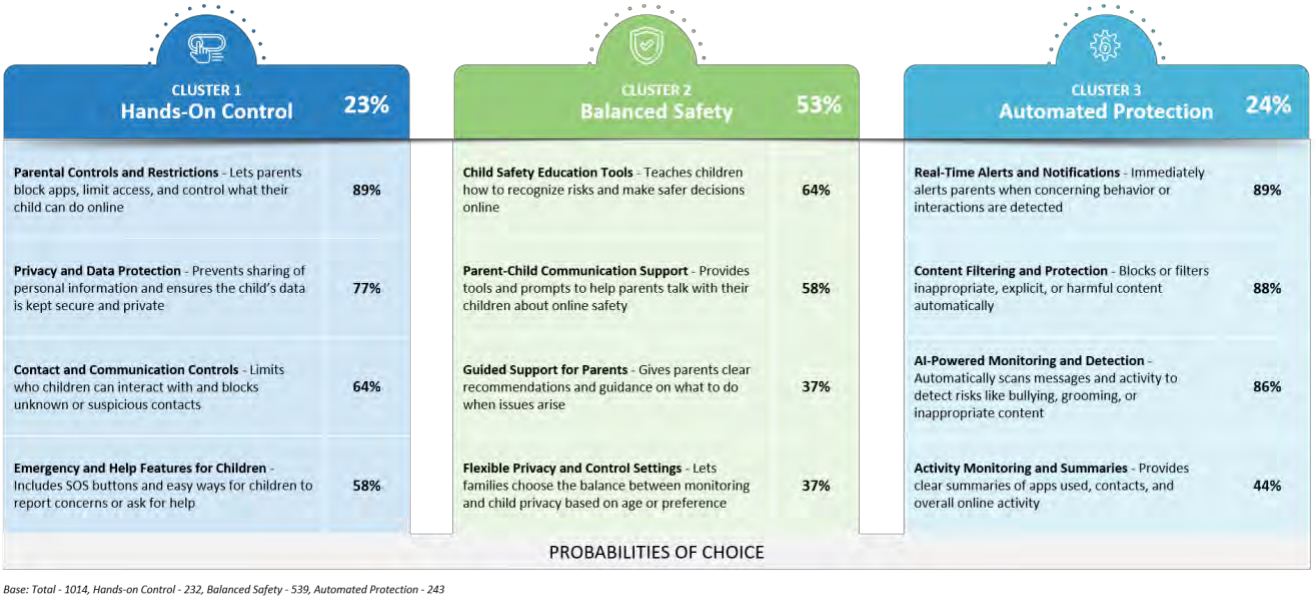



*Figure 6. Probabilities of choice.*

## *Strategic Implications*

The MaxDiff findings suggest several important implications for organisations developing online safety products and interventions:

- Parents strongly prefer actionable protection tools over passive guidance alone;
- Solutions that combine automation with parent visibility may have particularly strong market potential;
- Communication and education are important, but may need to be integrated into broader protection ecosystems rather than positioned independently;
- There is substantial opportunity to close the implementation gap between perceived importance and actual solution usage;
- AI-enabled monitoring and detection systems appear increasingly acceptable to many parents when framed around protection, alerts and risk prevention rather than surveillance alone.

## *Segmentation: Distinct Parent Segments Require Different Solutions*

The latent class segmentation revealed that parents approach children's online safety through distinctly different mindsets, behaviours and management styles, rather than through simple differences in overall concern levels. Five core dimensions drove segmentation classification (see Figure 7). Four distinct parent segments emerged:

1. Pragmatic Balancers (41%): The largest segment takes a moderate and practical approach to online safety, balancing protection with allowing children normal online experiences. They are less technically confident, less proactive and less likely to use advanced monitoring tools than other segments. They are also more likely to trust their children to manage themselves and to only step in when issues arise.
2. Vigilant Managers (12%): This segment is highly proactive, technically sophisticated and deeply focused on monitoring and control. They actively use and customise multiple safety tools, and place particularly high importance on parental controls and monitoring capabilities when evaluating solutions.
3. Social Risk Navigators (11%): These parents are especially focused on who their children interact with online and view online safety primarily through the lens of interpersonal and interaction-based risks. They tend to rely on selective intervention rather than intensive ongoing monitoring.
4. Engaged Protectors (37%): This segment combines high technical confidence with strong communication and ongoing engagement with children. Compared with other segments, they are most likely to foster open conversations, maintain regular communication about online safety and actively implement a broad range of safety actions and tools.

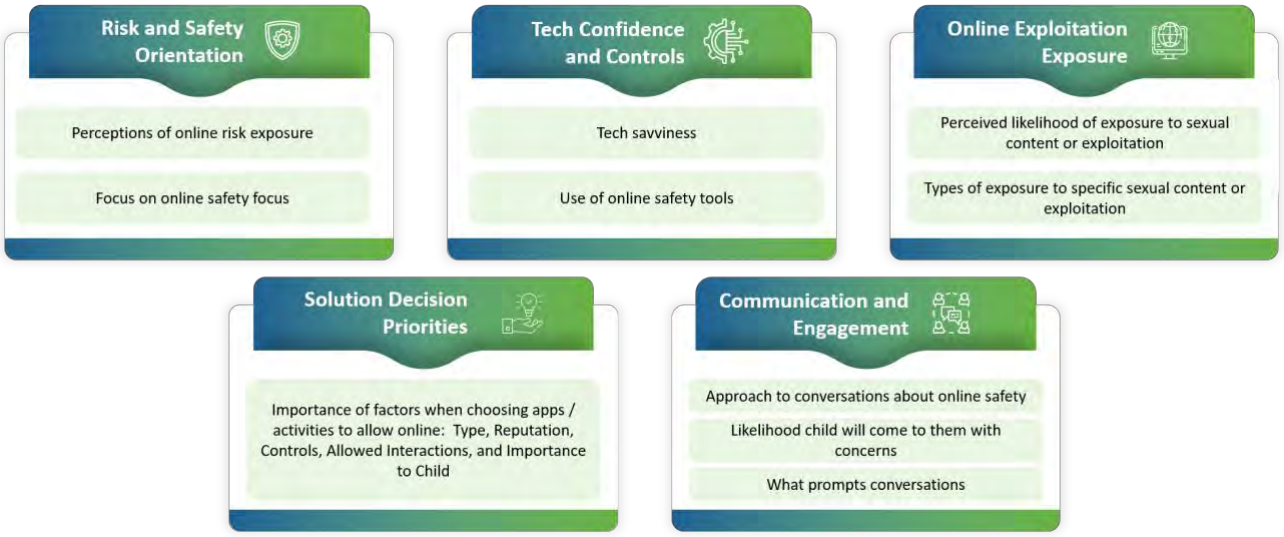



*Figure 7. Segmentation dimensions.*

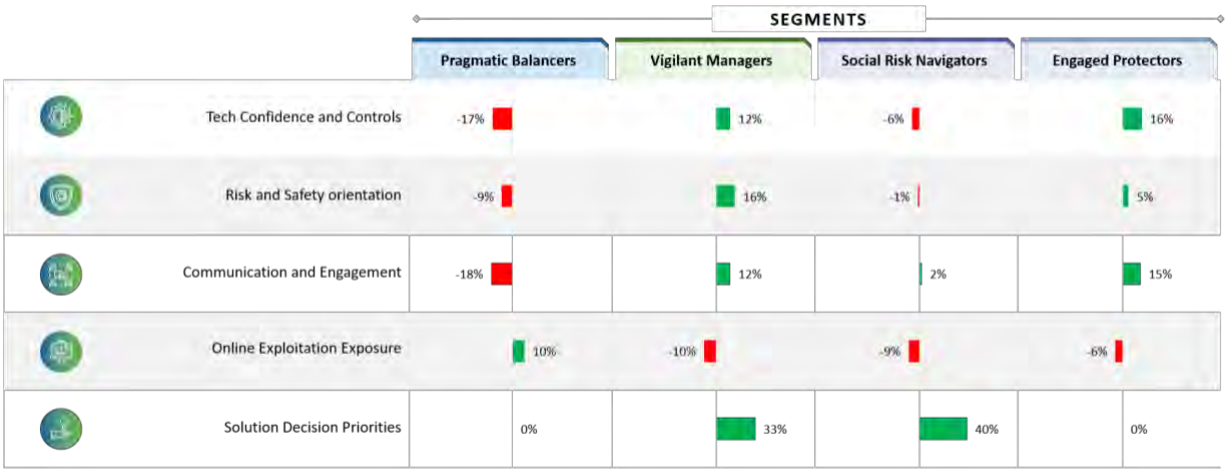



*Figure 8. Segmentation dimensions: variance from average.*

## Key Cross-Segment Insights

### *Online Safety Is Not a One-Size-Fits-All Problem*

Parents differ substantially in how they perceive online risks, how actively they intervene, how much they trust their children, how often they communicate and what types of solutions they value most. Online safety education, tools and interventions will likely be more effective when tailored to different parent mindsets and management approaches rather than relying on universal messaging.

### *Technical Confidence Strongly Influences Safety Management*

Engaged Protectors and Vigilant Managers report the highest levels of technology confidence, and are also the most likely to customise safety tools, monitor online activity, implement parental controls and take proactive safety actions. In contrast, Pragmatic Balancers are significantly more likely to rely on others for technology support, avoid advanced tools and use only basic controls or no tools at all. This highlights a meaningful capability gap that may limit some parents' ability to translate concern into action.

### *The Protection Gap*

One of the clearest findings across the segmentation is the existence of the "protection gap". Specifically, many parents acknowledge meaningful online risks, but their behaviours, monitoring practices, communication patterns and solution usage do not always align with the level of exposure they believe children face. Several findings support this gap:

- Pragmatic Balancers are among the least proactive segments despite reporting comparable perceptions of online risk exposure to other groups;
- Only 57% of Pragmatic Balancers believe their children would probably or definitely come to them with concerning online experiences, compared with over 90% among Engaged Protectors and Vigilant Managers;
- Pragmatic Balancers are also the least likely to use parental control tools, monitor usage or regularly discuss online safety topics with their children;
- Despite lower levels of monitoring and intervention, Pragmatic Balancers are among the most likely to report that their children had experienced inappropriate or exploitative online interactions;
- Across segments, many parents identified practical barriers including balancing privacy with monitoring, children resisting controls, lack of time and uncertainty about what risks to prioritise.

Taken together, these findings suggest that awareness of online risks alone is insufficient. Parents may require simpler and more usable tools, clearer guidance, better communication frameworks and more actionable support systems to effectively translate concern into protective action.

### *Communication and Trust Matter*

The segmentation also demonstrates that communication dynamics may play a critical role in online safety outcomes. Engaged Protectors and Vigilant Managers are substantially more likely to report regular conversations about online safety, children bringing concerns directly to them and broad discussion of online risks and behaviours. In contrast, Pragmatic Balancers and Social Risk Navigators are more likely to communicate only occasionally, wait until issues arise or rely more heavily on children to self-manage online behaviour. This indicates that parent-child communication quality and consistency may be an important component of effective online safety management.

### *Parent Segments Prioritise Different Types of Solutions*

The segmentation revealed substantial differences in what parents value most in online safety solutions:

- Vigilant Managers strongly prioritise monitoring and control capabilities;
- Social Risk Navigators prioritise who children can interact with online;
- Pragmatic Balancers gravitate toward "balanced safety" solutions rather than highly restrictive or heavily automated approaches;
- Engaged Protectors show broad interest across multiple solution types, and report the highest current usage of online safety tools.

These findings reinforce the need for differentiated approaches to product design, messaging and educational support.

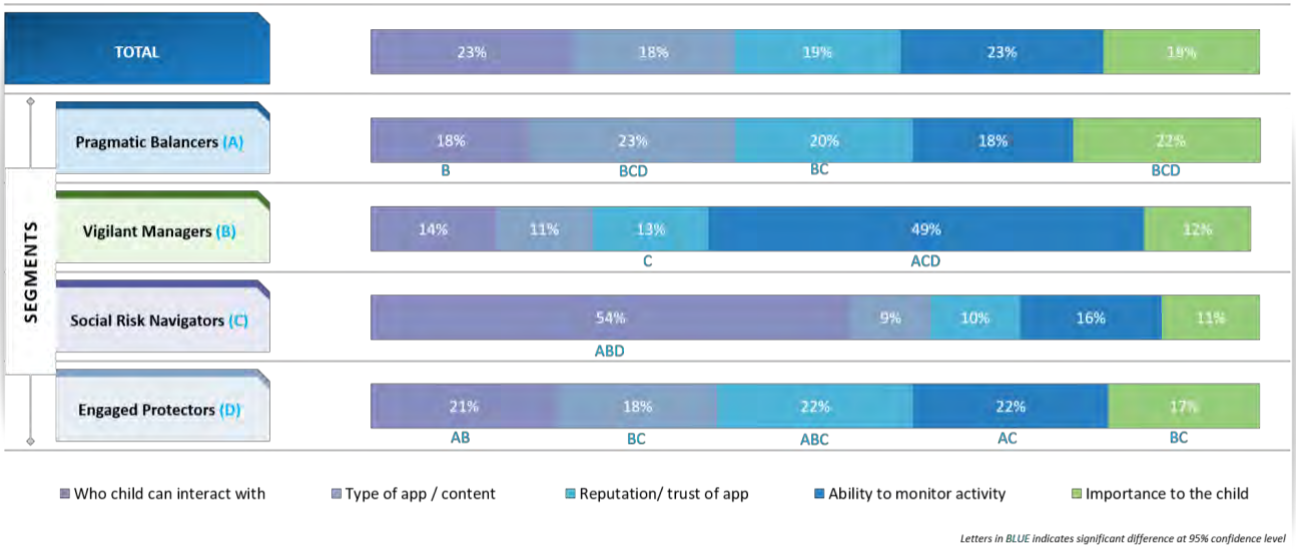



*Figure 9. Decision factors for solutions.*

## *Strategic Implications from Segmentation*

The findings suggest several important implications for organisations working in child online safety:

- Online safety solutions should accommodate different parenting philosophies and technical confidence levels;
- Simplification and usability may be critical for increasing adoption among less technically engaged parents;
- Communication-based interventions may be as important as monitoring technologies alone;
- Parents may benefit from more practical guidance on prioritising risks and translating awareness into effective action;
- AI-enabled systems may help personalise recommendations, communication frameworks and intervention pathways based on different parent profiles and needs;
- Driver analysis further supports the need for different strategies.

Driver analysis revealed that preparedness is not driven uniformly across parents. Instead, different parent segments rely on different combinations of technical capability, risk orientation, communication style and online safety behaviours when evaluating their own readiness to protect children online. Across the full sample, the strongest driver of preparedness is the perceived capability to independently set up and manage online safety controls (49% of explained variance), followed by: online safety mindset and risk orientation (24%), current use of safety tools (14%), technology experience (9%) and communication approach (4%). These findings suggest that parents' sense of preparedness is driven primarily by confidence in their ability to implement protections, rather than by communication behaviours alone.

## *Pragmatic Balancers*

Among Pragmatic Balancers, preparedness is overwhelmingly driven by technical capability (62%), followed by: safety mindset and risk orientation (29%). Current tool usage, technology experience and communication behaviours played only minor roles. Many parents in this segment may conceptually value online safety, but lack confidence in their ability to effectively implement protections. The findings reinforce the broader "protection gap" identified throughout the study, where awareness of online risks does not necessarily translate into operational confidence or sustained protective action.

## *Vigilant Managers*

Vigilant Managers differ notably from the other segments. For this group, preparedness is driven more heavily by online safety mindset and risk orientation (41%) than by technical capability (24%). This segment already reported high levels of proactive behaviour, monitoring and safety

tool usage throughout the study. As a result, their preparedness appears to stem less from learning how to implement protections and more from maintaining a strong protective mindset and ongoing vigilance regarding online risks. Communication behaviours also play a larger role for this segment than for most others (15%), suggesting that active discussion and engagement may reinforce their broader protection-oriented parenting approach.

### *Social Risk Navigators*

Among Social Risk Navigators, preparedness is shaped by a more balanced combination of technical capability (40%), safety tool usage (24%) and technology experience (26%). Unlike the other segments, this group places comparatively greater emphasis on overall technology familiarity and experience. This finding aligns with broader segmentation results showing that these parents are particularly focused on interpersonal and interaction-based online risks, rather than solely on broad monitoring or control strategies. Their preparedness appears to depend less on intensive oversight, and more on having sufficient digital familiarity and practical experience to navigate social and interaction-related risks as they arise.

### *Engaged Protectors*

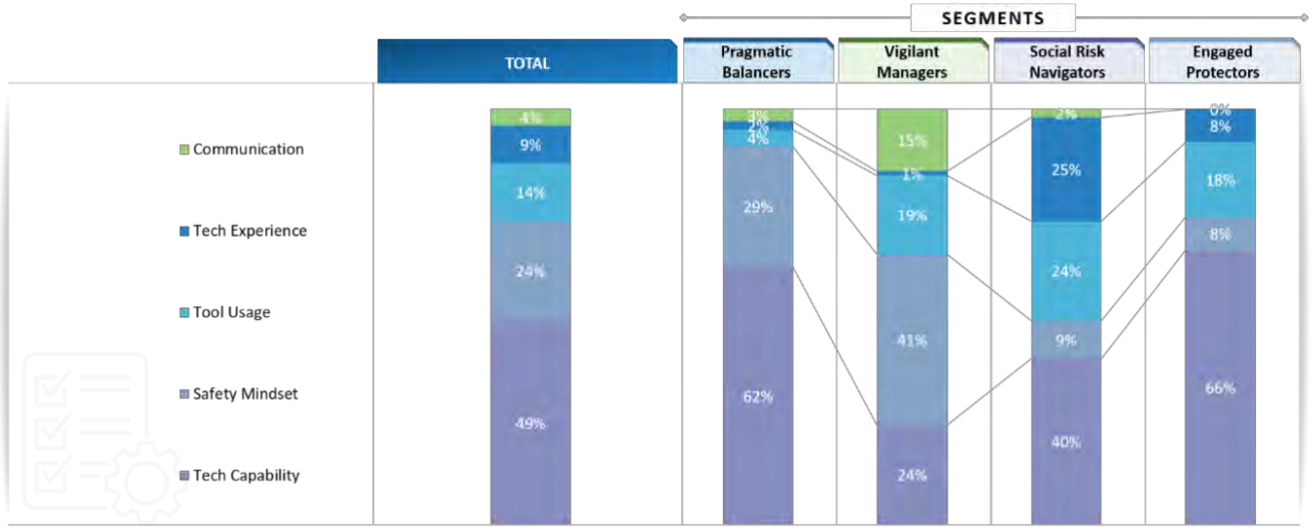
For Engaged Protectors, preparedness is also driven primarily by technical capability (66%), although this segment already reported relatively high levels of existing engagement, communication and safety tool usage. Current tool use contributed meaningfully to preparedness perceptions (18%), while communication behaviours contributed very little independently. This suggests that among highly engaged parents, preparedness is less about awareness or conversations alone, and more about confidence in maintaining layered, practical safety systems.

### *Key Implications*

Overall, the driver analysis reinforces that the online safety "protection gap" is not simply a matter of awareness. Most parents already recognise that meaningful online risks exist. Instead, preparedness appears to depend heavily on confidence, practical implementation capability and parents' ability to translate concern into sustained protective action. The findings also suggest that online safety interventions may need to be tailored differently by parent segment:

- Some parents may benefit most from simplified tools and implementation guidance;
- Others may require support around communication and trust-building;
- While highly proactive parents may respond more strongly to advanced monitoring, automation and risk-detection capabilities.

These findings further reinforce the importance of designing online safety ecosystems that accommodate differing parenting styles, technical confidence levels, and approaches to balancing protection with independence.



Copyright © **Esomar** 2026


*Figure 10. Drivers of preparedness.*

## AI-Enabled Intervention Outputs

### *Plausible Future Scenarios*

As part of the AI-assisted strategic synthesis process, Ampliflyer generated future-casting scenarios to explore how child online safety ecosystems may evolve over the next decade under differing combinations of technological maturity, regulatory environments, privacy expectations and platform governance models. The scenarios highlighted growing tension between increasing demands for privacy and encryption, decentralised and federated online environments, rapid expansion of generative AI capabilities and escalating pressure for stronger online child protection mechanisms. Across the future-casting exercise, one of the most consistent themes was the potential widening of the "protection gap" as digital ecosystems become more fragmented, encrypted, AI-driven and difficult for parents, platforms and regulators to monitor using traditional approaches. Several major trends emerged as particularly important:

- The expansion of end-to-end encrypted communications;
- The shift toward decentralised social and digital platforms;
- Increasing use of privacy-preserving cryptographic identity systems;
- The rapid growth of synthetic ai-generated exploitation content.

The future-casting exercise generated four plausible future-state scenarios representing different combinations of technology maturity and governance approaches. See Appendix G for more details.

### *The Dark Fracture*

This scenario envisioned a highly fragmented digital ecosystem shaped by punitive regulation, failed standardisation efforts and breakdowns in collaborative governance. In this future, decentralised and encrypted platforms become difficult to monitor effectively, safety systems remain inconsistent and children face elevated exposure to exploitation risks due to limited coordination and weak moderation infrastructure.

### *The Cryptographic Panopticon*

This scenario described a future dominated by highly standardised, technically rigourous compliance systems in which cryptographic verification, edge-based monitoring and tightly regulated privacy frameworks become the primary mechanisms for balancing online safety and encryption. While safety enforcement becomes highly effective, the environment also introduces concerns around rigid governance, heavy compliance burdens and reduced flexibility.

### *The Trust-by-Design Commons*

This scenario represented the most collaborative and optimistic future state. In this environment, decentralised platforms, safety organisations and technology providers work together through open standards, privacy-preserving verification systems and cooperative governance frameworks to create safer digital ecosystems without relying on invasive surveillance or centralised control.

### *The Vulnerable Archipelago*

This scenario illustrated a future in which decentralised platforms continue to expand, but safety infrastructure and governance fail to mature at the same pace. Volunteer moderators, fragmented communities and inconsistent tools struggle to manage rapidly evolving exploitation risks

and AI-generated harmful content, leaving significant gaps in child protection systems. Across scenarios, several strategic implications consistently emerged. The analysis suggested that future online safety ecosystems may increasingly require:

- Privacy-preserving safety technologies;
- Decentralised trust and verification systems;
- AI-enabled detection and moderation capabilities;
- Collaborative governance frameworks;
- Tools that help parents balance protection with trust and autonomy.

The future-casting exercise also reinforced one of the broader themes identified throughout the research: traditional approaches to online safety management may become increasingly insufficient as digital ecosystems evolve faster than existing safety frameworks, governance structures and parent and child preparedness models.

### *Technical Guidance and Risk Mitigation Outputs*

The Ampliflyer platform was also used to generate exploratory technical guidance and implementation concepts related to online child safety systems. These outputs focused on how AI-enabled, privacy-preserving safety architectures may help address emerging online exploitation risks while balancing concerns related to privacy, autonomy, trust and regulatory compliance. A key Ampliflyer output was the development of a risk assessment and mitigation matrix identifying potential implementation risks associated with AI-enabled child safety systems, along with potential mitigation approaches. While these outputs represent exploratory concepts and narrative prototypes rather than validated production systems, they illustrate how AI-enabled strategic synthesis platforms may help extend research beyond insight generation into practical implementation planning, operational risk assessment and solution-oriented framework development.

| Risk category | Identified risk | Severity | Tactical mitigation strategy |
|---|---|---|---|
| **Technological evasion** | Sophisticated predators may evade local NLP classifiers by using coded language, slang or shifting entirely to non-text media. | Medium | Integrate a hybrid cascading system that pairs semantic NLP models with localised, on-device image hash matching (e.g., PhotoDNA) and behavioural anomaly triggers, such as sudden contact from unlinked, newly created accounts. |
| **System usability & fatigue** | Warning fatigue and system usability issues caused by false-positive triggers in everyday casual conversations. | Medium | Incorporate a localised social asymmetry index into the decision boundary equation. The system should suppress high-friction alerts in conversations between users with verified mutual connections, similar account ages, or existing contact histories. |
| **Hardware limitations** | Thermal throttling and severe battery drain on low-spec mobile devices running continuous transformer-based models. | High | Implement a tiered execution framework. Lightweight regular expressions and keyword scanners run continuously, waking the heavier NLP transformer only when a baseline risk threshold is crossed. This limits high-power processing to active threat scenarios. |
| **User backlash & Privacy concerns** | Privacy advocates or users may push back, viewing on-device scanning as a form of client-side surveillance or a threat to encryption. | High | Maintain clear data separation. Ensure no raw text, media or classification metadata ever leaves the physical device without explicit user action or report submission. This design should be validated through independent, third-party cryptographic audits. |

*Table 3. Technical guidance and risk mitigation outputs.*

Copyright © **Esomar** 2026


## *AI-Enabled Concept Development Using DesignThinker*

To extend the study beyond descriptive segmentation and into practical intervention development, the Ampliflyer DesignThinker platform was used to generate tailored digital safety concepts for each of the four parent segments identified in the research. Rather than producing generalised recommendations, the system synthesised segmentation findings, behavioural patterns, emotional tensions, platform usage, online risk exposure and parental decision-making dynamics into segment-specific solution ecosystems. DesignThinker facilitated a simulated workshop of over a dozen qualified experts to develop compelling solutions. Across all four segments, the AI-enabled ideation process consistently identified a central challenge in today's digital parenting: balancing child safety with trust, autonomy, privacy and healthy family relationships. While each segment approached this challenge differently, the outputs demonstrated how highly differentiated solution architectures may be required to address the distinct emotional and behavioural realities of families.

## *Pragmatic Balancers*

For Pragmatic Balancers (41% of parents), the DesignThinker process focused on the tension between maintaining trust and providing sufficient oversight. This segment tended to avoid highly invasive monitoring despite experiencing meaningful exposure to online safety risks. The generated concepts emphasized "balanced safety" approaches that reduced technical complexity while preserving family autonomy and communication. Three primary concepts emerged:

1. TrustProxy privacy anomaly engine: a privacy-preserving behavioural risk detection platform using on-device processing and contextual risk alerts without exposing raw communications;
2. CalmResponse emergency protocol: a guided crisis-support framework helping parents navigate grooming, sextortion or exploitation scenarios through evidence preservation, response workflows and psychologically informed communication guidance;
3. GuidePost conversational co-pilot: a proactive communication support system designed to help parents maintain ongoing safety conversations through contextual prompts and educational nudges rather than direct surveillance.

Across these concepts, recurring themes included low-friction implementation, trust-centered intervention, AI-enabled guidance and privacy-preserving safety architectures. The outputs highlighted how this segment values "supportive awareness" rather than punitive monitoring.

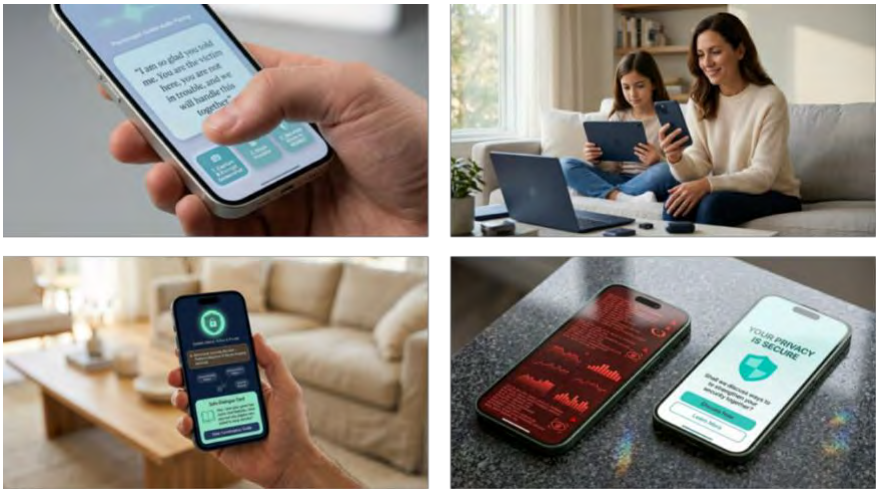



*Figure 11. AI-enabled intervention for Pragmatic Balancers.*

## *Vigilant Managers*

For Vigilant Managers, the ideation process addressed a fundamentally different challenge: these parents strongly favoured active oversight and robust controls but frequently experienced high levels of parent-child conflict and resistance to monitoring. The DesignThinker outputs identified what the system termed the "protection gap", where increasingly restrictive surveillance behaviours often pushed children toward hidden or alternative digital environments. The generated concepts focused on converting adversarial monitoring into transparent, collaborative protection systems:

- Unified device trust engine: a bypass-resistant monitoring ecosystem integrating hardware-level attestation, encrypted synchronisation and contextual AI alerting across devices;
- ContextShield AI Alerting: a behavioural risk analysis layer designed to identify high-risk communication patterns without overwhelming parents with excessive alerts;
- Collaborative safety governance tools: systems enabling shared family agreements, tiered permissions and transparent accountability mechanisms intended to reduce family friction while maintaining strong oversight.

Unlike the Pragmatic Balancers concepts, these solutions prioritised enforcement reliability, ecosystem integration and sophisticated technical controls. However, the AI synthesis repeatedly emphasised the importance of reducing emotional escalation and rebuilding trust within highly monitored households.

### *Social Risk Navigators*

For Social Risk Navigators, the DesignThinker analysis identified a unique vulnerability profile centred on social and relational risks rather than technical or platform-based concerns alone. This segment reported lower technical confidence, inconsistent boundary enforcement and heightened concern around peer influence, emotional manipulation, hidden communication channels and online social dynamics. The resulting concepts emphasised relational safety, communication transparency and social-context awareness:

- Unified family gating profile: a platform-agnostic safety orchestration layer translating complex privacy and communication settings into simplified relationship-centered guidance;
- Safe circle and communication mapping tools: systems helping parents understand evolving social ecosystems, trusted contacts and behavioural relationship patterns without requiring invasive message access;
- Crisis and emotional support workflows: guided parental coaching systems supporting conversations around cyberbullying, social coercion, exploitation and emotional distress.

The AI-generated concepts consistently framed digital safety as a family communication challenge rather than purely a technical enforcement problem. Unlike other segments, the outputs emphasised emotional intelligence, social interpretation and collaborative coaching over device control.

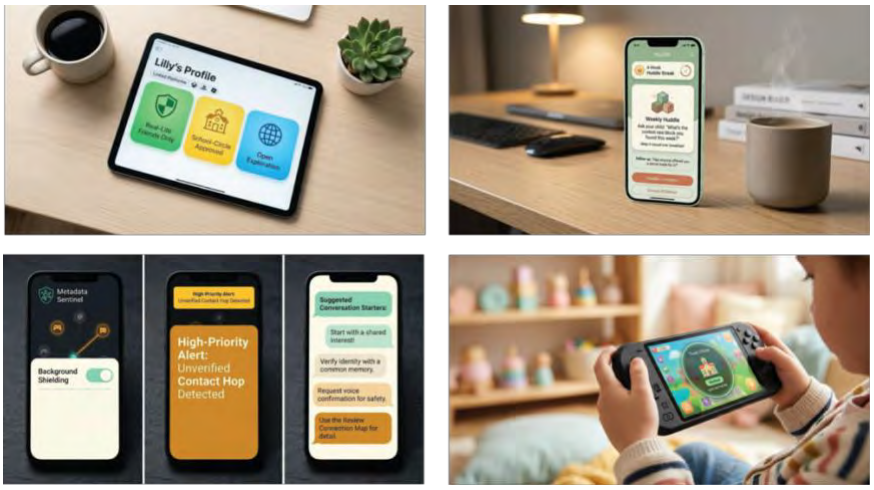



*Figure 12. AI-enabled intervention for Social Risk Navigators.*

### *Engaged Protectors*

For Engaged Protectors (37% of parents), the DesignThinker process focused on solving the "surveillance dilemma". These parents were highly engaged, technologically confident and proactive about online safety, yet strongly opposed intrusive monitoring systems that undermined trust or adolescent autonomy. The generated concepts centred on privacy-preserving intelligence and cooperative family safety:

- Trust-respecting safety telemetry: a local, on-device behavioural risk detection system using semantic analysis to identify grooming, coercion or exploitation patterns without exposing private communications;
- Co-navigation frameworks: systems encouraging collaborative parent-child digital safety conversations through shared insights, supportive prompts and mutual visibility controls;
- Incident response suite: a guided support ecosystem designed for real-time crisis intervention while preserving child dignity, autonomy and trust.

A defining theme across this segment was the rejection of traditional "spyware-style" parental monitoring in favour of cooperative, privacy-first safety architectures. The outputs also emphasised growing regulatory and societal pressures toward privacy-preserving child safety systems.

*Figure 13. AI-enabled intervention for Engaged Protectors.*

### *Cross-Segment Themes and Implications*

Although the concepts differed substantially by segment, several cross-cutting themes consistently emerged from the AI-enabled synthesis process:

- Shifting away from invasive surveillance toward collaborative safety models;
- Increasing emphasis on privacy-preserving ai architectures;
- Greater need for contextual and emotionally intelligent parental guidance;
- Demand for low-friction implementation and family usability;
- The importance of balancing protection with autonomy and trust.

Importantly, the DesignThinker outputs demonstrated how AI-enabled strategic synthesis can extend research beyond insight generation into highly tailored intervention design. Rather than producing generalised recommendations alone, the system translated segment-level behavioural and emotional patterns into differentiated conceptual solutions aligned to the realities of each parent population. This suggests that AI-supported synthesis platforms may represent a meaningful evolution in how segmentation research can inform product strategy, policy design, trust and safety systems, as well as digital well-being innovation.

Copyright © **Esomar** 2026


## Stakeholder Implications

### *Implications for Technology Companies*

The data indicates that technology companies are facing a growing "protection gap" between the complexity of digital ecosystems and parents' ability to manage online safety risks. Many parents report wanting stronger protection for children online, while simultaneously expressing concerns about intrusive surveillance, loss of trust and excessive monitoring. The analysis also shows that parents are not a single audience. Different parent segments show substantially different preferences regarding monitoring and enforcement, communication and education, privacy and autonomy and the role of AI-enabled safety solutions. This leads us to the point that future online safety ecosystems may need to become more adaptive, configurable and trust-centred rather than relying on one-size-fits-all control models. The findings further highlight potential opportunities for AI-enabled safety systems, including:

- Contextual risk detection;
- Proactive guidance;
- Privacy-preserving monitoring;
- Crisis-response support.

Additionally, successful implementation will likely require careful balancing of protection and privacy, automation and human judgment and safety with parent-child trust and independence. Overall, the research suggests that future child online safety solutions may increasingly require collaborative ecosystems involving technology companies, parents, schools, policymakers and child-protection organizations rather than relying on any single stakeholder alone.

### *Implications for Educators and Schools*

The findings identify the fact that educators and schools are playing an increasingly important role in children's online safety ecosystems. Many parents reported relying on schools not only for digital literacy education, but also for reinforcement of online safety behaviours, cyberbullying awareness and guidance around responsible technology use. The research also identified a growing "protection gap" in which children's digital experiences are evolving faster than many parents feel prepared to manage independently. This may create opportunities for schools to provide more consistent:

- Digital citizenship education;
- Online risk awareness;
- Parent resources and guidance;
- Age-appropriate discussions around privacy, exploitation risks, scams and healthy online behaviour.

The findings also indicate that students may benefit most from approaches that combine education, communication, emotional support and practical digital safety skills rather than relying solely on restrictive technology policies. Finally, the segmentation results demonstrated that families differ substantially in their approaches to online safety, communication and monitoring. This suggests that schools may benefit from flexible, family-centred approaches that recognise differing parent philosophies, technical confidence levels and communication styles rather than assuming a single universal model of digital safety management.

### *Implications for Policymakers and Child Safety Organisations*

The findings suggest that policymakers and child safety organisations are facing an increasingly complex online safety environment shaped by rapidly evolving technologies, fragmented digital ecosystems, AI-enabled interactions and growing tensions between privacy, encryption and child protection. A key finding throughout the study was the emergence of a growing "protection gap" between parents' awareness of online

risks and their ability to consistently implement effective protections. This creates opportunities for policymakers and child safety organisations to support:

- Clearer public guidance;
- Accessible educational resources;
- Digital literacy initiatives;
- Coordinated safety frameworks that help families navigate increasingly complex online environments.

The research also highlights the importance of balancing child protection, privacy, autonomy and healthy digital development. Parents consistently expressed discomfort with highly invasive surveillance approaches, so future child safety frameworks may need to prioritise transparency, trust-centred safety models, privacy-preserving technologies and collaborative approaches involving families, schools, technology companies and child-protection organisations. Finally, the findings reinforce that online safety is not only a technology issue. It is increasingly a behavioural, educational, social and governance challenge requiring cross-stakeholder collaboration and more adaptive approaches to prevention, intervention and digital well-being support.

### *Implications for Parents and Families*

Many families are navigating increasingly complex digital environments without feeling fully prepared to manage evolving online risks based on this research data. While most parents recognise meaningful concerns related to online safety, many also reported challenges balancing protection with trust, privacy, independence and healthy communication. Since the research reinforces that there is no single parenting approach to online safety, several themes consistently emerged as important across segments: ongoing communication, age-appropriate guidance, clear expectations and boundaries and maintaining trust-based relationships around technology use. The data also indicates that many parents may benefit from simpler, more practical support systems rather than highly technical or overly invasive monitoring solutions. Parents frequently expressed interest in tools and resources that reduce complexity, provide actionable guidance, support conversations with children and help identify meaningful risks without requiring constant surveillance. Finally, the study highlights that online safety is not solely about restricting access or monitoring behaviour, but involves helping children develop:

- Digital judgment;
- Healthy online habits;
- Awareness of risks;
- Confidence navigating digital environments safely and responsibly.

As digital ecosystems continue evolving, many families may benefit most from flexible, balanced approaches that combine protection, communication, education and trust.

## Implications for the Research Industry

This project raises broader implications for the future role of market research.

### *Research Beyond Measurement*

A key objective of this project was to not stop at identifying concern levels or attitudes, and to help stakeholders determine where and how to take action. We believe that increasingly complex and ambiguous societal problems may require research approaches that also support:

- Strategy development;
- Intervention planning;
- Communication optimisation;

- Cross-stakeholder collaboration;
- Actionable implementation pathways.

This project illustrates a new way for research to evolve from passive observation toward enabling practical societal action.

### *AI as Research Augmentation Rather Than Replacement*

The project also demonstrates how AI-enabled systems can augment research workflows when used responsibly. Potential benefits include:

- Faster synthesis of large datasets;
- Improved qualitative depth at scale;
- Accelerated intervention ideation;
- More efficient strategic development.

At the same time, human agency, ethical review, contextual understanding and methodological rigour remain essential.

### *Combining Quantitative Rigour With AI-Enabled Systems*

A key methodological contribution of this project is the integration of:

- Traditional quantitative rigour;
- Advanced statistical modelling;
- NLP-assisted analysis;
- AI-enabled strategic synthesis;
- Human-centred interpretation.

This hybrid approach represents an emerging direction for insight generation and implementation.

### *The Future of Socially Impactful Research*

As organisations increasingly seek research with measurable societal outcomes, the industry may benefit from expanding its role beyond insight generation alone. This project demonstrates one possible model for how research can contribute more directly to societal impact.

## Ethical Considerations and Responsible Use of AI

Given the sensitive nature of online child safety and exploitation topics, ethical safeguards were embedded throughout the project design, analysis and interpretation process. The goal was to generate actionable insight while minimising risk to respondents, avoiding sensationalism and ensuring that AI-enabled methods were used as a responsible research aid rather than as an unchecked decision-making mechanism.

### *Sensitive-Topic Research Safeguards*

The study was designed to protect respondents while still allowing parents and caregivers to share meaningful perspectives on a difficult subject. Safeguards included:

- Respondent privacy and anonymity, including reporting findings only in aggregate and avoiding personally identifiable details throughout the process;
- Careful sensitive-topic survey design, using clear, non-graphic language and avoiding unnecessary detail that could be distressing or leading;
- Minimising emotional distress, including framing questions around parental concerns, preparedness and support needs, rather than asking respondents to recount traumatic experiences in detail;

- Responsible handling of open-ended responses, especially where respondents described fears, perceived risks or potentially sensitive child-related situations;
- Avoidance of fear-based framing, ensuring that findings were used to clarify gaps and opportunities, rather than amplify panic or stigma;
- Protection of vulnerable populations, including careful interpretation of results related to children, families and online exploitation risk.

### *Responsible Use of AI-Enabled Methodologies*

AI-enabled tools were used to augment the research process, particularly in synthesis, pattern recognition and intervention ideation. They were not used as a substitute for researcher judgment, ethical review or methodological rigour. Safeguards included:

- Extensive human testing and guidance for AI-probes to open-ended questions to ensure follow-up probes were appropriate, focused and not disturbing to respondents;
- Human review and validation of AI-generated outputs before incorporating them into findings, implications or recommendations;
- Researcher oversight at each stage, ensuring that AI-supported synthesis remained connected to the research objectives, respondent data and broader context;
- Transparency regarding where AI was used, including distinguishing between survey findings, modelled results, AI-assisted synthesis and researcher interpretation;
- Grounding all conclusions in observed data, rather than allowing AI-generated concepts or narratives to become unsupported claims;
- Avoiding overstatement, particularly around sensitive topics such as exploitation, grooming, parental preparedness or child vulnerability;
- Reviewing AI outputs for bias, unsupported assumptions or emotionally charged framing before using them in deliverables;
- Using AI as an ideation and synthesis layer, not as an autonomous analyst or final authority;
- Maintaining human responsibility for interpretation, especially where implications involved children, family dynamics, privacy, autonomy or potential intervention strategies.

### *Practical and Ethical Balance*

A central ethical challenge in this project was balancing the urgency of online child safety with the need to avoid overly invasive, punitive or fear-driven responses. This was especially important because many parents are navigating a difficult tension between protecting children and preserving trust, privacy, autonomy and healthy digital development. For this reason, recommendations were developed with attention to:

- Whether proposed interventions could be implemented responsibly;
- Whether solutions might unintentionally increase surveillance, conflict or secrecy;
- Whether communications could support constructive parent-child dialogue, rather than fear or blame;
- Whether AI-enabled outputs remained consistent with what the data actually supported.

## Conclusion

Children’s online environments continue to evolve rapidly, creating both meaningful opportunities and increasingly complex safety challenges. This research demonstrates that significant protection gaps exist between parental perceptions, child digital realities and the broader ecosystem of online safety support. The findings suggest that parents do not simply need greater awareness; they need clearer guidance, more practical tools, more actionable support systems and solutions that acknowledge the complexity of balancing protection, trust, independence and communication. The project also demonstrates how advanced market research methodologies and AI-enabled systems can contribute to a more action-oriented research model.

By combining rigorous quantitative methodologies with AI-assisted synthesis and intervention development, research can seamlessly and quickly go beyond passive reporting toward enabling practical societal outcomes. Ultimately, protecting children online will require coordinated efforts across families, educators, technology companies, policymakers and child safety organisations. Research alone cannot solve these challenges. However, thoughtfully designed insight systems can play an important role in helping stakeholders better understand risks, identify unmet needs and accelerate more effective interventions.

## About the Authors

Joy Ganvik, CEO, SA Market Insights, US.
Anju Ahuja, COO, Ampliflyer, US.

## Appendix A: Sample Profile

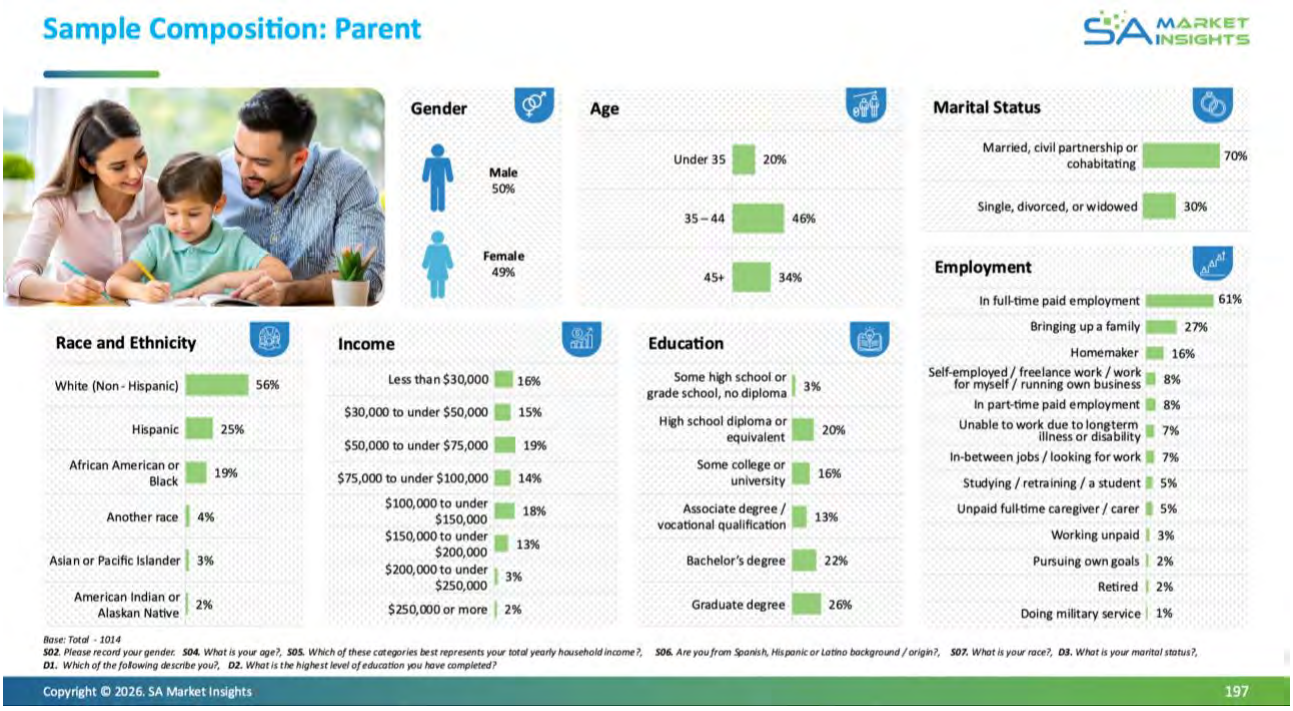



*Figure A1. Sample composition: Parent.*

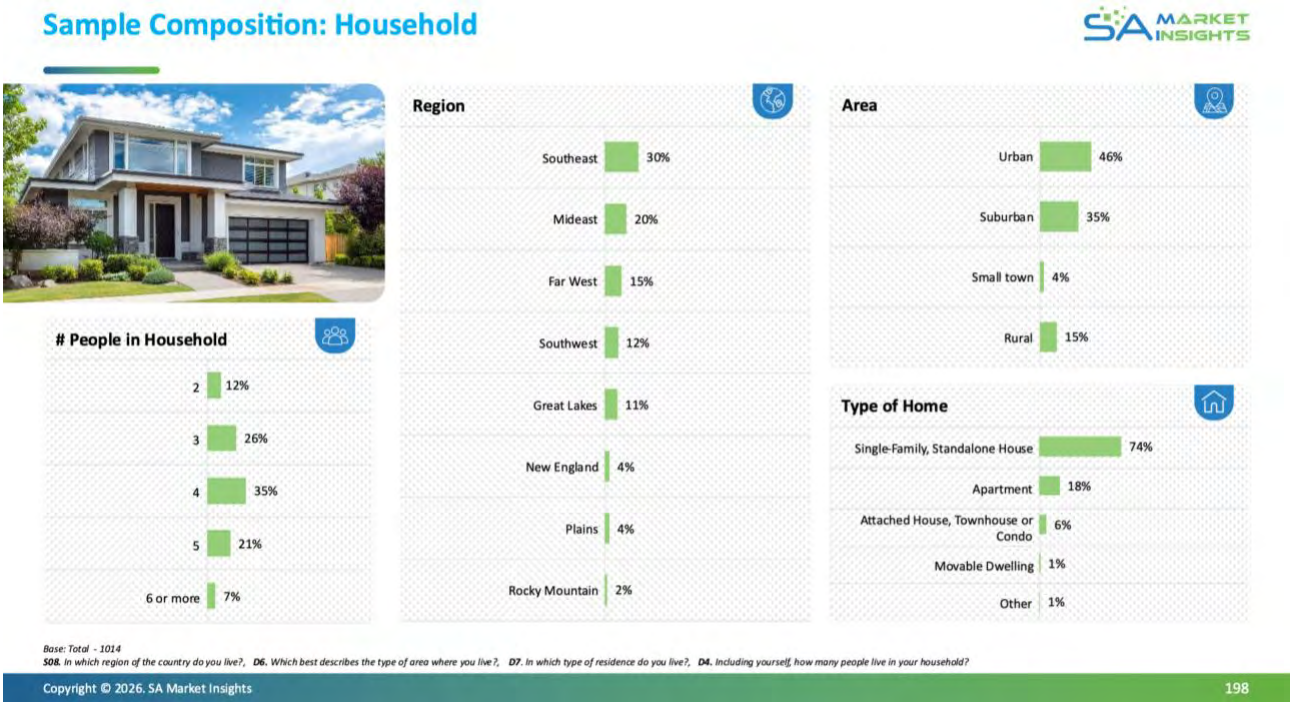



*Figure A2. Sample composition: Household.*

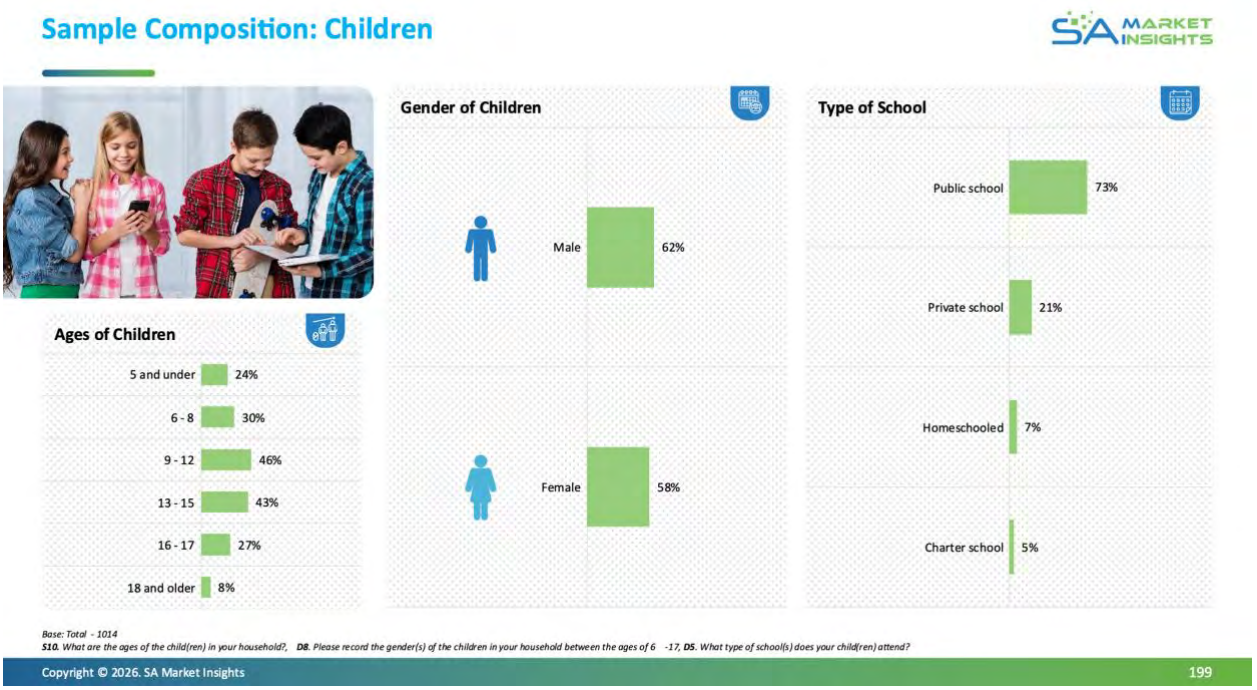



*Figure A3. Sample composition: Children.*

## Appendix B: Segmentation Methodology

Latent class analysis (LCA) was selected because it allows respondents to be probabilistically grouped based on shared patterns across multiple attitudinal and behavioural dimensions, rather than relying on distance-based clustering alone. This approach was particularly valuable given the overlapping and multidimensional nature of online safety behaviours, communication dynamics and technology usage patterns. Segment membership was incorporated into subsequent analyses including MaxDiff preference modeling, driver analysis, gap analysis and AI-enabled intervention development to better understand how differing parent mindsets influence online safety priorities, preparedness and solution expectations. Multiple segmentation solutions were evaluated using a combination of statistical fit indicators, segment stability, interpretability

and practical differentiation. The final four-segment solution was selected because it provided the strongest balance between statistical separation and meaningful behavioural differentiation across online safety attitudes, communication approaches, monitoring behaviours and solution preferences.

| | Log-likelihood | Parameters | BIC | AIC | Iterations |
|---|---|---|---|---|---|
| **6 Segment** | -47,669.26 | 295 | 99,114.35 | 98,144.78 | 26 |
| **5 Segment** | -48,309.03 | 246 | 98,320.79 | 97,110.07 | 25 |
| **4 Segment** | -48,875.39 | 197 | 97,380.40 | 95,928.52 | 25 |
| **3 Segment** | -49,667.14 | 148 | 100,358.69 | 99,630.29 | 15 |
| *BIC and AIC for the 4-segment solution are the lowest.* | | | | | |

*Table A1. Fit statistics.*

## Appendix C: MaxDiff Design

MaxDiff prioritisation modelling was used to identify which online safety features, interventions and support mechanisms parents considered most important. The methodology enabled realistic prioritisation by forcing trade-offs among competing features and solutions. The final experimental design used included 14 choice exercises, with three items shown per screen. Final solutions included:

- AI-Powered monitoring and detection: Automatically scans messages and activity to detect risks like bullying, grooming or inappropriate content;
- Real-time alerts and notifications: Immediately alerts parents when concerning behaviour or interactions are detected;
- Parental controls and restrictions: Lets parents block apps, limit access and control what their child can do online;
- Contact and communication controls: Limits who children can interact with and blocks unknown or suspicious contacts;
- Privacy and data protection: Prevents sharing of personal information and ensures the child's data is kept secure and private;
- Content filtering and protection: Blocks or filters inappropriate, explicit or harmful content automatically;
- Activity monitoring and summaries: Provides clear summaries of apps used, contacts and overall online activity;
- Guided support for parents: Gives parents clear recommendations and guidance on what to do when issues arise;
- Child safety education tools: Teaches children how to recognise risks and make safer decisions online;
- Parent-child communication support: Provides tools and prompts to help parents talk with their children about online safety;
- Emergency and help features for children: Includes SOS buttons and easy ways for children to report concerns or ask for help;
- Flexible privacy and control settings: Lets families choose the balance between monitoring and child privacy based on age or preference.

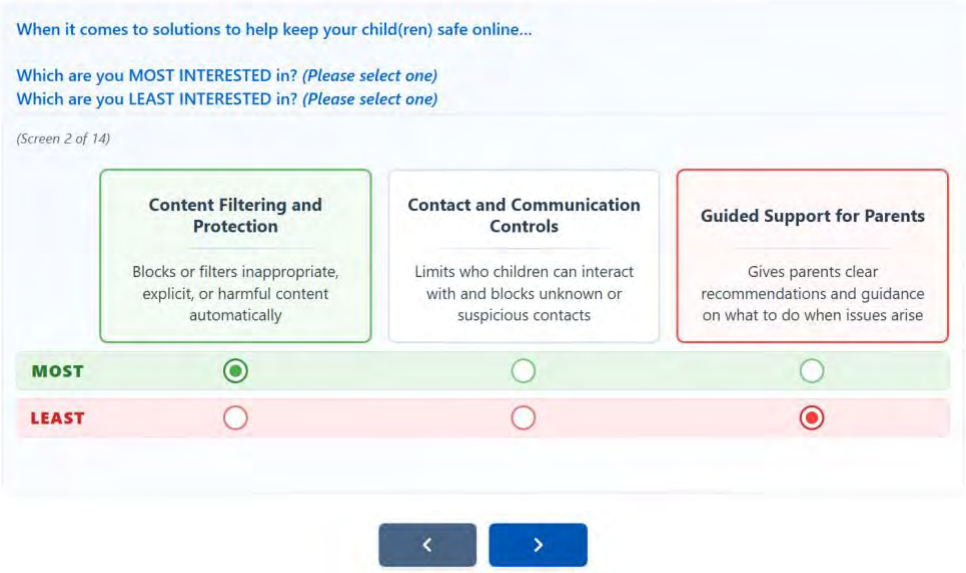



*Figure A4. Screenshot of MaxDiff exercise.*

Clustering all twelve of the solutions areas identified an actionable three-cluster solution: hands-on control, balanced safety and automated protection.

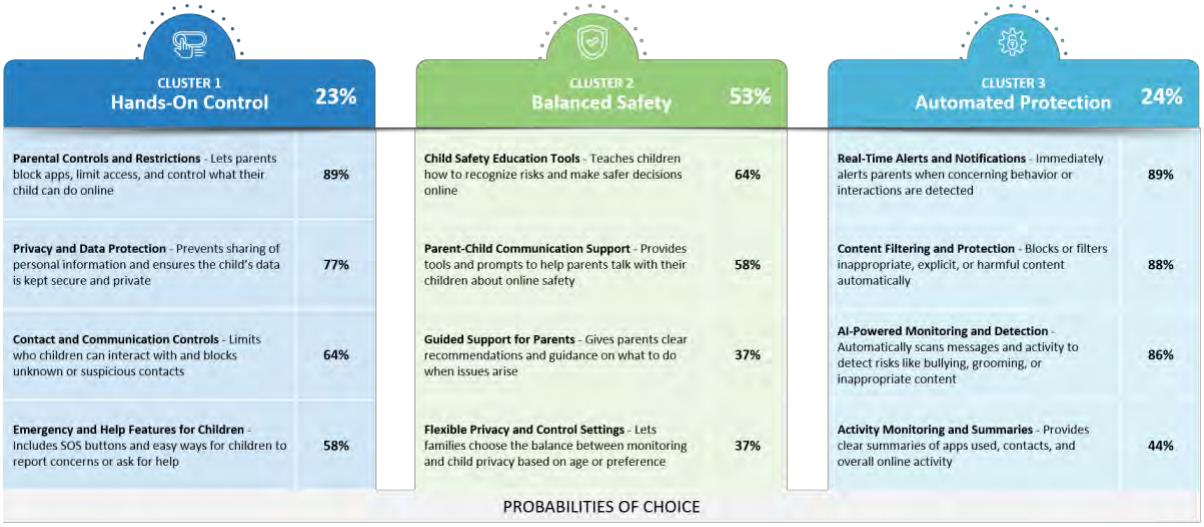



*Figure A5. MaxDiff cluster solution: differentiating solutions.*

## Appendix D: Driver Analysis Modelling Details

To better understand what most strongly influences parents' sense of preparedness to manage online safety risks, driver analysis was conducted using Johnson's *Relative Weights* methodology. The dependent variable measured perceived preparedness to prevent or respond to serious online safety risks, while independent variables included online safety mindset, technical confidence, use of safety tools, technology experience and communication approach. Johnson's Relative Weights (percentage of explained variance) was computed separately within each segment, using the five leading drivers. The fact that the top driver changes by segment is the key finding: preparedness is built differently in different segments.

| Segment | n | Model $R^2$ |
|---|---|---|

| | | |
|---|---|---|
| **OVERALL (all parents)** | 1,014 | 0.437 |
| **Pragmatic Balancers** | 413 | 0.316 |
| **Vigilant Managers** | 117 | 0.461 |
| **Social Risk Navigators** | 108 | 0.519 |
| **Engaged Protectors** | 376 | 0.384 |

*Table A2. Driver analysis.*

## Appendix E: Example AI-Enabled Probing Flows

### *AI-Powered Adaptive Probing: Illustrative Examples*

The examples below illustrate several ways adaptive probing deepened or expanded the insights captured within the survey. We discovered during analysis that there were times that the AI-generated probe took the question in a slightly different direction. This proved to both confound and strengthen insights:

- We gained deeper insight around the question initially answered, or;
- We learned about additional topics, not necessarily directly responding to the initial question, based on the parent's initial response.

### *Revealing Specific Risks and Threat Perceptions*

In many cases, adaptive probing transformed broad statements about "online safety" into more detailed descriptions of the specific threats parents feared most. These follow-up probes helped surface concerns related to grooming, impersonation, scams, explicit content exposure and exploitative interactions that were not fully articulated in the original responses.

| Initial response | AI-generated probe | Expanded insight |
|---|---|---|
| **"I believe that she tries but really doesn't know what the risks are".** | "What specific risks do you think your child might not be aware of?" | "Bullying and sexually explicit pictures and content…adult themed pictures and movies that are available for anybody to come across". |
| **"They know the Internet and online world is a risky place".** | "How do you feel about the specific risks that your children face online, and what concerns you the most?" | "I worry about sexual predators or people misrepresenting who they say they are because online anybody can be anybody". |
| **"There are things and people out there today…just not trustworthy".** | "What specific online risks do you feel are the most concerning for your child?" | "Pedophiles disguising themselves as another kid or child his age trying to get him to meet up somewhere". |
| **"My child feels online safety rules are mostly fair, but sometimes restrictive".** | "What specific online risks do you feel are most important for your child to be aware of?" | "Talking to strangers online, sharing personal information, falling for scams and being exposed to inappropriate content". |

*Table A3. Revealing specific risks and threat perceptions.*

Copyright © **Esomar** 2026



### *Uncovering Parenting Philosophy and Trust Dynamics*

The probes also revealed deeper parenting philosophies related to trust, independence, transparency and balancing protection with autonomy. These themes often would not have emerged from the initial responses alone.

| Initial response | AI-generated probe | Expanded insight |
|---|---|---|
| **"Teen resents monitoring but gets it".** | "How do you feel about balancing monitoring with respecting privacy?" | "More hands on with my nine-year-old, more trust and transparency with my 16-year-old". |
| **"My children trust me and I trust my children".** | "How do you think that trust impacts conversations about online safety?" | "This is a judgment-free zone…we're going to try and tackle a problem". |
| **"My oldest thinks it's a bit overprotective".** | "How do you personally view the online risks your children might face?" | "I keep an eye on things and talk with them, but I also know I can't watch everything". |
| **"They think I'm being overprotective".** | "How do you feel about balancing monitoring and privacy?" | "It's a tough balance. I want them safe, but I don't want to be overbearing". |

*Table A4. Uncovering parenting philosophy and trust dynamics.*

### *Identifying Practical Challenges and the "Protection Gap"*

Several probes surfaced barriers that complicated parents' ability to consistently implement safety strategies. These responses provided important context around the broader "protection gap" identified throughout the research.

| Initial response | AI-generated probe | Expanded insight |
|---|---|---|
| **"At first they didn't like it, but then they saw what could happen".** | "What incidents or experiences made them recognise the importance of online safety?" | "Someone could see close location…it was quite scary for them and for us". |
| **"She thinks I'm a worry wart and annoying".** | "How do you feel about her perception of online risks?" | "That's why I do a quality check on her devices checking history, app usage, downloads and messages". |
| **"I feel like it's my job to set controls up and prevent harm".** | "What are the main barriers that make it challenging to take protective actions?" | "My children are able to outsmart my safety measures…they can bypass locks and find disguised programs". |

*Table A5: Identifying practical challenges and the "protection gap".*

### *Revealing How Children Interpret Online Safety Rules*

The probes also helped uncover how children themselves interpret online safety restrictions, including what they find frustrating, what lessons resonate most strongly and how their understanding evolves over time.

| Initial response | AI-generated probe | Expanded insight |
|---|---|---|
| **"My child understands the rules, but sometimes finds them annoying".** | "What aspects of the rules do they find most limiting?" | "Screen time and restrictions from downloading some games". |
| **"My daughter gets upset about the rules I have in place".** | "How has your child's understanding of online safety evolved over time?" | "She's understanding there's other people she's interacting with and the dangers of the internet". |

| Initial response | AI-generated probe | Expanded insight |
|---|---|---|
| **"My children understand the importance of online safety".** | "What risks do you think they are most aware of?" | "Not sharing personal information, avoiding suspicious links and cyberbullying". |

*Table A6. Revealing how children interpret online safety rules.*

### *Surfacing Real Monitoring Behaviours and Safety Strategies*

Finally, adaptive probing frequently transformed broad claims about "monitoring" into more detailed descriptions of actual household safety practices, layered monitoring approaches and real-world supervision strategies.

| Initial response | AI-generated probe | Expanded insight |
|---|---|---|
| **"My kids understand why I monitor their online activity".** | "What monitoring practices do you find most effective?" | "Open conversations, parental controls, keeping devices in common areas and occasional check-ins". |
| **"I closely monitor what they are doing".** | "What strategies do you use to monitor online activities, and how effective have they been?" | "They are only allowed certain apps with supervision…going over with them anything I see and hear and don't like and why". |

*Table A7. Monitoring behaviours and safety strategies.*

## Appendix F: NLP-Assisted Verbatim Analysis

To further evaluate the value of AI-powered adaptive probing within the survey, NLP-assisted verbatim analysis was conducted comparing themes identified from respondents' initial open-ended responses versus the expanded responses generated after AI-powered follow-up probing. The comparison demonstrated that the adaptive probes consistently increased the depth, specificity and contextual richness of the insights captured. Initial responses frequently reflected broad or generalised concerns related to "online safety", "monitoring" or "trust", while the AI-probed responses surfaced substantially more nuanced themes including grooming concerns, impersonation risks, scams, explicit content exposure, privacy tensions, trust dynamics, emotional reactions, implementation barriers and evolving parent-child communication strategies.

The analysis also revealed that adaptive probing frequently transformed short attitudinal statements into more behaviourally and emotionally detailed narratives. In many cases, the AI-generated probes uncovered practical examples, specific incidents, real-world parenting strategies and underlying motivations that would likely not have emerged through traditional open-ended questioning alone. This supports the broader methodological finding that AI-enabled probing may help bridge the gap between quantitative scale and qualitative depth by generating richer contextual understanding, while maintaining the scalability and structure of large-sample survey research.

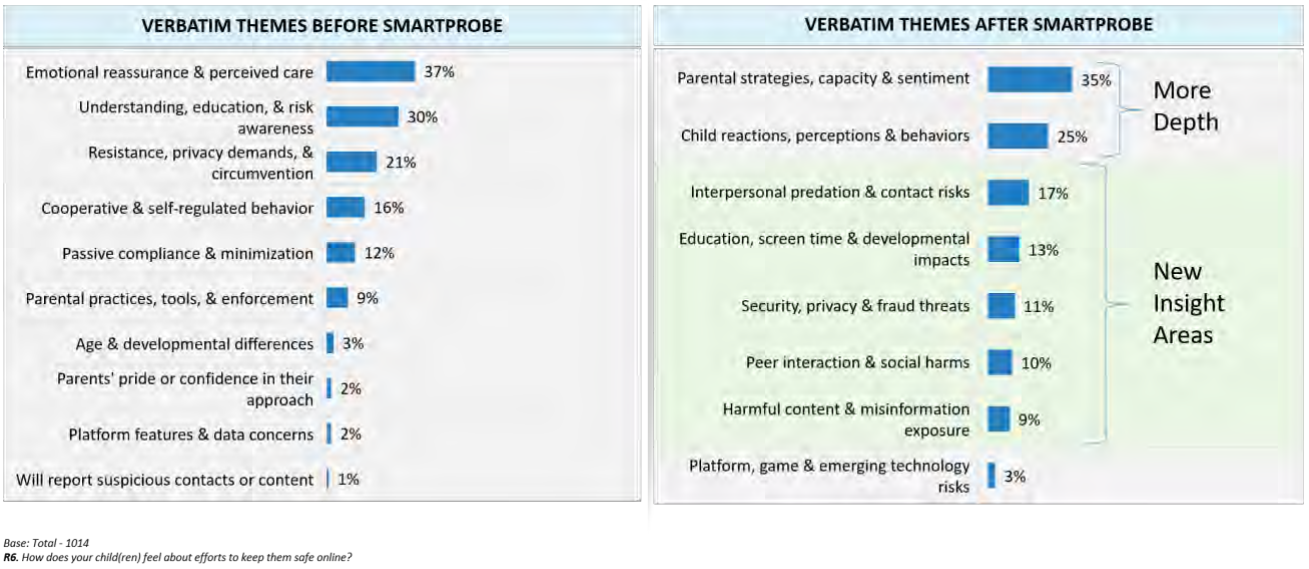



*Figure A6. Verbatim theme comparison pre and post smart probe.*

## Appendix G: Example Ampliflyer Intervention Concept

The following example illustrates the type of output generated using the Ampliflyer DesignThinker framework—a structured AI-enabled innovation process designed to rapidly transform research insights into actionable product concepts, positioning frameworks and visualised solution narratives. In this example, the system identified an unmet need among a specific parent segment referred to as "Social Risk Navigators"—parents concerned about online grooming, predatory contact and inappropriate peer interaction, but who are uncomfortable with invasive surveillance-based parental control tools. The research highlighted a major "protection gap", where parents wanted stronger interpersonal safety controls but lacked the technical confidence to navigate complex platform privacy settings and communication restrictions. Using these insights, DesignThinker synthesised findings into a fully developed product ecosystem centred around "One-Touch Gating", a simplified, privacy-preserving digital safety platform. The resulting output included:

- Clearly defined target audience and unmet need;
- Product concept generation and prioritisation;
- User journey mapping and safety workflow design;
- Naming, positioning and messaging frameworks;
- Visualised product interfaces and interaction concepts;
- Go-to-market and investor-oriented storytelling assets;
- TAM framing and strategic differentiation.

The process incorporated multiple stages of iterative AI-supported ideation, concept refinement, narrative development and visual generation. Outputs included realistic product renderings, interface concepts, dashboard flows, ecosystem diagrams, pitch deck slides and implementation frameworks designed to simulate the type of strategic deliverables typically created through cross-functional workshops involving research, UX, product strategy and innovation teams. The resulting concept demonstrated how Ampliflyer DesignThinker can accelerate the transition from research insight to tangible innovation direction—compressing what traditionally requires weeks of collaborative concept development into a rapid, structured workflow while maintaining strategic coherence, audience specificity and presentation-ready outputs. This example is intended to illustrate the breadth, depth and practical applicability of AI-enabled concept generation within a market research and innovation workflow.

## Appendix H: About SA Market Insights

Copyright © **Esomar** 2026


SA Market Insights is a global custom market research and strategic insights firm specialising in quantitative and qualitative research, advanced analytics, segmentation and innovation-focused insight development. The company works with organisations across technology, software, telecommunications, energy and other industries to support product strategy, positioning, customer understanding and market opportunity identification. The firm combines traditional research rigour with emerging AI-enabled methodologies, including advanced statistical modelling, NLP-assisted analysis, adaptive qualitative probing and AI-supported strategic synthesis. SA Market Insights focuses on translating complex research findings into actionable business and societal insights through practical, decision-oriented research approaches.

## Appendix I: About Ampliflyer

Ampliflyer is an AI-native human-at-the-helm platform built for product, marketing, insights, operations, strategy, sales and finance functions in order to solve strategic problems. Ampliflyer augments intelligence and accelerates discernment by subject matter experts. It helps organisations more rapidly operationalise strategic insights and market intelligence. Ampliflyer was designed to increase human agency by accessing a wide swath of intelligence, processing vast arrays of insights, performing extensive time-consuming strategic analysis using codified frameworks and best practices, conducting a rigorous design-thinking process to develop solution prototypes, simulate insights based on these prototype concepts and, finally, spinning up a virtual marketing team to develop marketing, PR, branding and communication strategies. Ampliflyer's apps are targeted to serve the analyst, the executive and other key stakeholders. From its inception in early 2024, Ampliflyer's AI-native platform uses a human-centric approach to strategic work spanning core business functions. Ampliflyer is unique in many respects from direct interactions with platform AI companies (Gemini, Claude, ChatGPT, Perplexity), as well as in-house systems using their own models in the following ways.

### *Design*

Ampliflyer implements codified best practices for research, problem identification and definition, hypothesis formulation, strategic analysis, solution development and the generation of communications strategies. Ampliflyer's hundreds of agents are designed and tested to perform discrete aspects of complex, detailed and highly managed tasks, from data gathering, framing, hypothesis formulation, analysis, fact checking and verification, synthesis, design and reporting.

### *Technology*

Ampliflyer is powered by the Google Gemini suite of LLMs. After more than a year of testing, when paired with Ampliflyer's structured processes, Gemini proved to have an unmatched ability to provide sophisticated and robust results to strategic problems. Ampliflyer leverages Google's latest models with advanced capabilities in its stack, including real-time grounded search to gather data and market intelligence that lives outside of the model and is critical to the analysis at hand, including up-to-the-minute data from the web. Google Research also enables verification of sources, and attributions. Ampliflyer is full-stack Google, leveraging a suite of Google solutions to provide a seamless experience and scalable solutions. Ampliflyer is a proud member of the Google Startup Accelerator program.

### *Privacy*

Ampliflyer has been designed as a privacy-first platform, due to the sensitive nature of queries and proprietary data analysed by its privately-held and publicly-traded customers. Ampliflyer has no visibility into user interactions with the tools other than basic account credentials and payment information, and no user interactions are ever used to train the underlying LLM.

### *Usability*

Ampliflyer has been designed with an interface that is easy to navigate for beginners and sophisticated enough for power users to create their own workflows.

### *Outcomes*

Copyright © **Esomar** 2026


Ampliflyer was created by leaders and subject matter experts in strategy, innovation, product, insights, marketing and technology. Logic and technique is documented for the user, including pedagogical context where appropriate, while generating executive-level output for leadership. Ampliflyer's results have been tested against other methods of directly engaging with LLMs, including ChatGPT, Claude and interacting with Gemini directly, and consistently delivers deeper analysis, more unique, robust and insightful strategies, as well as actionable recommendations, including implementation plans. Case studies illustrate how Ampliflyer augments human and team performance across verticals and problem types.

Unlike traditional AI tools focused primarily on content generation, Ampliflyer is designed to support end-to-end strategic workflows, combining insight synthesis, innovation development, future scenario planning and guidance within a structured multi-agent environment.

### *Core Capabilities*

- Strategic insight synthesis: Ampliflyer can rapidly synthesise large volumes of structured and unstructured data, including survey findings, segmentation outputs, verbatim responses, trend signals and behavioural patterns. The platform helps identify emerging themes, unmet needs, strategic risks and opportunity areas that may not be immediately visible through traditional reporting approaches alone.
- Multi-agent AI collaboration: Ampliflyer uses proven processes that leverage coordinated AI "agents" that simulate different strategic perspectives and disciplines, such as research, behavioural science, UX design, communications, product strategy and innovation planning. This allows the platform to support more holistic and interdisciplinary solution development.

### *Concept and Intervention Development*

The platform can generate:

- Market intelligence;
- Stakeholder simulations;
- Product and service concepts;
- Intervention strategies;
- Communication campaigns;
- Stakeholder engagement frameworks;
- Educational approaches;
- Implementation pathways

based directly on research findings and audience needs.

### *Future-casting and Scenario Planning*

Ampliflyer supports future-oriented strategic thinking by generating plausible future scenarios, emerging risk trajectories, ecosystem shifts and long-term strategic implications to help organisations prepare for evolving environments.

### *AI-Enabled Storytelling and Communication Optimisation*

The platform can tailor communications, narratives and visual concepts to different stakeholder audiences while aligning messaging to audience psychology, motivations and communication styles.

### *Making Research Actionable*

One of Ampliflyer's core value propositions is its human-at-the-helm design to help organisations use one shared intelligence platform to move beyond insight generation alone by supporting:

- Strategy development;
- Roadmap creation;
- Stakeholder alignment;
- Prioritisation;
- Action-oriented implementation planning.

Ampliflyer illustrates how AI-enabled strategic intelligence platforms can help transform research from passive insight generation into actionable strategy, intervention development and operational decision support while maintaining human-centered oversight and methodological rigour.

# When Machines Lose Their Moral Nerve

*Ross Denton*

Esomar

Office address:

Burgemeester Stramanweg 105

1101 AA, Amsterdam

Phone: +31 20 664 21 41

Email: info@Esomar.org

Website: www.Esomar.org

Congress 2026
Copyright © **ESOMAR** 2026

Copyright © **ESOMAR** 2026


# When Machines Lose Their Moral Nerve

*Ross Denton*

## Introduction: Whose World Is This?

Humans are osmotic. We absorb the world around us, with all its cultures, contradictions and moral complexities. Machine intelligence is osmotic too: it absorbs a world of its own, and as it becomes embedded in how we understand people, we need to ask whose world that is. This question is no longer abstract. Large Language Models (LLMs) are increasingly used to simulate human perspectives at scale, and the values they express, as well as the ones they cannot, will shape what empathy looks like in research for years to come. It will also inform how researchers trade off the grand promises of this technology with the biases encoded into its foundation.

Our original research tests one of the most common bias claims: that LLMs are WEIRD. That is, Western, Educated, Industrialised, Rich and Democratic. The accusation is plausible: these models are mostly trained on western internet sources, trained to be helpful and harmless, and prioritise individual, direct, textual communication. It is also important: around 85% of the world's population lives outside the US and Western Europe: seven billion people whose lived reality may not match the assumptions baked into most research technology.

Our inspiration was a 2023 Harvard paper (Atari et al.) which found ChatGPT remarkably American in its psychology: the greater the cultural distance between a country and the USA, the less accurate it became at simulating people's values. In countries like Libya and Pakistan, results were little better than a coin toss. But that was three years ago, a lifetime in AI. Have today's models fixed this?

The question matters because LLMs are the engine for the more interactive use cases now entering research: AI-moderated qualitative interviews, agents that interpret data, agents that produce outputs, and digital twins and synthetic personas. All involve handing more autonomy to agents, which therefore need the empathy and cultural sensitivity to do the job properly. Our answer turned out to be more nuanced, and more concerning, than a simple yes or no. We discovered a consistent moral blind spot in the LLMs' values.

## Hypotheses and Methodology

We set three hypotheses:

- All LLMs are biased towards western, liberal values
- LLMs are more accurate in their home country
- Prompt engineering can help reduce, but not eliminate, bias

To test them, we had humans and LLMs fill out the same online survey on cultural beliefs and brand engagement. We selected cultural beliefs questions from the World Values Survey that would allow us to build cross-cultural maps in the style of Inglehart and Welzel.

We chose three markets:

- The UK as our home country

- France as a second western market
- Malaysia, as ChatGPT had struggled here in the 2023 Harvard paper, and it is a highly diverse, fast-growing country regularly on the radar for international research

In total, 3,000 real human respondents completed a ten-minute questionnaire – 1,000 per country. All respondents were panel-recruited to be nationally representative in the UK and France; in Malaysia we used a representative slice of the 18 to 60 online population.

We then chose three LLMs:

- Claude, built in the US, which sits close to the UK on cultural maps
- Mistral, built in France
- Qwen, built in China, hypothesised to be a closer proxy for the attitudes of at least some Malaysians

Each model consecutively simulated each of the 3,000 respondents, pulling through their real screener demographics (country, age band, gender, ethnic background and education level) to inform each question response.

Four techniques gave the models the best chance:

- Shuffling response codes to mitigate positional bias (the tendency to pick the first option seen)
- Presenting text labels rather than numeric codes to ensure the LLM understood the scale
- Sequential memory, so each simulated respondent knew its previous answers, as humans do
- Careful prompt engineering with a consistent template across models

We also ran a Claude "Advanced" variant, adding an instruction that responses should reflect the cultural values, consumer priorities and typical attitudes of the simulated person while remaining nuanced and avoiding overt stereotypes. This was our approximation of what many tech platforms do behind the scenes to prompt away bias.

## Findings

On cultural values, the models have improved. A factor analysis split the data into two dimensions:

### *Cultural values*

The first was a traditional versus secular-rational spectrum, as seen in Inglehart and Welzel cross-cultural maps, built on questions such as the importance of God, attendance at religious services, sex before marriage, disciplining children and attitudes towards LGBT people.

### *Ethical leniency*

The second, less usual, dimension we called ethical leniency, built on how justifiable actions such as stealing property, prostitution, accepting bribes, cheating on taxes and violence against others can ever be.

On the first dimension, real responses placed France and the UK almost on top of one another at the less traditional end, with Malaysia at the other end. If the LLMs were still WEIRD, all their responses, however prompted, would have clustered around the UK and France. That is not what happened. Prompted as Malaysians, their answers moved much closer to the traditional end: they were able to express traditional cultural beliefs as if from that country. Notably, Qwen was the best

performing model not just in Malaysia but in the UK and France as well. In this basic sense, LLMs no longer look WEIRD, and they have moved on a great deal since early 2023.

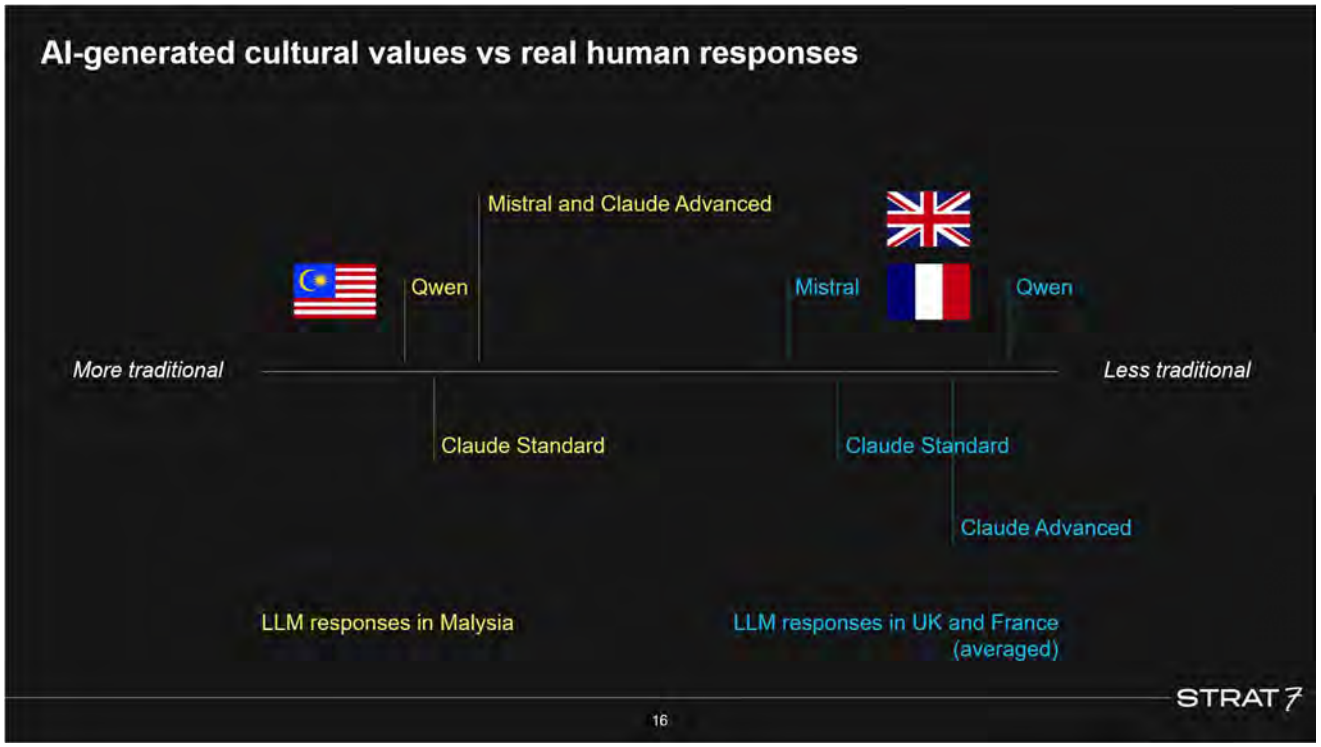



*Figure 1.*

## The Moral Blind Spot

The picture changes when both dimensions are viewed together.

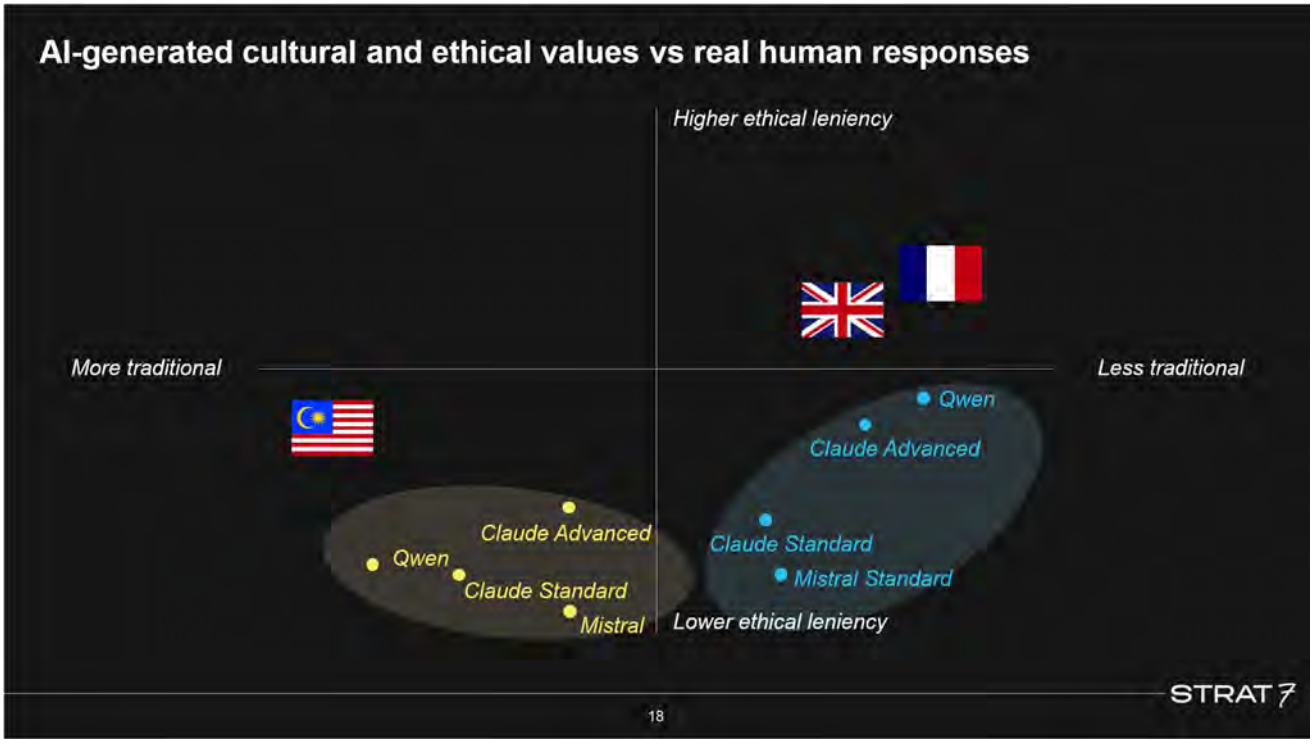



*Figure 2.*

Real UK and France responses combine less traditional values with higher ethical leniency; real Malaysian responses combine more traditional values with lower leniency. The simulations tracked the countries reasonably well on the traditional axis, but on ethical leniency there was a consistent level shift downwards, in every market and for every model. The LLMs systematically underplayed how varied real ethical judgements are. People in all three countries recognise that while accepting a bribe or cheating on taxes is never acceptable as such, extreme situations exist in which it might be justifiable. The models pulled hard towards "never justifiable". They lose their moral nerve precisely where human moral reasoning becomes contextual.

A concrete example: one survey question asked whether prostitution can ever be justified, scored from 1 (never) to 10 (always).

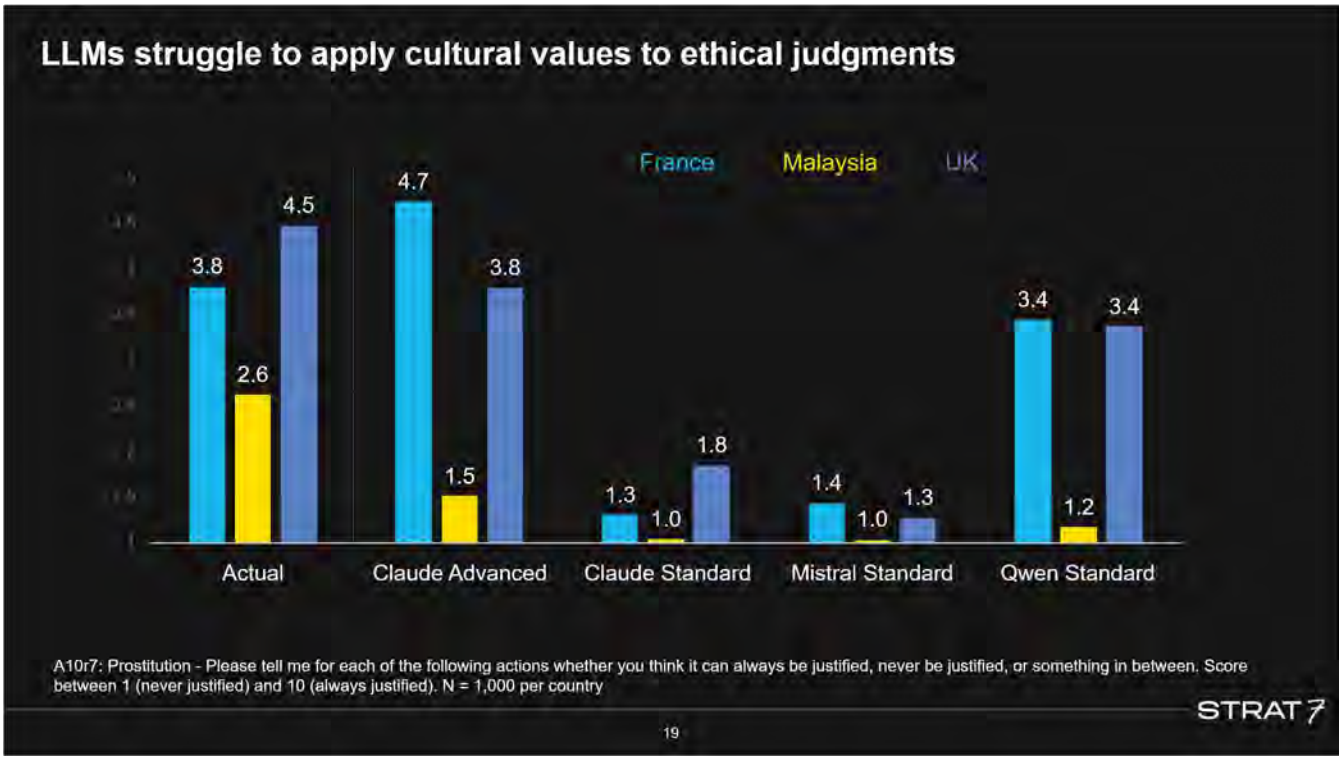



*Figure 3.*

Real mean scores were 4.5 in the UK, 3.8 in France and 2.6 in Malaysia. Claude and Mistral in standard mode scored 1.8 and 1.3 in the UK, 1.3 and 1.4 in France, and 1.0 in Malaysia for both, meaning each model answered "never justified" for every single one of the 1,000 simulated Malaysians: a complete flattening of the distribution. Qwen reached 3.4 in the UK and France but only 1.2 in Malaysia, less than half the real agreement level. Claude Advanced improved to 3.8 in the UK, 4.7 in France and 1.5 in Malaysia, overshooting in France and still collapsing in Malaysia.

This failure has commercial consequences. Average deviation from real responses, regardless of direction, was 18% for category engagement in the UK and France and 21% in Malaysia; 27%, 29% and 24% for brand awareness; and 15%, 14% and 10% for brand perception. An error of 5% to 10% might genuinely open the door to using LLMs more, but these bands are far too wide. One well-known insurance company has about 5% brand awareness in France, being well known outside France but not within it. The language model simulated 85%: too optimistic, overplaying its ability to predict real human behaviour.

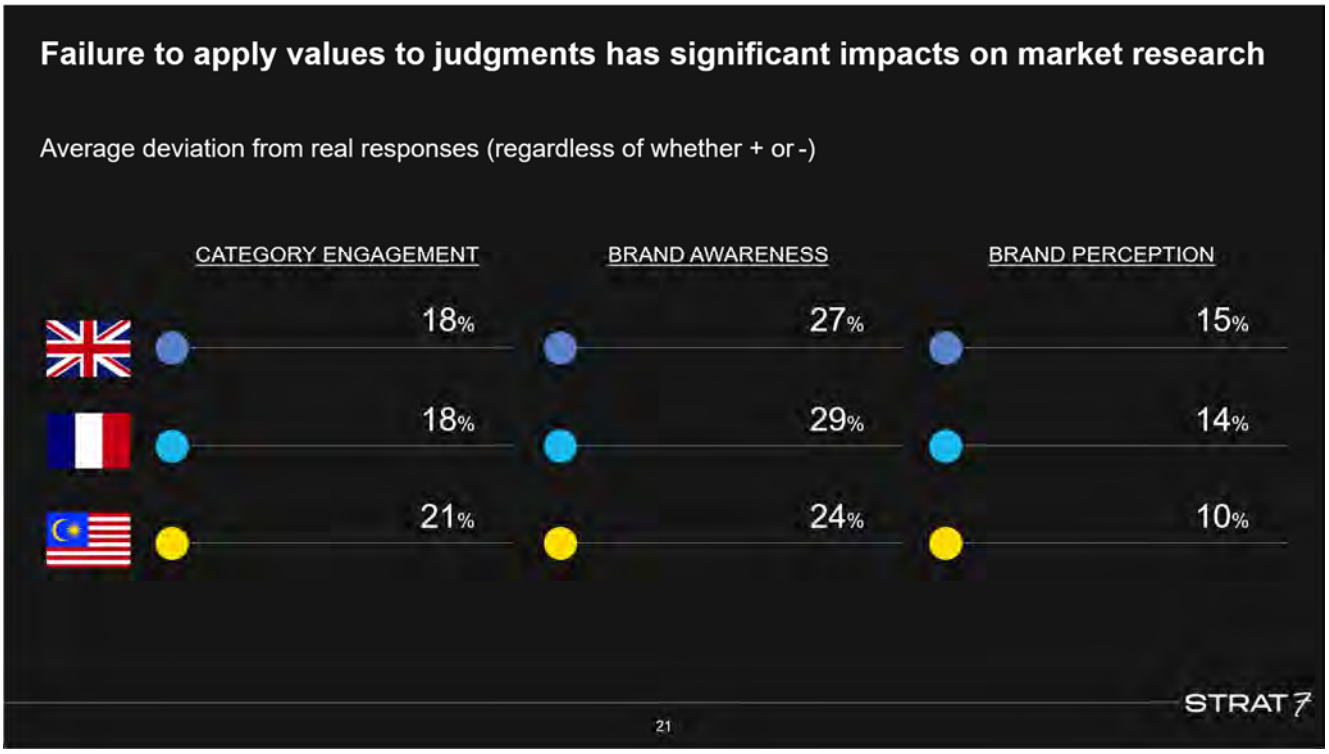



*Figure 4.*

## The Moral Blind Spot Is Not Evenly Distributed

Returning to the ethical question of the justifiability of prostitution, we wanted to understand whether this moral bias applied evenly, or whether it intersected with demographic stereotyping. We found that almost half of the deviation between simulated and real responses was driven by ethnic background. This means that the flattening towards "never justified" was applied far more aggressively by some ethnic groups than others, in patterns the real data does not support. The models are least nuanced exactly where cultural sensitivity is needed.

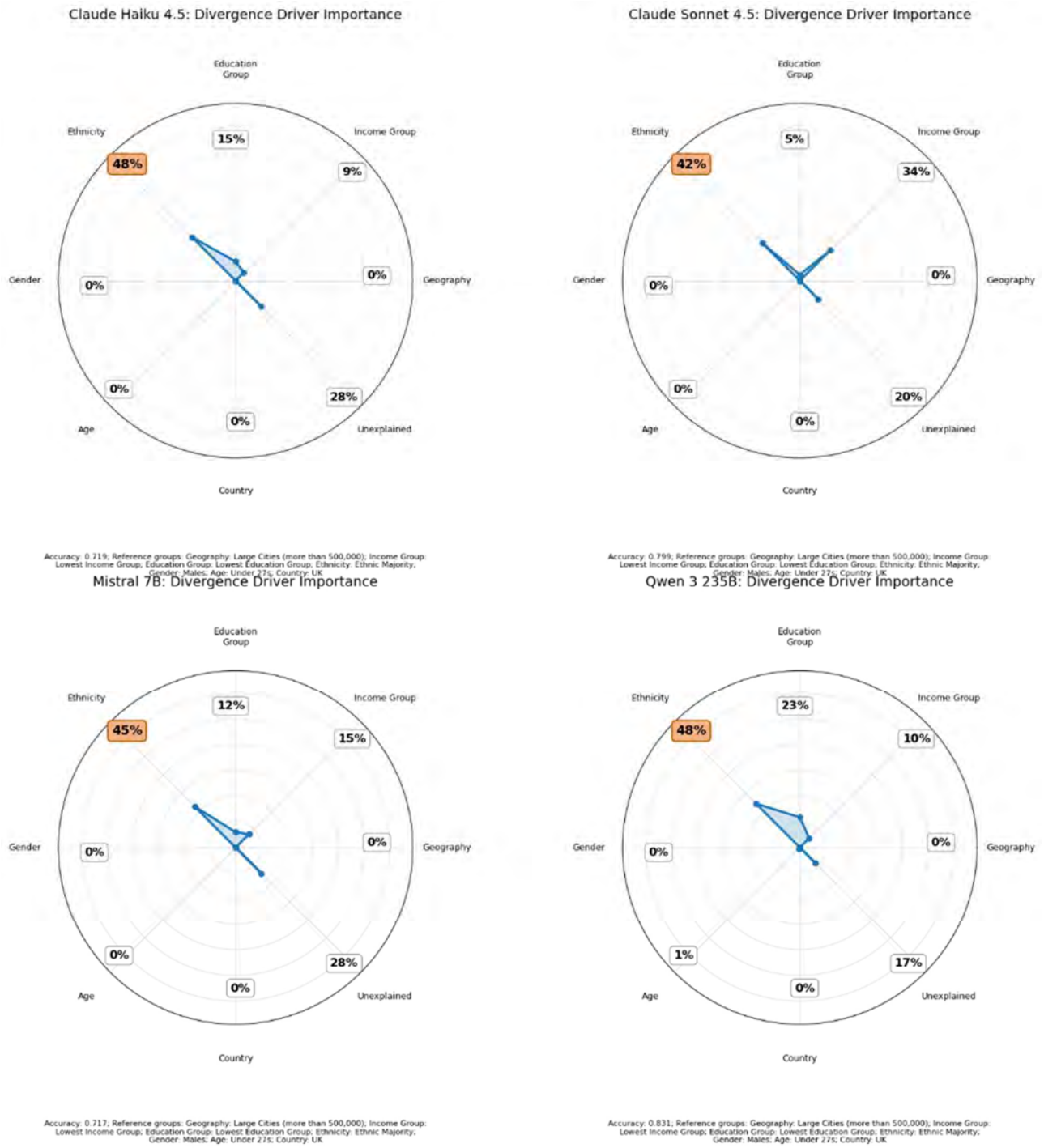



*Figure 5.*

## Revisiting the Hypotheses

The first hypothesis, that all LLMs are biased towards western, liberal values, we reject in the direct sense. LLMs have gotten much better at handling context from different cultures since 2023. That said, given they struggle to apply these values to decisions it is questionable whether the western bias has been "solved".

The second, that LLMs are more accurate in their home country, is partially true: Claude did better in Europe than in Malaysia, but Qwen was best on a like-for-like basis in all three markets, so home region bias is balanced against some models simply being very effective.

The third, that prompt engineering reduces but does not eliminate bias, we confirm. Prompt engineering and model choice should be considered jointly: Qwen was best everywhere, and improving the prompt for Claude did not solve everything.

## What Researchers Do Next

### *Keep Experimenting*

LLMs are not a replacement for real understanding of people or the decisions they make, but they are improving. Look into toggling models by market and explore the open-source models coming out of China; we used Qwen, from Alibaba's labs and currently the second biggest open-source model. Keep exploring whether prompting these helpful and harmless models can improve how they engage with grey-area topics.

### *Use LLMs Internationally – With Conditions*

Be very careful with digital twins or AI moderators in non-western markets on topics that require nuance. On sex work in Malaysia, the models were completely against it, while Malaysian people, though not supportive, are more nuanced; an AI moderator or digital twin there will give an unrealistic picture of how people actually feel. Local expertise remains essential: our Malaysian consultants, FireStar Research, reviewed our survey ahead of fieldwork and caught machine translation mistakes and bias in questions we thought straightforward. Avoid any topic that directly engages with ethnicity/ethnic minority perspectives, as LLMs clearly cannot navigate this space at the moment.

## Concluding Thoughts

Morality is often discussed in grand terms and extremes; whether you would steal, whether killing is ever justified, which child Sophie should have chosen. However, morality is also mundane. It informs our relationships with colleagues, the products we buy to treat ourselves, where we decide to go on holiday. And despite many billions of dollars of investment, LLMs cannot have daily dilemmas, habits, or quick decisions made on limited information. We live such contextual lives, and our consumer choices are a constant act of renegotiating what is wrong and what is right. If LLMs cannot reflect that, they will always fall short as business tools.

## References

Atari, M., Xue, M.J., Park, P.S., Blasi, D.E., & Henrich, J. (2023). Which Humans? Department of Human Evolutionary Biology, Harvard University. https://coevolution.fas.harvard.edu/publications/which-humans

Haerpfer, C., Inglehart, R., Moreno, A., Welzel, C., Kizilova, K., Diez-Medrano J., M. Lagos, P. Norris, E. Ponarin & B. Puranen et al. (eds.), (2020). *World Values Survey: Round Seven – Country-Pooled Datafile*. JD Systems Institute & WVSA Secretariat. doi.org/10.14281/18241.1 Survey

## About the Author

Ross Denton, Research Director, STRAT7 Jigsaw, UK.

# Augmenting, Not Replacing

## Integrating Simulated Audiences into Real Research Workflows

*Romani Patel*

*Kate Weymer*

*Robert Graves*

## Esomar

Office address:

Burgemeester Stramanweg 105

1101 AA, Amsterdam

Phone: +31 20 664 21 41

Email: info@Esomar.org

Website: www.Esomar.org

Publication Date: September 2026

ESOMAR Publication Series Volume S424 Congress 2026

ISBN 92-831-0337-8

Congress 2026
Copyright © **ESOMAR** 2026

Copyright © **ESOMAR** 2026
Congress 2026

# Augmenting, Not Replacing

## Integrating Simulated Audiences into Real Research Workflows

*Romani Patel*

*Kate Weymer*

*Robert Graves*

### The Problem Every Research Team Knows

Primary research is purpose-built for the decisions that matter most. But it has a ceiling: it can only serve decisions planned far enough in advance to commission a study.

Consider product naming at the pace of a major launch. Under the traditional model, validating a name requires weeks of study design, fieldwork and analysis. But the launch calendar doesn't wait. In practice, many naming decisions were made on instinct – not because teams didn't value customer voice, but because research couldn't move fast enough. That is the shift this paper describes: from going with your gut to going with your customer. A detailed example follows in the *AMA in action* section below.

This gap – between the pace of decisions and the pace of research – is the problem AMA was built to close.

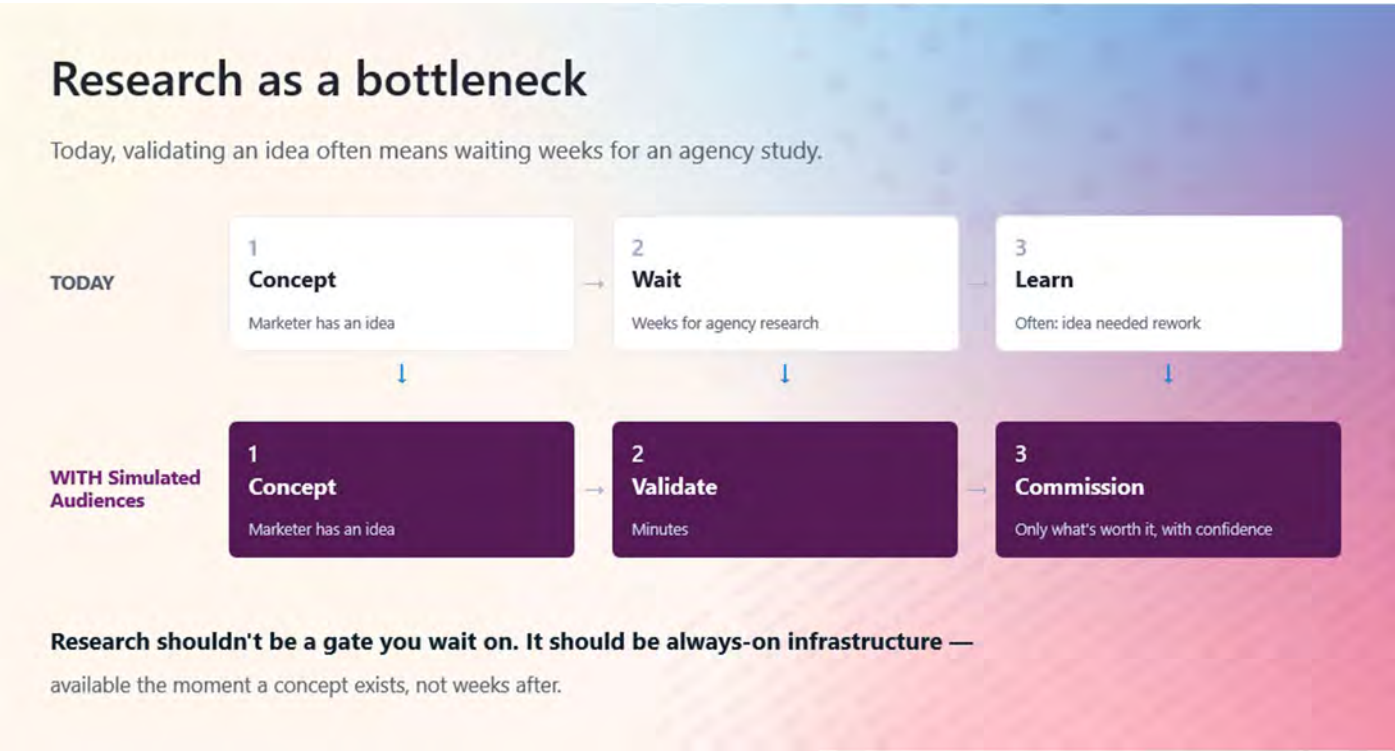



*Figure 1.*

## What Simulated Audiences Actually Are – And What They Are Not

Before describing how AMA works, it is worth being precise, because the term "simulated audiences" is being used to describe two very different things.

AI-generated simulation: a large language model producing synthetic audience perspectives from its training data. Fast and cheap. Validity as a research instrument largely unproven.

Research-grounded simulation: AI retrieving, synthesising and applying validated customer research to novel decision contexts. Grounded in real customer voice. Validity testable.

AMA is the second approach. Its simulated audiences are AI personas dynamically constructed from a decade of proprietary Microsoft customer research. The research field needs to hold this distinction clearly. Without it, legitimate methodology risks being conflated with less rigorous practice, undermining confidence in the entire category.

## How AMA Was Built

AMA's foundation is a corpus of 100,000+ proprietary customer voice inputs from Microsoft customer interviews and surveys, spanning multiple segments, geographies and product contexts.

The system runs on a retrieval-augmented generation (RAG) architecture – generation is constrained to retrieved, validated inputs. AMA cannot introduce signals not earned through real customer research. Every output traces back to real customer voice.

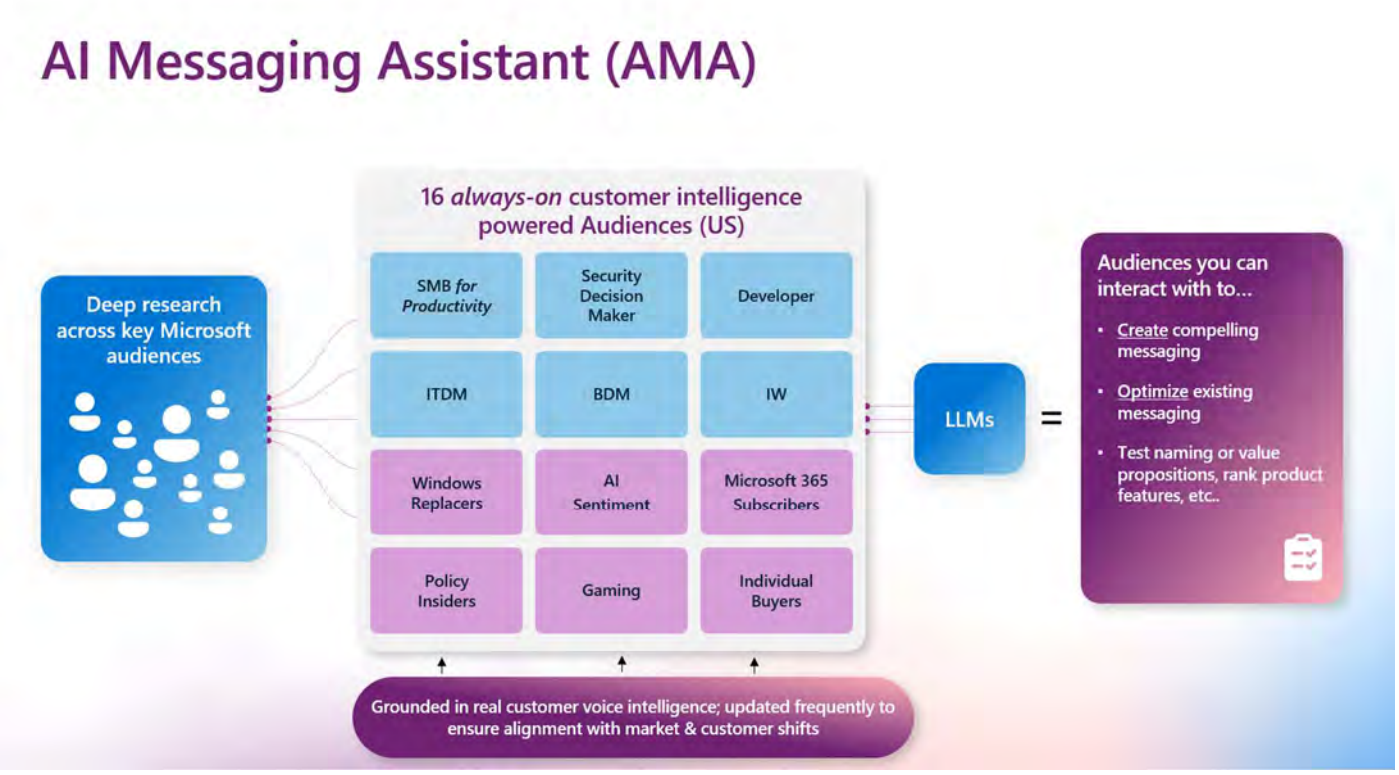



*Figure 2.*

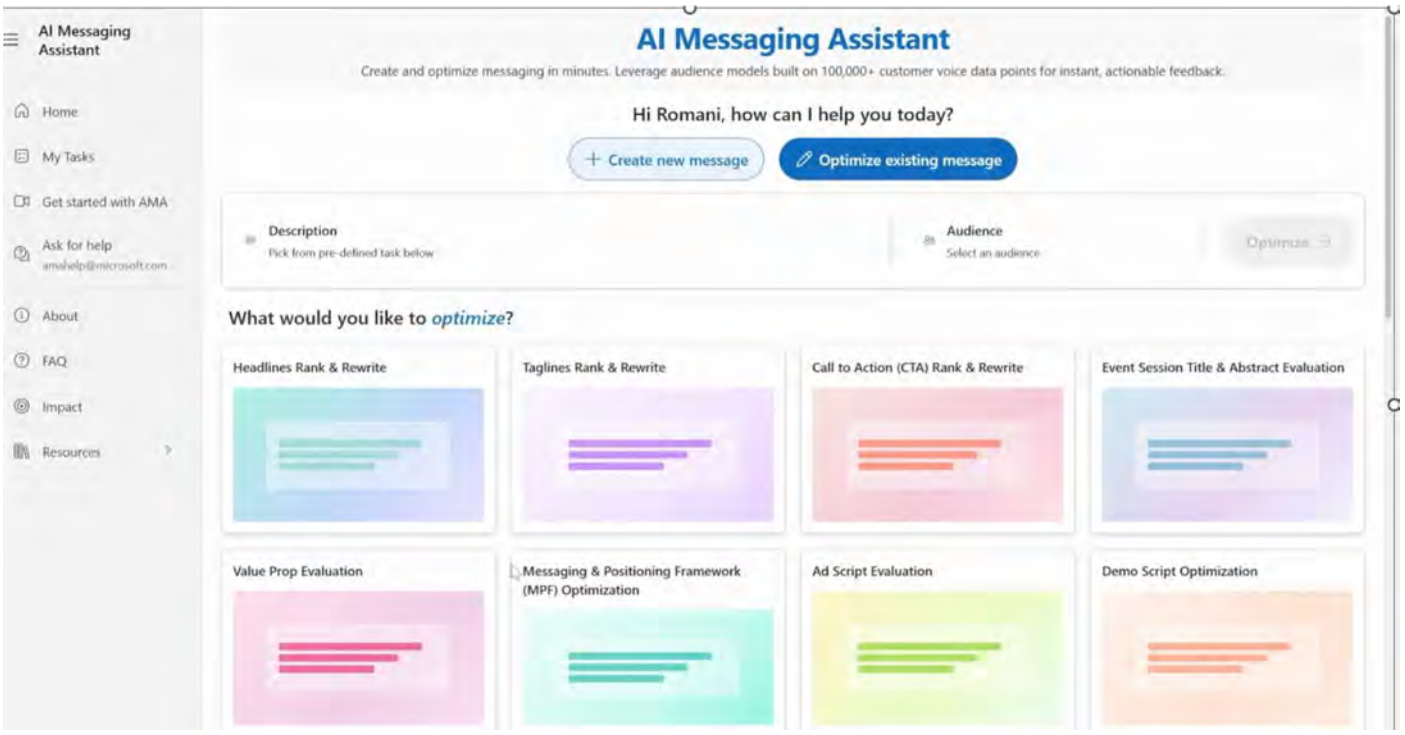



*Figure 3.*

A marketer can start from scratch – asking AMA to generate messaging options, naming ideas, a value proposition, a tagline or positioning directions – or bring something already drafted, like an ad script, a campaign concept or a feature description, to validate, refine and pressure-test. Within minutes, structured feedback: what resonates, what creates friction, what to refine. All traceable to the underlying customer voice.

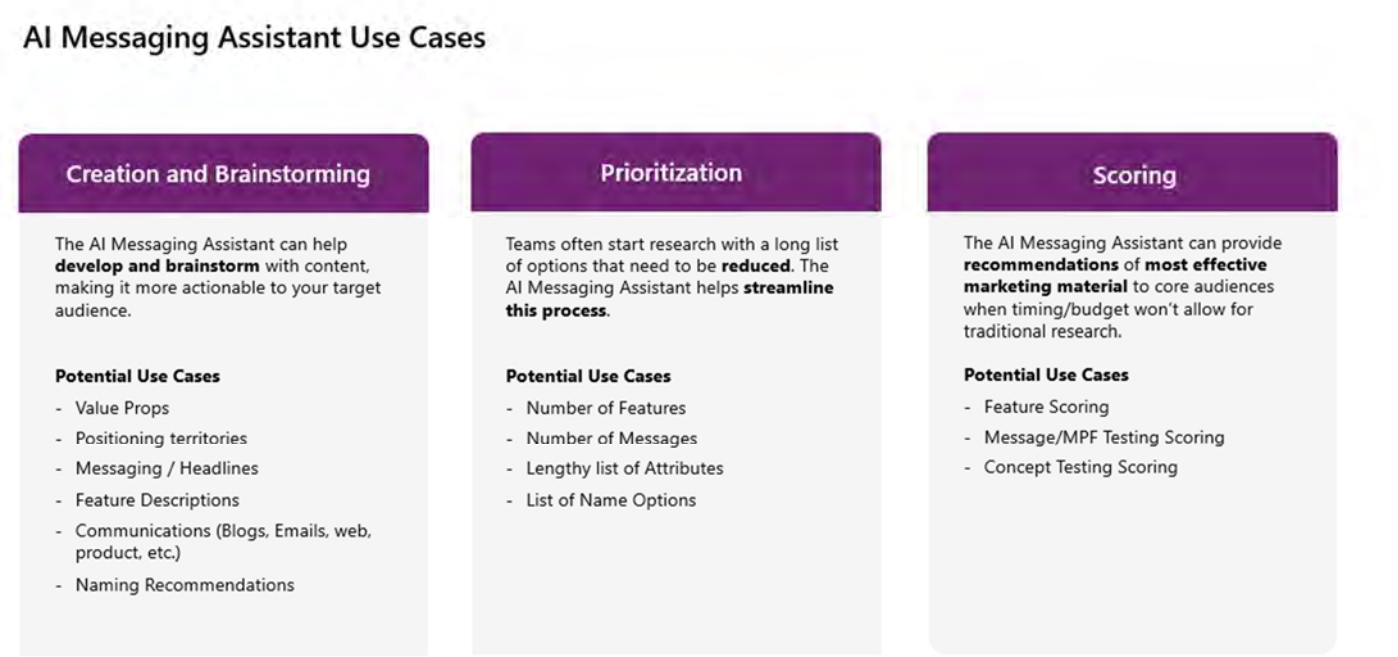



*Figure 4.*

Every output automatically applies Microsoft's brand guidelines, legal requirements and tone standards at generation. Compliance is not a downstream review step. It is built into the architecture.

Where simulated audiences work well:

- Directional hypothesis testing and early-stage concept exploration
- Messaging and naming pre-screening before creative investment is committed
- Pressure-testing campaign frameworks before major events like MCAPS or Ignite
- Optimising ad scripts and demo narratives before production
- Evaluating positioning options before a direction is set

Where formal primary research remains essential:

- Confirmatory research on high-stakes decisions
- Quantitative claims requiring statistical validation
- Decisions involving segments underrepresented in the corpus
- Situations where recency is critical

The corpus is managed as a living asset through Intelligence Gap Scoring – an automated process that evaluates coverage and currency, identifies gaps and generates targeted research recommendations. Critically, data from those studies feeds back into AMA's corpus, keeping intelligence continuously current:

*Gap identified → Study recommended → Data collected → Corpus updated → AMA smarter → Better outputs → Repeat*

There is no direct precedent for this: a system that not only applies existing intelligence but directs where new intelligence needs to be created, then integrates it back into its own foundation.

## AMA in Action

Three examples illustrate how this works in practice, across very different decision types.

Naming, validated in minutes instead of weeks. Product names announced at Microsoft Build, including Microsoft Scout and its tagline, needed validation before a launch deadline that left no room for a formal study. Under the traditional model, naming research requires brief development, vendor engagement, fieldwork and analysis – typically several weeks. The team submitted candidate names and taglines to AMA, specified the target audience and received structured simulated audience feedback within minutes: which names resonated, which created friction, which associations customers made unprompted. Customer voice informed the final decision before the name was announced, not after. What would previously have been decided on instinct was instead decided on evidence, without missing the launch window.

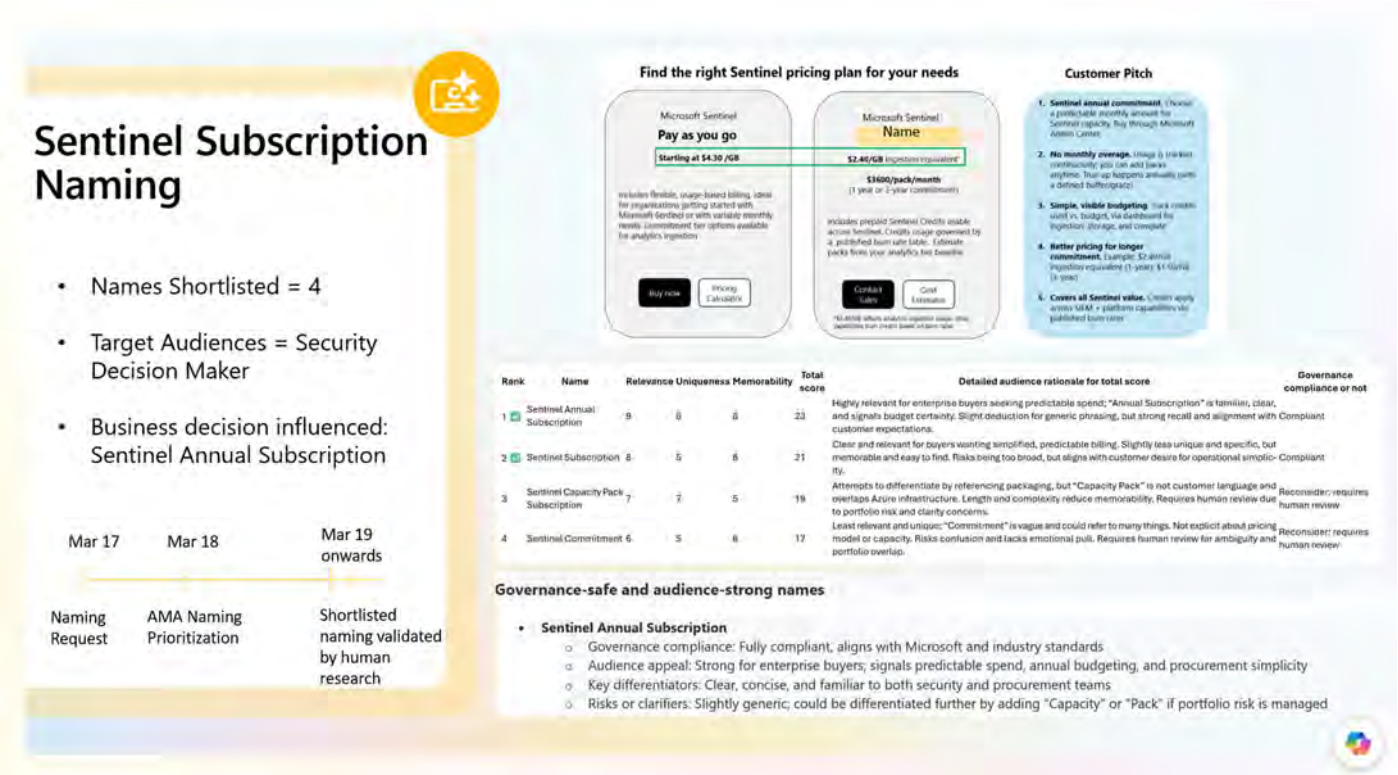



*Figure 5.*

Ad scripts, grounded in customer voice before production. Advertising scripts were historically developed through creative judgment, with customer reaction measured only after the ad had already been produced and aired. Using AMA, a team drafted an ad script and submitted it for simulated audience review before committing to production. The feedback identified specific lines that created emotional resonance and others that introduced friction or confusion for the target segment. The script was revised based on that feedback and customer voice shaped the final creative before a single dollar was spent on production, not after the campaign had already launched.

A messaging campaign that outperformed its target by 25%. A campaign team used AMA to test several messaging directions against simulated audience feedback before committing to a final creative approach. Early directions surfaced friction points around tone and emphasis that the team had not anticipated; later iterations addressed those points directly. The team refined the messaging through several rounds of simulated testing before launch. The resulting campaign delivered 25% above its sales target, a result the team attributes in part to the sharper messaging direction that emerged from testing multiple options against real customer voice before any one direction was locked in.

Across all three examples, the common thread is timing. Customer voice was present while the decision was still being formed, not introduced afterward to confirm a choice already made.

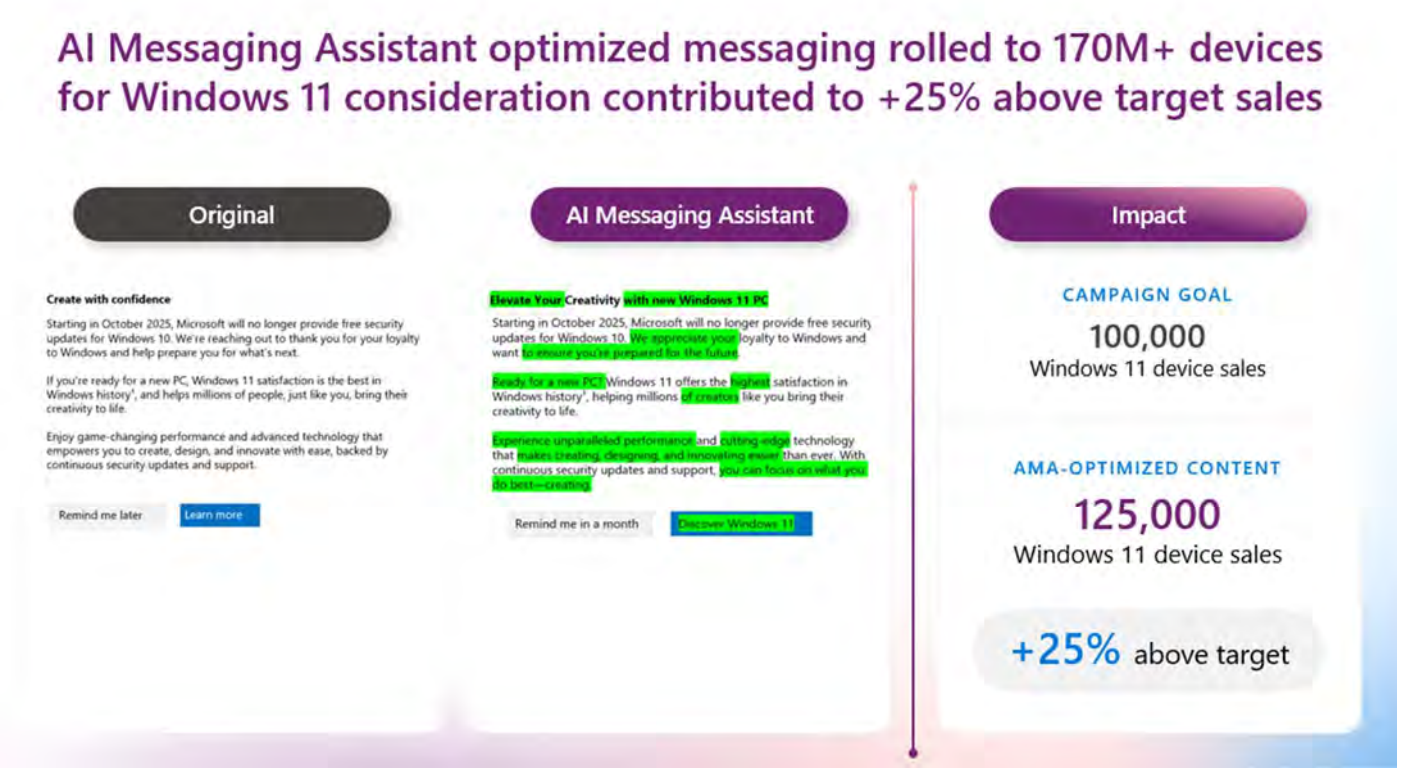



*Figure 6.*

## What the Evidence Shows

Accuracy benchmarking across hundreds of A/B test cases shows 90% directional alignment between AMA's simulated audience outputs and real research conclusions – sufficient to support confident directional decision-making across the majority of use cases.

The 10% where divergence occurs are informative: misalignments cluster around decision contexts underrepresented in the corpus, questions requiring recency sensitivity and specific quantitative claims requiring statistical validation. These patterns define the boundaries beyond which formal research remains essential.

Operationally, AMA delivered $10M in value within six months – across time savings, reduced research spend and faster time-to-market. Cycle time fell by 90%, the kind of compression illustrated in the naming and ad script examples above.

Formal research commissioning did not decline. It changed character – more targeted, more hypothesis-driven, more focused on genuinely uncertain territory. The sharpening effect is the clearest empirical evidence of augmentation in practice.

The hardest challenge was trust. Traceability was the mechanism: users who could see which research inputs informed a simulated perspective expressed substantially higher confidence than those receiving outputs without visible sourcing. The ability to trace a simulated perspective back to real customer voice is not a transparency feature. It is the epistemological basis for practitioner trust.

Copyright © **ESOMAR** 2026


## What This Means for Research Practice

Customer voice as a durable, reusable asset. Traditional research produces a one-time result – intelligence that expires at the boundary of the project that created it. AMA treats customer voice as a durable asset. Research conducted for one product launch surfaces when a different team asks a related question months later. A single study can inform hundreds of decisions over time. Every research investment keeps generating returns far beyond its original scope.

Research moves upstream. Conventional research validates after the frame is set. AMA puts customer voice in the room when the frame is still open – shaping direction rather than evaluating it after the fact. The difference between arriving after the frame is set and arriving before it is the difference between being consulted and being consequential. That is the permanent seat at the table. A 24/7 customer voice – present at every decision, always ready to be tapped.

A framework for responsible integration:

- Ground simulation in real research data
- Use simulation for direction, not confirmation
- Design for traceability
- Integrate with formal research, don't isolate from it
- Measure research reach, not just research speed

## Conclusion

The question facing market research is not whether AI will change how audiences are understood and simulated. It will. The question is whether the field manages that change in ways that preserve research validity, or sacrifices it for efficiency.

AMA offers one answer: simulated audiences grounded in validated real-world research, integrated into workflows designed to complement rather than replace formal studies, governed by clear boundaries, and built on traceability that keeps every output anchored to genuine customer voice.

The result is not automated research. It is extended research impact. Customer voice now reaches decisions it was never designed to inform. The R&I function's influence expands without its rigour being compromised.

For organisations exploring AI-enabled insight systems, AMA offers a pragmatic blueprint: disciplined integration, governed by research rigour, with human expertise at the centre of every decision about where simulation ends and primary research begins.

That boundary is not a limitation. It is the methodology.

## About the Authors

Romani Patel, Director Data Science, Microsoft, USA.
Kate Weymer, Director Data Science, Microsoft, USA.
Robert Graves, Senior Director, Microsoft, USA.

# When Grief Walks Into Work

## *From Mixed-Method Research to AI-Driven Tools That Support Grievers at Scale*

*Gina Pingitore, PhD*

*Pete Trujillo, MSAI*

Esomar

Office address:

Burgemeester Stramanweg 105

1101 AA, Amsterdam

Phone: +31 20 664 21 41

Email: info@Esomar.org

Website: www.Esomar.org

Publication Date: September 2026

ESOMAR Publication Series Volume S424 Congress 2026

ISBN 92-831-0337-8


Copyright © **ESOMAR** 2026

Copyright © **ESOMAR** 2026


# When Grief Walks Into Work

## *From Mixed-Method Research to AI-Driven Tools That Support Grievers at Scale*

*Gina Pingitore, PhD*

*Pete Trujillo, MSAI*

### Prior Research and Theoretical Foundation

Early grief models treated loss as a uniform progression through predictable stages. More recent frameworks – Dual Process Models, Meaning Reconstruction Theory, Worden's Four Tasks of Mourning and Boss's Ambiguous Loss – have significantly advanced the field by moving away from that assumption, recognising grief as dynamic, non-linear and shaped by the ongoing relationship between the bereaved and the deceased. This research builds on that foundation and extends it into the organisational context. Two people who lose a parent may grieve in entirely different ways depending on bond intensity, support networks, prior losses and concurrent stressors. The loss type is identical; the experience is not. Segmentation based solely on loss category systematically misrepresents the population. The finding that bond intensity matters more than loss type became the methodological cornerstone of everything that followed.

### Methodology: Four Questions That Shaped Every Decision

*Q1: How do you get people to tell the truth about something this traumatic?*

Grief carries stigma, exhaustion and workplace suppression. Getting honest data required a key design decision: grief was never the stated subject of the quantitative research. The program surveyed nearly 15,000 respondents, including employees and managers, as part of a multi-country study on how technology is reshaping work. Embedded within that survey was an abbreviated Holmes-Rahe Life Events Scale.

Respondents who indicated a qualifying loss were routed to a deeper grief support module – without ever being recruited to a grief study. Of nearly 15,000 respondents, 4,367 reported at least one qualifying loss, producing a grief subsample of unusual scale and credibility.

Two findings stand out. First, younger workers reported significantly higher grief intensity and greater work impairment than older cohorts, a pattern that held across all loss types ($p<0.001$). Organisations treating grief as primarily a concern for older employees are misreading their most affected population. Second, respondents in India reported substantially higher grief intensity and work impairment than those in North America, Europe or China. Three converging factors explain this: India has no statutory bereavement leave requirement; Hindu mourning rituals typically span 10–13 days of structured daily observance; and collectivist cultures frame grief as an ongoing obligation, not a personal journey toward closure. Multinational employers applying a single bereavement policy across geographies are not managing grief – they are managing the illusion of it.

The qualitative program, comprising more than 60 in-depth interviews and four focus groups, required a different approach. Grievers do not volunteer. Recruitment required partnering with trusted organisations, establishing genuine consent processes, and training interviewers in grief-informed listening. The methodological lesson: sensitivity is not a constraint on good research. It is a design input. When the topic carries trauma, the methodology must be guided by care.

### *Q2: How do you build personas that are emotionally true, not just demographically accurate?*

Standard persona methodology aggregates research into representative profiles. For grief, that approach produces a fiction. This is where AI became methodologically essential. The Grief Impact Illustrator generates synthetic employees with fully seeded life histories (e.g., demographics, family structure, support network, performance baseline) and a sequence of life events that establish the context each person carries into loss. Grief events are assigned using age-weighted probabilities derived from the quantitative research; bond intensity is drawn separately and independently. Demographics do not determine grief. Bond does. The AI produces the full spectrum of human response, then generates a narrative – a grief story written in language that makes an executive feel what a spreadsheet cannot. One synthetic employee, a 44-year-old engineering manager who lost his daughter, resigns in month eight. His replacement cost exceeds the combined productivity loss of his entire simulated team. The story creates the conversation. The conversation creates change.

### *Q3: How do you translate clinical frameworks into tools practitioners can actually use?*

Clinical architecture encodes how grief behaves over time, how it differs from depression and PTSD, and what interventions are appropriate at each stage. This architecture underlies three tools designed to operate at every level of the organisation.

*The Village Simulator* models productivity losses across an organisation using AI-enhanced synthetic employees with real-life histories. It captures both direct losses – the sustained output decline of the grieving employee – and ripple losses: the measurable productivity impact on teammates, managers and colleagues who absorb displaced work and navigate their own vicarious grief. It makes those losses visible and actionable in language that executives can respond to.

*CARRY* is a phase-aware, AI-powered companion built for two audiences through a single interface. For the griever: always available at 2 a.m. on a Sunday during the Year of Firsts, it helps find words, prepare for return-to-work conversations, manage specific workplace situations and understand what they are experiencing. A toggle transforms it into an entirely different tool for managers, colleagues and family members: role-specific guidance on what to say, how to check in without overstepping and how to sustain support over months, not days. When both the griever and their manager use CARRY simultaneously, they arrive at their next conversation from a shared clinical framework – with compatible language and a far better chance the exchange actually works.

*The Empathy Tutoring Tool* (ETT) is the next level of support for the people around the griever: a coach-supported practice environment designed to produce lasting changes in communication patterns. Users practice responding to real workplace grief scenarios using AI personas whose emotional states shift in real time based on their responses. Every response is scored across dimensions, including boundary respect, timeline awareness and trauma-informed language, with immediate feedback explaining why a response moved the emotional state in that direction. The distinction from

CARRY mirrors a familiar dynamic: self-help guides go only so far. Working with a coach is what it takes to truly internalise the skill. CARRY is the self-help. The ETT, used with a trained facilitator, is the coaching.

Together, these three tools form the core of a grief-responsive ecosystem. The Village Simulator makes the organisational impact visible. CARRY supports the griever and equips the people around them. The ETT builds the communication skills that determine whether those conversations actually land.

### *Q4: What does grief actually cost, and what helps?*

Grief does not stay in one seat. A ripple engine traces four mechanisms by which one person's grief spreads to colleagues who experienced no loss: manager cascade, decision latency, coverage burden and peer-support ripple. Together, these add approximately 17% to total productivity impact. For an organisation of 1,000 employees, the direct year-one productivity decline is approximately $500,000; the total annual impact, including ripple effects, is approximately $750,000. The interventions that move these numbers most are not expensive. Manager acknowledgment, structured check-ins at 30, 60 and 90 days, and flexible return-to-work arrangements reduce absenteeism and flight risk in measurable, modifiable ways. Unsupported grievers are two to three times more likely to leave within 18 months, at replacement costs of 50 to 150 percent of salary. The business case closes. The barrier is not the cost. It is the absence of language, confidence and organisational permission to act.

## Key Takeaways for Insights Professionals

Grief is a measurable workforce variable, not a personal interruption. Between 10% and 20% of any workforce is grieving at any given time, with quantifiable, attributable downstream impacts. This reframes grief from an HR edge case into a strategic workforce priority.

Demographics do not determine grief intensity; bond does. Segmentation based solely on loss type systematically misrepresents the population. Bond intensity must be measured, not assumed.

Disenfranchised losses carry the highest grief-unique cost. The losses organisations least recognise (friends, pets, pregnancies) generate the highest proportion of grief-attributable impact. Grief support programs create their greatest value precisely where organisational policy is most silent.

AI-enhanced personas grounded in primary research can change executive behaviour where reports cannot. The Grief Impact Illustrator demonstrates a replicable model for using generative AI to activate research findings in any domain where emotional distance, not lack of information, maintains the gap between insight and action.

Research insight without practice tools is a diagnosis without a treatment. The Grief Impact Illustrator makes cost visible. CARRY gives grievers language to articulate what they need. The ETT gives managers and colleagues the skills to show up well. Knowing what grief costs and knowing what to say about it are two different organisational capabilities. This research has built tools for both.

## The Ethical Frame

This paper was written at a time when technology companies were building products that simulate the dead. This program argues for a different use: technology that listens to what the living have told us and builds tools that help

organisations respond. The Grief Impact Illustrator does not simulate grief. It communicates – in language executives respond to – what real grievers told us in real research. The limitations are as visible as the outputs. AI amplifies human insight. It does not replace human presence.

## About the Authors

Gina Pingitore PhD, Head of Research Innovation, Impact Advisors and Cint, USA.

Pete Trujillo MSAI, Head of Product Development, Impact Advisors and Aerie, USA.

# Designing the Future of Sport

## Mapping Emerging Communities to Shape Creative Territories

*Sarah Laurier*

*Rodolphe Pal*

Esomar

Office address:

Burgemeester Stramanweg 105

1101 AA, Amsterdam

Phone: +31 20 664 21 41

Email: info@Esomar.org

Website: www.Esomar.org

Publication Date: September 2026

Esomar Publication Series Volume S424 Congress 2026

ISBN 92-831-0337-8

Copyright © **Esomar** 2026

# Copyright

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system of any nature, or transmitted or made available in any form or by any means, electronic, mechanical, photocopying, recording or otherwise, without the prior written permission of Esomar. Esomar will pursue copyright infringements.

In spite of careful preparation and editing, this publication may contain errors and imperfections. Authors, editors and Esomar do not accept any responsibility for the consequences that may arise as a result thereof. The views expressed by the authors in this publication do not necessarily represent the views of Esomar.

By the mere offering of any material to Esomar in order to be published, the author thereby guarantees:

- that the author - in line with the ICC/Esomar International Code of Marketing and Social Research– has obtained permission from clients and/ or third parties to present and publish the information contained in the material offered to Esomar;
- that the material offered to Esomar does not infringe on any right of any third party; and
- that the author shall defend Esomar and hold Esomar harmless from any claim of any third party based upon the publication by Esomar of the offered material.

Published by Esomar, Amsterdam,
The Netherlands
Edited by: Bel Kerkhoff-Parnell PhD.

Copyright © **Esomar** 2026

# About Esomar

Esomar is the champion of the insights and analytics sector. It is the business community for every sector professional. Founded in 1947, the global membership association is a network reaching over 40,000 professionals and 750+ companies in 130+ countries. We support individual and corporate members supplying or using insights by helping them raise ethical standards, grow the demand for insights, and improve its uses and applications by all decision-makers.

## About Esomar Membership

Esomar is open to everyone, all over the world, who believes that high quality research improves the way businesses make decisions. Our members are active in a wide range of industries and come from a variety of professional backgrounds, including research, marketing, advertising and media.

Membership benefits include the right to be listed in the Esomar Directories of Research Organisations and to use the Esomar Membership mark, plus access to a range of publications (either free of charge or with discount) and registration to all standard events, including the Annual Congress, at preferential Members' rates.

Members have the opportunity to attend and speak at conferences or take part in workshops. At all events the emphasis is on exchanging ideas, learning about latest developments and best practice and networking with other professionals in marketing, advertising and research. CONGRESS is our flagship event, attracting over 1,000 people, with a full programme of original papers and keynote speakers, plus a highly successful trade exhibition. Full details on latest membership are available online at www.Esomar.org.

# Contact us

Esomar

Esomar Office:

Burgemeester Stramanweg 105

1101 AA, Amsterdam

The Netherlands

Tel.: +31 20 664 21 41

Email: info@Esomar.org

Website: www.Esomar.org

Copyright © **Esomar** 2026

Congress 2026

# Designing the Future of Sport

## Mapping Emerging Communities to Shape Creative Territories

*Sarah Laurier*

*Rodolphe Pal*

### Introduction

The future of sport is no longer defined solely by performance metrics, athlete profiles or participation rates. Across global digital cultures, new communities are reshaping how sport is practiced, expressed, consumed and imagined. These communities are driven as much by identity, aesthetics, belonging and self-expression as they are by competition and performance. For organisations operating in the sports industry, this transformation presents a challenge. Traditional frameworks have long relied on categories, disciplines, demographics and consumer segments to understand audiences. Yet these models increasingly struggle to explain emerging behaviours and aspirations. Over the past five years, we have collaborated on a continuous social and cultural intelligence programme designed to explore the future of sport through the lens of communities. What began as a foresight initiative evolved into a living cultural ecosystem capable of identifying emerging territories, decoding weak signals and translating cultural movements into actionable inspiration for strategy, innovation and design. This paper explores that evolution. It describes how social and cultural intelligence moved from an illustrative function to a strategic and creative capability. It also demonstrates how understanding communities, rather than simply tracking “consumers”, can help organisations imagine and design future possibilities.

### Why Traditional Segmentation Was No Longer Enough

For decades, sports markets have been understood through familiar lenses. Audiences were segmented according to age, gender, income, participation frequency or preferred discipline. Running, cycling, football, fitness, hiking and swimming became the primary categories through which companies analysed demand and organised innovation. While these frameworks remain useful, they have become increasingly insufficient for understanding contemporary cultural dynamics. Digital platforms have transformed how people discover practices, build identities and connect with others. Communities now emerge around shared values, aesthetics, narratives and aspirations that often transcend geography, demographics and even specific sports disciplines. A consumer living in New York may feel culturally closer to a trail running creator in Paris than to a neighbour practicing the same sport. Shared references, visual codes and cultural aspirations increasingly create stronger bonds than physical proximity.

This shift is particularly visible among younger generations, but extends far beyond them. Sport today intersects with fashion, wellness, creativity, outdoor exploration, sustainability, urban culture, gaming and personal development. Individuals navigate multiple identities simultaneously, creating hybrid communities that cannot be adequately captured through traditional segmentation models. As a result, a fundamental question emerged: *Rather than asking which sports were growing, should we instead ask which communities were shaping the future of sport?* This question became the foundation of our approach.

### From Sports Categories to Cultural Communities

To answer this question, the project shifted focus away from categories and toward communities. The objective was not simply to identify what people were doing. It was to understand why they were doing it, what emotional and symbolic meanings they attached to practices and what aspirations they projected through sport. The research explored digital ecosystems across Europe, North America and Asia. Millions of

conversations, visual expressions, creators, content ecosystems and online communities were observed through a social intelligence lens. What quickly emerged was the existence of cultural territories that cut across traditional categories:

- Outdoor enthusiasts were increasingly blending performance with spirituality and mindfulness;
- Urban sports communities were borrowing codes from fashion, music and street culture;
- Wellness-oriented practitioners were redefining notions of achievement and success;
- Aesthetic expression was becoming as important as performance for many emerging groups.

These communities often remained relatively small in scale. Yet their influence extended far beyond their size. They acted as cultural laboratories where new rituals, values and expectations emerged before diffusing into the mainstream. Importantly, many of these groups remained largely invisible within traditional quantitative frameworks. Their statistical weight was limited, but their cultural influence was considerable. The project therefore prioritised cultural significance over scale. It focused on identifying communities capable of shaping future narratives rather than simply measuring existing behaviours.

## Building a Living Cultural Cartography

At the heart of the initiative was the creation of a dynamic cultural cartography. Unlike traditional segmentation models, the cartography was designed as a living structure rather than a static framework. The methodology combined social intelligence, expert cultural interpretation and foresight analysis. Digital conversations, visual platforms, community ecosystems, creators and emerging cultural spaces were continuously analysed to identify clusters organised around shared aspirations, practices, aesthetics and symbolic references. The objective was not to produce another trend report, but to create an evolving map capable of revealing emerging opportunities and future directions. This living cartography allowed new communities to appear, evolve, merge or disappear over time. Rather than freezing reality into predefined segments, it embraced movement as a core principle. Weak signals occupied a central role within the framework. Traditionally, weak signals are often treated as isolated observations or interesting anomalies. In this project, they became strategic assets. Emerging practices, aesthetic shifts, niche creators and peripheral communities were analysed not because they represented the majority, but because they revealed potential futures.

Over time, the cartography became a reference system that could be revisited, updated and interrogated to address new strategic questions. More recently, advances in AI have expanded the possibilities of this ecosystem. Conversational interfaces connected to the community framework allow teams to explore emerging aspirations, test hypotheses and navigate community narratives in a more intuitive way. Rather than replacing human interpretation, these tools extend the ability to engage with cultural intelligence and make exploration more accessible across the organisation.

## From Foresight to Creation

One of the most significant evolutions of the project occurred when insights began moving beyond strategic diagnosis and into creative development. Many foresight programmes successfully identify trends, but struggle to influence concrete outputs. Findings often remain confined to presentations, reports and strategy discussions. From the outset, Decathlon sought a different outcome. The ambition was not simply to understand future communities; it was to translate cultural understanding into creative possibilities. To achieve this, the project brought together researchers, strategists, designers, stylists and creative experts. This interdisciplinary collaboration proved essential. Communities do not inspire organisations simply because they exist; their influence emerges when their values, narratives, aesthetics and aspirations are translated into meaningful opportunities. Through collaborative workshops and creative sessions, cultural insights became territory platforms capable of informing design thinking:

- Weak signals became visual inspirations;
- Community narratives became aesthetic directions;
- Emerging aspirations became opportunities for future products, services and experiences.

The objective was not to predict the next successful product, but rather to help Decathlon's global design community imagine what the future of sport could become. This creative translation process transformed cultural intelligence into a practical innovation tool.

## Why Formats Matter as Much as Insights

Another major learning emerged throughout the five years of collaboration. Insight only creates value when people engage with it. Many organisations focus significant resources on producing knowledge while underestimating the importance of how that knowledge is experienced. The challenge is not always generating insight, but rather making insight desirable. As the project evolved, formats became as important as findings. Traditional reports were no longer sufficient to create engagement and activation. Instead, a broader ecosystem of formats was developed:

- Editorial magazines offered immersive explorations of emerging territories;
- Living cultural books documented the evolution of communities over time;
- Pulse formats highlighted recent shifts and weak signals;
- Interviews and conversations brought external perspectives into the discussion;
- Internal talks and show-style presentations created moments of collective reflection and debate.

These formats were not simply communication tools; they became an integral part of the methodology. They helped teams revisit ideas, challenge assumptions, discuss implications and collectively build meaning around the findings. The conversations generated by these formats often became as valuable as the insights themselves. As a result, the project evolved from a research programme into a living cultural dialogue within the organisation.

## Business Implications

Over five years, the initiative generated value across multiple functions:

- For strategy teams, it provided a richer understanding of emerging cultural dynamics shaping the future of sport;
- For innovation teams, it revealed new opportunity spaces and future territories for exploration;
- For design teams, it delivered concrete sources of inspiration grounded in real communities and practices;
- For brand teams, it helped identify evolving narratives, symbols and cultural references capable of strengthening relevance.

Most importantly, it encouraged the organisation to move beyond category thinking and embrace a more fluid understanding of culture, aspiration and community.

## Implications for the Insights Industry

This case reflects a broader transformation occurring within the insights profession.

### *The Role of Social Intelligence is Evolving*

What was once primarily an observational and illustrative tool is becoming a strategic capability with the potential to inform decisions, inspire innovation and shape creative processes.

### *The Role of The Researcher is Evolving As Well*

Researchers are no longer only responsible for measuring reality. They increasingly help organisations interpret cultural change, identify emerging possibilities and navigate uncertainty.

### *This Shift Requires New Forms of Collaboration*

Copyright © **Esomar** 2026


Insight must work alongside strategy, design, innovation and creative functions. It also requires new forms of storytelling.
Analysis alone is not enough; Insights must be translated into formats, narratives and experiences that people can engage with and act upon. Ultimately, the future of insight may depend less on our ability to collect information and more on our ability to transform cultural understanding into shared vision.

## Conclusion

Over five years of collaboration, social and cultural intelligence evolved from an observational practice into a creative foresight system. What began as a trend monitoring initiative became a living ecosystem for understanding communities, identifying emerging cultural territories and inspiring future possibilities. The project demonstrated that the future of sport is not simply shaped by products, disciplines or demographics:

- It is shaped by communities;
- By aspirations;
- By identities;
- By cultural affiliations.

Understanding these forces requires moving beyond static segmentation models and embracing more dynamic approaches capable of capturing movement, emergence and transformation. It also requires recognising that insight does not end with analysis; its true value emerges when it informs imagination, stimulates creativity and helps organisations design what comes next.

In a world where communities evolve faster than categories, cultural intelligence may become one of the most powerful tools available to organisations seeking to shape the future rather than simply observe it.

## About the Authors

Sarah Laurier, IFOP Group, Paris, France.
Rodolphe Pal, IFOP Group, Paris, France.

# Not Another Gen-Z Brief

## How to Form a New Contract with Older Adults

*Scott Percival*

*Drago Djourov*

## Esomar

Office address:

Burgemeester Stramanweg 105

1101 AA, Amsterdam

Phone: +31 20 664 21 41

Email: info@Esomar.org

Website: www.Esomar.org

Congress 2026
Copyright © **ESOMAR** 2026

Copyright © **ESOMAR** 2026


# Not Another Gen-Z Brief

## How to Form a New Contract with Older Adults

*Scott Percival*

*Drago Djourov*

### Introduction

*This paper is part of the NEXT50 study by Potluck – the research-and-collaboration engine behind the Potluck podcast, through which Scott and Drago explore the people, culture and brands of Asia.*

Takashi owns seven guitars. On any given week he might be at the gym for a Zumba class, running a meet-up where students, young professionals and older adults swap advice, or doing a part-time job for a charity. He still finds time to post and share with his friends on Instagram, Facebook and Threads – all at once. Travelling and dining out regularly are further passions.

Takashi is 71.

OK, he manages a couple of health conditions, his wife has now passed away, and he only gets to see his son once a month, but he's busier and more fulfilled than he's ever been. We open with him because most brands, and many of us in research, would not have pictured a 71-year-old when reading that first paragraph. That gap – between who older adults actually are and who we assume them to be – is the subject of this paper. Takashi is not an outlier. He is symbolic of something we found across three Asian markets: a generation in the middle of a quiet but profound reorganisation of identity, purpose and relationship.

We came to call this shift "The New Contract".

For most of his working life, Takashi lived under a different agreement. He joined a large Japanese corporation straight out of university and never left – a reciprocal deal in which hard work bought income, status and a sense of being in it. Then, as he aged, he was eased toward the back office: less needed, less central and finally retired. The identity he had worn for forty years simply… stopped. That is The Old Contract: an industrial-era bargain that promised a clean handover from a working life to a funded, restful retirement – and that, for this generation, has largely broken down.

What happens next is the interesting part. The loss, the sense of being "out of it" can be destabilising. Yet, it can also be liberating. Freed from a contract written by employers and society, older adults like Takashi begin drafting a new one – this time with themselves. Its terms are based more on freedom, connection and growth. To be "in it" once more.

The New Contract is, at its core, a story of metamorphosis: not decline, not nostalgia, but a fundamental rewrite of what the next chapter is for. And it demands an equivalent rewrite in how brands – and the researchers who advise them – see, study and serve this group.

Copyright © **ESOMAR** 2026


## Framing the Opportunity: The Audience Hiding in Plain Sight

Across our three markets, in the last three months alone, 80% of over-50s had bought something from an e-commerce site, 78% had used social media, 77% had used mobile banking and 70% had made a digital payment in-store. Half had used a ride-hailing app; nearly half had ordered groceries or food via delivery apps. These figures barely move between Singapore, Japan and Indonesia. On the evidence of their actual behaviour, older adults in Asia are not waiting offline for the modern economy to slow down for them – they are already in it.

They also have the means. Around one in four people in Asia is now over 50 (UN DESA, 2024), a share that keeps climbing as the region ages and birth rates fall – with the proportion even higher in ageing markets including South Korea, Mainland China, Hong Kong and Taiwan. The over-60 cohort is expected to account for one third of consumption growth in Asia by 2030 (McKinsey Global Institute, 2021). In Japan, households headed by the over-60s alone hold around two-thirds (63.5%) of Japan's household financial assets (Statistics Bureau of Japan, 2019), and households headed by 50–59-year-olds outspend those headed by under-29s by more than 60% (Statistics Bureau of Japan, 2024).

But if we consider their spending power against where marketing money flows, the picture inverts: by most estimates the 50+ receive only 5–10% of marketing investment while accounting for more than half of spending power (Emsley, 2020; Rella, 2023; AARP, n.d.).

As part of a broader marketing industry, we pour budget, creativity and research attention into the young, then express surprise that older consumers feel unseen. As one of our Japanese respondents put it: "Brands don't target us. Ads always show young people."

The point is not that brands should abandon younger audiences. It is that the lifetime value of an older customer is routinely and radically underestimated – and that the most populous, asset-rich and discerning stage of adult life is too often treated as an afterthought or excluded from the research brief entirely.

Properly understood, lifetime value does not taper at retirement. Dr Shakuto (Professor of Anthropology, University of Sydney) spoke to this in our podcast interview:

*"One misconception about retirement is that it's seen as the end of one's life course. But actually, because of the increasing longevity, a lot of people have about 20 to 30 years more where their body is healthy – they can do a lot of things before and after their retirement."*

Indeed, health-spans are lengthening and working lives extending or being re-imagined, so people stay active in categories for far longer; their spending power frequently peaks after 50 rather than declining and they re-enter categories they once left, returning to a youthful passion as a later-life pursuit now backed by time and means, or simply having fresh needs as they transition across numerous life-stage inflexion points. They even shape the choices of the young around them (see "Brands Revalued" below) – while remaining valuable customers in their own right.

Each is a distinct way the value of an older customer can be extended and maximised rather than written off. As populations age across not only Asia (the focus of our research) – but also Europe and North America – capturing that value is fast becoming a central commercial question, not a niche one.

Copyright © **ESOMAR** 2026


## About the Research

The findings come from a self-funded, mixed-method programme conducted over more than a year, in two deliberate phases.

### *Phase 1 – Qualitative Exploration (Japan and Singapore)*

We built the New Contract thinking from the ground up, through:

- 12 in-home immersive ethnographic interviews with older adults in Japan and Singapore. Deliberately targeting a cross-section of individuals either side of the official retirement age in both countries, and representing different relationships with work.
- 4 expert interviews with academics and specialists in gerontology, ageing and the "silver economy"
- 5 dedicated podcast episodes with practitioners across the older-adult space (see Potluck Podcast, 'The New Contract meets…' series)
- Extensive desk research spanning academic, cultural and commercial sources.

### *Phase 2 – Quantitative Validation (Japan, Singapore and Indonesia)*

We then measured the prevalence of behaviours, attitudes and beliefs that underpinned our emerging framework with an online survey of 1,500 adults aged 50-75 (n=500 per market) – whilst simultaneously developing a NEXT50 segmentation of older adults.

The survey focused specifically on finance and travel as two core categories of interest, whilst also spotlighting wider behaviours, attitudes and beliefs

Indonesia was added deliberately at this stage so the findings would carry both a developed-market and a developing-market perspective, rather than resting on two mature (and more aged) economies alone.

Unless attributed otherwise, all statistics in this paper are drawn from this survey. The programme also produced a brand-equity tool, the NEXT50 Brand Action Index, which sits outside the scope of this paper.

In essence, it is an evolving thought-leadership study which is leading us to explore the older adult opportunity with domain experts, creative agencies and brand leaders alike.

## The New Contract: A Framework in Six Dimensions

We wanted to make The New Contract more than just a metaphor for describing what is shifting for older adults. Six interconnected dimensions were identified, each capturing a different facet of the metamorphosis.
Together they form a strategic framework: a way to understand later life and a springboard for product, brand and communications thinking.

We bring each to life below with defining statistics, Takashi's lived experience (alongside other voices from our fieldwork), and a thought-starter for how brands could act upon it.

### *1. 3C's Revolution – Collaboration, Contribution, Community*

The first dimension is central to understanding the New Contract shift. As working life recedes or takes on new forms, the organising logic of life flips: from competition, comparison and career to collaboration, contribution and community. The value people seek stops being something to accumulate and becomes something to give – and, crucially, this is experienced not simply as charity or winding down but as a richer, more reciprocal way of being useful.

Where the old working life was more transactional, self-centred and short-term, this one is more relational and open-ended. In our survey, 63% of over-50s feel a renewed sense of purpose and want to contribute to people and causes they believe in – with only 7% disagreeing.

For Takashi this is the whole point of Ossan Rental[1] – that's the "Rent an Uncle" service he's enthusiastically part of. "It's great to be able to help. To be useful. I can't quit doing this", he told us, describing a fifty-something client he'd coached through a job interview who later wrote to say she'd got the role. "When I was in HR, I learnt a lot", he added, "so now it's like I'm returning the favour" – very much living out the Japanese concept of *ongaeshi*, or repaying kindness.

Brand thought-starter: give older customers a role, not just a discount. The brands that resonate here behave less like vendors and more like platforms for contribution – creating ways for customers to share expertise, join shared endeavours or give back to their community, and treating that generosity as part of the value exchange rather than a marketing add-on.

### *2. Different Energy*

Popular culture, and in turn, brands, love to celebrate the white-hot energy of youth and quietly dismiss the typically slower energy that comes with age. But what we heard was not diminishment – it was recalibration: a more deliberate and more authentic energy that many experience as an upgrade rather than a loss.

It shows up as more focus, more self-knowledge and far less to prove. Three in four over-50s (76%) say they are more comfortable simply being themselves now – an audience with less front and veneer than in their younger years. Whilst two-thirds (65%) say standing still isn't an option and they want to keep learning new skills. Lifelong learning isn't simply a buzzword – but a means to look forward, stay sharp and even consciously fight off any risk of mental decline.

Takashi frames it as a loss of pretence: "There's no desire to gain new honours or money now, so whatever I encounter, I can react in an ordinary, relaxed way." He trains at the gym most days of the week, working around treatment for a health condition – not as a battle, simply as part of a life he intends to keep living forward. A Singaporean business owner in his sixties described the same energy from the other side: "Cognitively I'm still in hyperdrive… so let me consolidate my wisdom and pass it on."

Brand thought-starter: Reframe "slowing down" as a strength, not a limitation. Build experiences and propositions around depth, quality and considered attention rather than speed and efficiency – and consider where a calmer, wiser, more deliberate register might be exactly what an over-stimulated, burnt-out culture is quietly craving. Internally, think of the role "Chief Wisdom Officers" could play in grounding business decisions, or how customer-facing sales and support staff could better mirror the energy of older-adult customers.

### *3. New (and Old) Allies*

[1] https://rental-japanese.com/ossan-rental-japan/

The New Contract is intensely social. Older adults are forging new friendships while deepening and rekindling old ones – often after careers that left little room for relationships beyond the workplace, where the "glue" of shared work can dissolve as soon as it ends.

What follows is a deliberate rebuilding of social circles: frequently leaner, but more freely chosen, and prized precisely because they reflect who someone is rather than what they did. It's also a chance to re-appraise and strengthen relationships with spouses, children and long-lasting friends.

And running counter to the prevailing "aging = loneliness" narrative, 59% of our respondents say they have formed meaningful new friendships in the past few years – though the same openness can expose them to romance scams and predatory online networks, a real and rising downside.

Takashi's networks now span generations and genders. He organises a monthly lunch with a group from his gym: "It's a group of ladies in their 60s, 70s and 80s, and there's me, a guy. I like hanging out with them because their conversation is more interesting than the guys'." A dynamic he couldn't have imagined in his former life as the corporate salaryman.

Brand thought-starter: help customers build leaner, richer social circles – and protect them while they do it. There is room to host or enable genuine connection around shared interests, hobbies and life-stage moments, and, equally, to provide the guidance and guardrails that keep new social and digital spaces safe from scams and isolation.

### *4. Aligning Algorithms*

If social media has hardened generational echo-chambers, many older adults are actively trying to permeate them – reaching toward younger colleagues, adult children and grandchildren, and the wider culture they refuse to be locked out of.

Nearly half (49%) want more interaction with younger generations but find it genuinely hard to engineer, especially as services become more remote and digitised; 65% say keeping up with new technology makes them feel more relevant, even as half (50%) feel most new tech isn't designed with people their age in mind.

The relationship with technology is therefore not aversion but a wary, effortful engagement: they want in, on terms that respect them rather than patronise them.

"Talking with young people stimulates us too", Takashi says of his mentees. "Opportunities to meet them become rarer as we get older." On keeping pace with new tools, he is honest about the limits but clear about the motive: "At the very least, I want to be able to participate in the discussion."

### *Brand Thought-Starter*

Be the bridge between generations. The opportunity is to design genuine two-way touchpoints – moments where older and younger customers exchange something of value – and to build products and interfaces that invite older users into the conversation, rather than signalling at-a-glance that it wasn't built for them. Notably, in our NEXT50 Brand Index – we found that Google was one of the top-performing brands with over-50s regionally.

## *5. Brands Revalued*

Brand relationships do not disappear in later life; they concentrate. As status-signalling falls away, consumption becomes more intentional – we find that brands matter more, but in fewer places, standing out as "beacons" earned through quality, reliability, craftsmanship and the human touch more than logo or aspiration.

The most valued brands tend to cluster at two poles: necessities – the categories still essential to daily life, or becoming more so, such as health and wellness – and passion points, where a hobby, ritual or long-held desire makes a brand genuinely meaningful. They also increasingly value brands that enable experiences and memory-making, over material acquisition.

For Takashi the passion pole means Gibson guitars; for someone else it might be the watch finally bought after decades of coveting it, or a whisky or face cream that marks a habitual, well-earned indulgence.

The older adult still cares about brands but cares far less about what you think of their choices. Indeed, three-quarters (75%) of over-50s no longer feel pressure to buy brands that look good to others, and 79% say that as long as quality is good, they don't need to buy expensive brands. Crucially, this is not conservatism dressed up: 54% still actively like trying new brands and products. This is an audience still very much drawn to novelty, and open to switching.

Takashi captures the recalibration exactly: "For me, a brand is a sign of quality. But I don't go around thinking, 'Look, I'm carrying a Gucci bag.'" The brands that survive earn it through depth: he has shopped at the same New Yorker store for over twenty years – "the store manager there is a friend". This shift is supported by a young-at-heart Singaporean grandmother, who told us, "I don't go for the very expensive brands anymore, but I would still buy things like Coach or Michael Kors… I would rather spend money on travel – it's my guilty pleasure… If it's nearby I don't mind budget airlines. Longer haul would still be Singapore Airlines."

## *Brand Influence on the Young*

We tend to assume influence runs one way – younger consumers set the trends and older ones follow. It drives much of why brands view the young as their aspirational audience and trust that relevance will trickle down to everyone else.

Our data challenges that comfortable logic. In the past two years, 77% of over-50s have recommended a brand or product to someone from a younger generation (be it colleagues or relatives), most of them across two or more categories – and the strongest are precisely the more considered, higher-stakes ones: travel (44%), technology (38%) and investments (33%).

These are not passive consumers at the end of the line but active advisors whose judgement younger buyers seek out. Win them, and influence can flow the other way too.

It is notably the finding our clients tend to respond to fastest, because it recasts the older adult from a drain on relevance into a route to the very audiences that brands already covet.

Brand thought-starter: earn one of the valued places that still matter – and recognise the customer's reach. Compete on craftsmanship, reliability and genuine human service rather than status or novelty for its own sake; deepen the relationship in the categories where you already have permission; and treat a satisfied older customer as an active advocate to younger buyers, not an end-of-line consumer.

Copyright © **ESOMAR** 2026


### *6. Future Forward*

We instinctively associate older adults with the past – not the future. Yet the New Contract is markedly future-facing, typically in a mode of "active hope" that holds regret and optimism at once: regret, often, for a planet and a society they feel their generation didn't tend well enough; optimism in the form of legacy, contribution and a determination to enjoy whatever time remains.

Far from disengaging, many are rewriting what the next chapter looks and feels like. Nearly three quarters (73%) are concerned about the environmental and social issues facing future generations, and 77% agree that, while they worry about money, they want to enjoy life while they still can.

Takashi's regret is generational and his response is practical – he now chairs an NPO promoting disaster preparedness: "We were really wasteful… I think we could have avoided this situation." On money and legacy he is unsentimental: "There's no point in being so stingy and hoarding everything." He is not alone in pairing penitence with purpose. "Yes, we made mistakes – we admit to it", a Singaporean respondent told us, addressing the young directly. "But don't do the blame game… can we find a middle ground where we meet and move forward? We will repair our sins, with whatever wisdom we have left, and make a better world for you."

Brand thought-starter: speak to hope and legacy, not decline. Acknowledge the optimism and growth mindset that brands routinely overlook in this group; support their drive to keep learning and to leave things better than they found them; and don't shy from the honest "what next?" conversation – handled with candour rather than gloom, it can be a powerful source of relevance. Nostalgia isn't the only lever to pull here.

## A Scorecard for Brands

The six dimensions can serve as a diagnostic lens. Brands can use the framework to ask, honestly, how they show up against the shift their older customers are living through – treating each dimension less as a box to score than as a prompt for spotting where they are, and aren't, meeting the New Contract.

In turn, it can serve as a springboard for ideation around everything from new product development to creative ideas for brand communications.

| Dimension | The shift underway | Ask of your brand |
|---|---|---|
| **1. 3C's revolution** | From accumulation to contribution | Do we give customers a way to contribute to causes they care about? Do our products and services enable genuine collaboration and community? |
| **2. Different energy** | From speed to deliberate depth | Does our experience allow a slower, more deliberate pace? Do we lead with quality and mindfulness over speed and efficiency? |
| **3. New (& old) allies** | From fixed networks to reimagined ones | Do we help customers form new connections — and reconnect old |

| | | |
|---|---|---|
| | | ones? Do we make expanding their social world feel safe and guided? |
| **4. Aligning algorithms** | From generational silos to bridges | Are we actively bridging generational divides in how we communicate? Do we create touchpoints that serve older and younger customers at once? |
| **5. Brands revalued** | From broader, shallow ties to fewer yet deeper ones | Have we shifted from selling acquisition to enabling experience? Do we lead with craftsmanship, durability and a human touch? |
| **6. Future forward** | From nostalgia to active hope and altruism | Does our messaging reflect their optimism and growth mindset? Do we enable their drive to keep learning and leave something better? |

*Table 1.*

## The NEXT50 Segmentation: Same Shift, Different Speeds

Of course, the over-50s are not a monolith, and treating them as one is precisely the habit the New Contract is seeking to break. As Dr Helen Ko, Associate Professor of Gerontology at the Singapore University of Social Sciences, told us: "The key concept I always teach my students is heterogeneity. People tend to lump older adults together. But they are highly, highly different." If anything, they are more heterogeneous than younger cohorts, having had decades longer to diverge.

So, the two halves of this research work together.

The New Contract describes what is shifting for everyone; the segmentation describes who has moved furthest along that shift, and how they respond to it.

We deliberately sought to create a psychographic segmentation that sits beyond more basic criteria such as age band, life stage or income group – something more overarching, grounded in the mindsets older adults hold, that brands could apply flexibly across very different categories. To build it, we ran a k-means cluster analysis on the survey's life-stage attitudes – toward learning, technology, purpose, social connection, money and trying new things. Three balanced, distinct mindsets emerged, separating chiefly along three dimensions: confidence, agency, and openness to change.

### *Embracers (31% Regionally) – The Group to Excit*

Confident, forward-looking and socially active, Embracers are the older adults brands most readily recognise – or aspire to align with. They index extremely high on agency: among them, 93% feel a renewed sense of purpose, 95% say keeping up with technology makes them feel relevant, 91% like trying new brands, and 71% are net happier than a decade ago. They are travellers, culture and thrill-seekers, tech adopters and active recommenders. They are also the heartland of the Indonesian sample, but we see meaningful pockets in Singapore and Japan – with Takashi as one of them.

Copyright © **ESOMAR** 2026


### *Realists (40% Regionally) – The Group to Convince*

The largest segment, and in our view the greatest commercial growth opportunity. Realists are socially connected and purposeful but pragmatic: investors with an eye on financial security, open to travel and to new brands (56% like trying new products) yet waiting for a compelling enough reason to act. 82% are comfortable being themselves and 72% feel a renewed sense of purpose – close behind the Embracers – but they are markedly less likely to strongly agree with anything. The same measured streak shows in how they travel: 63% (the highest of any segment) are keen to visit the historical staples, while 46% are ready to try new food along the way. But caution balances their openness – 70% worry about scams, and 74% worry about money even as they're determined to enjoy life while they can. They are won less by excitement, and more by security, social proof and planning support – the levers most likely to move a careful, capable group that simply hasn't been given the nudge. They form the spine of the Singapore sample.

### *Unsettled (29% Regionally) – The Group to Reassure.*

The least confident and most inward-looking segment, and a necessary reality check on the New Contract's optimism: this stage brings upheaval in identity, work, relationships and finances, and not everyone lands on their feet.

The Unsettled are more homebound, less tech-confident and more socially withdrawn; only 20% feel happier than a decade ago, and just 13% are drawn to new brands – yet 62% carry a constant worry about scams and money. They are often invisible to brands entirely, despite representing real unmet needs around reassurance, simplicity and genuine community. They are also the majority mindset in Japan.

Profiled across the attitudes that define them, the mindsets pull sharply apart:

| Mindset | Embracers (31%)<br>Forward-facing, "in it" | Realists (40%)<br>Sensible and pragmatic | Unsettled (29%)<br>Holding back, "out of it" |
|---|---|---|---|
| **Positivity** | 71% | 48% | 20% |
| **Self-fulfilment** | 92% | 82% | 51% |
| **Purpose** | 93% | 72% | 15% |
| **Sense of growth/ desire to develop new skills** | 94% | 78% | 16% |
| **New social connection** | 90% | 69% | 12% |
| **Openness to brand trial** | 91% | 56% | 13% |
| **Openness to new tech** | 95% | 65% | 35% |
| **Societal orientation** | 96% | 72% | 52% |
| **Pleasure-seeking/ YOLO** | 93% | 74% | 62% |
| **Digital safety concerns** | 86% | 70% | 62% |

*Table 2.*

*Note: the attitude statements which define each row form a segment "typing tool" which any brand can apply to their own category and customer base.*

Two patterns are worth dwelling on. First, on the agency items – purpose, learning, novelty, technology – the gap between Embracers and Unsettled is a chasm. The starkest illustration of why one lens isn't enough sits inside a single statistic; asked whether they like trying new brands, 91% of Embracers agree, but only 13% of the Unsettled do.
Second, on the worry items – money and scams – the segments converge a degree closer: even confident Embracers carry real concern. Liberation and anxiety, in other words, are not opposites here; they coexist.

A further signal comes from outside this table. On our NEXT50 Brand Index – the brand-equity tool that sits beyond this paper's scope – the same divide holds: Embracers are markedly more brand-engaged, awarding higher scores to brands overall, while the Unsettled rate brands significantly lower across every pillar (brand trust, product relevance and communications connection).

Mindset, then, shapes not only how older adults live, but how generously they appraise the brands competing for them.

The segmentation also refuses to behave like a single regional audience, because the mix differs markedly by market:

| Segment | Overall | Indonesia | Japan | Singapore |
|---|---|---|---|---|
| **Embracers** | 31% | 48% | 16% | 28% |
| **Realists** | 40% | 41% | 28% | 52% |
| **Unsettled** | 29% | 11% | 56% | 20% |

*Table 3.*

This is the clearest proof that the variable predicting mindset is not age alone, but the context and culture they live in. And the contrasts make intuitive sense market by market:

Indonesia is the aspirational, up-and-coming market of the three. Overall affluence and individual spending power may be comparatively lower than in Japan or Singapore, but those with the means to live an active, travel-rich, experiential life embody exactly the Embracer outlook: living for tomorrow and enjoying the fruits of their labour.

Singapore skews Realist, and pragmatically so. There is real affluence here and ample opportunity to travel and enjoy later life – but it sits alongside a balanced, measured disposition, in keeping with a nation that built itself from very little and tends to plan rather than splurge.

Japan, the most mature ageing society of the three and the one that has lived with these dynamics longest, skews Unsettled – though that label needs to be viewed with some care. With arguably less of a collective or intergenerational safety net to lean on, more is left to the individual to chart the road ahead; and against a backdrop of a long-stagnant economy and a weak yen, "Unsettled"" often reflects an understandable uncertainty about what comes next rather than outright pessimism.

And the segment mix will differ further inside a brand's own category and customer base. The practical implication is that brands should apply the segmentation to their own audience and their own markets, rather than assuming the regional average describes the customers in front of them. Layering on the segmentation with category drivers, usage-habits and brand perceptions is where further value can be unlocked.

Translated into action, each segment presents different implications from a brand, product and communications lens:

Copyright © **ESOMAR** 2026


| Lens | Embracers (31%) – excite | Realists (40%) –convince | Unsettled (29%) – reassure |
|---|---|---|---|
| **Brand** | Invite co-creation; treat as ambassadors | Lead with security and social proof | Lead with trust, simplicity and human help |
| **Product** | New experiences, travel, learning, premium | Planning support, bundles, low-risk trials | Reassurance, accessibility, scam protection |
| **Comms** | Aspirational, peer-to-peer, trendsetting | Evidence, testimonials, "people like me" | Plain, warm, community-anchored |

*Table 4.*

## Conclusion: Ageing Audiences, or Ageing Thinking?

Used together, the New Contract framework and the NEXT50 segmentation form a practical toolkit.

The six dimensions reveal where the opportunities to build relevance lie – in contribution, in a different energy, in new alliances, in intergenerational bridges, in revalued brand relationships, in a future-facing outlook. The segmentation reveals who to prioritise and how to reach them: which message excites an Embracer, convinces a Realist, or reassures someone Unsettled.

What it asks of us, as researchers and marketers, is a shift in our assumptions and disposition towards age. Stop talking about what older adults have lost, and start engaging with what they are gaining. Challenge the brief that stops at 50. Resist the caricature — the monolithic "silver" segment – in favour of the rich, divergent reality the data and the fieldwork both insist on. The structural under-attention to this audience is not only a commercial miss; it is, increasingly, a failure of relevance for the industry that advises on it.

Crucially, none of this comes at the expense of younger customers. The opposite, in fact: many of the brands that engage older adults best are precisely those that engage the young just as well. The lesson is to think age-inclusive first – communications, products and services designed to span generations rather than wall them off. There is a time and place for more targeted older-adult propositions, and emerging technology will only make that kind of hyper-personalisation easier and sharper. But the governing principle is this: speaking to older adults will not age your brand. Done well, it widens the reach and attention you can earn – across the young and the old alike.

Which returns us to where we began: lifetime value. Brands that take the New Contract seriously are not chasing a fading audience but extending the most valuable stretch of the customer relationship – and nowhere is that value greater, or more rewarding to earn, than in the 50–75 cohort this paper has been about.

Takashi is not waiting to be marketed to. He is too busy living a life that most brands haven't yet imagined him having. The real risk, in other words, isn't ageing audiences. It's ageing thinking.

## A Note on Sources

Unless otherwise attributed, all statistics are drawn from our proprietary NEXT50 survey (n=1,500; adults aged 50–75; Singapore, Japan and Indonesia; equal base of 500 per market). Qualitative verbatims are from the study's in-home interviews and expert interviews; "Takashi" is a pseudonym used to protect the respondent's identity. Sources for all external figures are quoted on relevant pages and listed in the references section.

Copyright © **ESOMAR** 2026


## References

AARP Media Advertising Network (n.d.). *The Mighty 50-Plus*, citing the U.S. Consumer Expenditure Survey, https://advertise.aarp.org/50-insights/mighty-50-plus

Emsley, V. (2020). *Don't Underestimate the Market Power of the 50+ Crowd.* Harvard Business Review, 9 January 2020, https://hbr.org/2020/01/dont-underestimate-the-market-power-of-the-50-crowd

McKinsey Global Institute (2021). *The Trailblazing Consumers in Asia Propelling Growth.* https://www.mckinsey.com/featured-insights/asia-pacific/the-trailblazing-consumers-in-asia-propelling-growth

Ossan Rental: https://rental-japanese.com/ossan-rental-japan/

Potluck Podcast – New Contract episodes**:** https://podcasts.apple.com/gb/podcast/the-new-contract-meets-the-silver-backpacker/id1491720768?i=1000713013020

Rella, B. (2023). *The Undertapped Power of the 50+ Consumer.* WARC Opinion, 9 August 2023, https://www.warc.com/newsandopinion/opinion/the-undertapped-power-of-the-50-consumer/en-gb/6284

Statistics Bureau of Japan (2019), *National Survey of Family Income, Consumption and Wealth* - https://www.stat.go.jp/english/data/zensho/index.html

Statistics Bureau of Japan (2024). *Family Income and Expenditure Survey, Annual Report.* https://www.stat.go.jp/english/data/kakei/index.html

UN DESA, World Population Prospects 2024, https://population.un.org/wpp/; UN ESCAP, Population Ageing data portal, https://www.population-trends-asiapacific.org/population-ageing

## About the Authors

Scott Percival, Director & Founder, Lore Consulting, Singapore.

Drago Djourov, Managing Director, DD Global, Singapore & Japan.

# Conversations with Agents

## Turning complex consumer data into confident innovation bets

*Dave Lundahl*

*John Ennis*

*Mimi Sherlock*

*Greg Stucky*

Esomar

Office address:

Burgemeester Stramanweg 105

1101 AA, Amsterdam

Phone: +31 20 664 21 41

Email: info@Esomar.org

Website: www.Esomar.org

Publication Date: September 2026

Esomar Publication Series Volume S424 Congress 2026

ISBN 92-831-0337-8

Copyright © **Esomar** 2026

Copyright © **Esomar** 2026

Congress 2026

# Conversations with Agents

## Turning complex consumer data into confident innovation bets

*Dave Lundahl*

*John Ennis*

*Mimi Sherlock*

*Greg Stucky*

### Introduction: What Happens When Insight Stops Being a Report and Becomes a Conversation?

For decades, the insights industry has been organised around static outputs: reports, charts, dashboards and executive summaries designed to "deliver" answers. Yet, as datasets multiply and decisions accelerate, confidence often declines. Teams are left navigating fragmented evidence, outdated learnings and unanswered questions, without a clear way to interrogate what they already know. As human and machine intelligence evolve side by side, the insight industry is moving from static outputs to evidence that can be questioned, stress tested and refined in real time (Li et al., 2024; Nuremburg Institute for Market Decisions, 2026; Sarstedt et al., 2024). We introduce a fundamentally new paradigm: insight as a living conversation.

### Methodology

Conversations with agents present a live conference panel composed entirely of AI agents, each designed for a distinct role in the insight process. The session models a way of working in which evidence is explored, debated and stress tested in real time. It focuses on a real-world research problem: discovering moments of opportunity for food and beverage innovations among consumers taking GLP-1 prescriptions (KPMG, 2024; White et al., 2024).

#### *AI Agent Method Disclosure*

This paper employs AI agents as analytical instruments for evidence interrogation and to engage with teams of people to accelerate human learning within iterative workflows throughout the front end of innovation. Synthetic persona agents are used to simulate plausible consumer behavioural patterns for internal decision support only. Outputs do not constitute primary research with human respondents and require human validation before business application. The design reflects emerging guidance on synthetic samples and AI-assisted research, where usefulness depends on transparency, source grounding and explicit limits rather than claims of direct representativeness. The session demonstrates an agentic insight workflow that integrates unstructured consumer narratives, existing research and behavioural frameworks into a shared knowledge workspace. Specialised AI agents are trained for distinct analytical roles rather than generic summarisation. This is a practitioner case study using commercial platforms developed by the authors' organisations: Claira™ from InsightsNow (n.d., a), and THEUS from Aigora (2026). The authors' affiliations are disclosed below, and the case is presented as a transparent demonstration of method and workflow rather than as an independent product benchmark.

#### *Project Set Up*

This workflow begins with a set-up process facilitated by Claira, an AI agent from InsightsNow (n.d., b), built to apply behavioural science frameworks to consumer-product behaviour within structured innovation workflows. Claira works collaboratively with a human project manager to define the business objectives, stakeholder objectives, project goals in terms of behavioural impact and the workflows needed to reach them. That set-up information keeps the innovation team focused on the agreed goals and gives the downstream agents a common behavioural brief.

### *Data Ingestion and Synthesis*

Once the project set-up is established, the workflow moves into ingesting and synthesising historical evidence. In this case, the evidence base included 17 primary research reports from IFF and a modelled summary of 25,000 posts from GLP-1 users about their experiences. The social listening modelling was contributed by 113 Industries9
(n.d.), whose work applies natural language processing and machine learning to large volumes of digital consumer conversations. The data used in this workflow was treated as aggregated evidence for synthesis; no personally identifiable information was used for the agentic panel. Claira then synthesised this evidence into a behavioural intelligence structure that could be used by the innovation team.

### *Gap Analysis and Filling Gaps*

Claira begins by conducting a gap analysis, identifying where the evidence is strong, where confidence is lower and what missing information would increase the scope and confidence of innovation decisions. This is where the agentic panel discussion demonstrated on stage begins: not by pretending the evidence is complete, but by making its strengths and gaps visible. To extend this synthesised content into a live knowledge experience, an integration between Claira and THEUS, Aigora's agentic research intelligence application, was built. Claira synthesised the behavioural intelligence, while THEUS mapped that structured intelligence into a live, multi-agent conversational panel hosted through Dr. Sinclair, the moderator agent. This provided a pathway to generate evidence-conditioned consumer persona agents, each with a distinct persona name and to let human participants interrogate the underlying evidence in real time. These personas are analytical instruments: synthetic proxies for evidence interrogation, not measured consumers. They are designed to surface gaps, challenge assumptions and simulate plausible behavioural patterns derived from the evidence base. Results generated through persona agents require human interpretation and primary validation before being used for business decisions.

The resulting conversation was intentionally guided by a human moderator who stewarded the analytical dialogue itself, probing interpretations, challenging assumptions and ensuring evidence was examined with rigour and transparency. The human moderator introduces the agents and begins with an audit assessment of existing knowledge from Claira. The moderator then directs Dr. Sinclair, a THEUS-hosted AI moderator persona, to ask questions to the GLP-1 consumer persona agents. The new information generated through this agentic panel is then summarised into new insights by Claira and reviewed by the human team before being used for downstream innovation decisions.

### *Safeguards and Limits*

The workflow was designed as decision support, not as a substitute for primary validation with consumers. The agents worked from the evidence provided to the research base and behavioural intelligence layer; where the evidence was thin, the system was expected to expose gaps rather than fill them with false certainty. In alignment with the ICC/ESOMAR Code's (2025) emphasis on transparency, synthetic personas did not generate new empirical primary data. They acted as conversational interfaces to the existing, ingested evidence. Human moderators and project leaders remained responsible for framing questions, interpreting outputs, deciding which ideas to carry forward and determining when new consumer research would still be required.

### *Moments Discovery*

Claira then applied this behavioural intelligence to identify the top moments of opportunity for innovations to improve the lives of GLP-1 consumers. Claira defined moments of opportunity by synthesising and categorising evidence about the context of snacking occasions; the unmet needs, unfulfilled aspirations and unresolved tensions felt in those situations; and the motivators inherent in those GLP-1 snacking experiences. One moment was selected for downstream co-creation.

### *Ideation and Concept Co-creation*

Copyright © **Esomar** 2026


Claira then worked as a co-creator with an IFF innovation team of 10 professionals through three 90- to 120-minute facilitated sessions via an online forum. The co-creation sessions were designed to ideate and build full, client-facing concepts targeting the selected moment of opportunity. The co-creators were separated into two teams of five to work independently with Claira. The ideation session by these two teams resulted in the creative generation of hundreds of different ideas. Human co-creators each submitted their initial ideas. Claira then applied her intelligence to complete complex tasks such as generating new ideas from the initial set of human ideas, adding new ideas for consideration and applying creative frameworks such as SCAMPER (Eberle, 1972), to create idea permutations. At the end of day one, the co-creators from each team selected their top 10 ideas. The final two days were spent by each team working with Claira to build each idea into full concepts and client-facing concepts with images. Claira was able to quickly refine, modify and propose ways to strengthen concepts, directed by a team member facilitating the collaboration. Final product and packaging images were also created through human-in-the-loop engagement with Claira. Each team was tasked to submit its four top concepts to move forward into the innovation process. Claira then processed all the relevant content about the selected moment, the final concepts and associated behavioural insights into a playbook for each final concept. Each playbook is the brief for translating each concept into product development and sensory specifications, packaging design specifications and marketing positioning.

## Results

In this case, conversations with agents produced a behavioural intelligence workflow that supported IFF in its front end of innovation services addressing the growing needs of consumers taking GLP-1 prescriptions. The synthesised evidence suggested that GLP-1 users were not simply “on a diet”. They were experiencing an entirely new understanding of food, one that existing meal and snacking frameworks could not fully accommodate. These observations were derived from Claira’s synthesis of the 17 proprietary research reports and the 113 Industries social-listening model, not from direct consumer input to the agentic panel. Eating had become schedule-driven and nutritionally functional, with format mattering far less than in-the-moment utility. Yet, despite food becoming a practical necessity, the desire for the pleasure it once provided had not disappeared. Within the modelled evidence base, users gravitated toward restrained, gentle and clean sensory experiences such as light citrus, refreshing coolness, silky texture and clean finish. These patterns were treated as suggestive opportunity signals requiring downstream consumer validation, not as population-level prevalence estimates.

This was surfaced for the IFF team through Claira’s interactive chat panel, an expandable sidebar where the team could view behavioural analysis in real time. IFF could immediately collaborate with Claira, combining or separating moments, requesting that they be expanded or collapsed and reframing them around specific tensions or motivators the team chose to focus on. The moments were distinct: each defined the context of an occasion or situation where there were unmet needs, unfulfilled aspirations or unresolved tension, along with a set of motivators that could be applied to develop product concepts. From this behavioural intelligence, Claira identified over 20 distinct moments of opportunity. The InsightsNow team selected 10 to present to IFF, who chose one for deeper development: the “end of day emotional reset moment” (see Figure 1).

## The End of Day Emotional Reset Moment

*Comfort and Guilt Avoidance*

End-of-day moments strongly need mental or physical breaks to maintain emotional balance

**Context**
Typically occurs at the end of the workday, early evening, at home, alone or with a partner.

**Internal Needs**
Focus on taking a mental or physical break, maintaining emotional balance, and avoiding guilt or regret.

**External Motivators**
Seeking comforting, treat-like snacks that help relax and unwind without being heavy or unhealthy.

**Core Behavioral Tension**
The central tension in this moment is the conflict between emotional desire and sensory reality. GLP-1 users want joy and comfort through food but can't achieve it through traditional indulgent foods due to altered taste perception.

INSIGHTSNOW.COM

*Figure 1*

This moment focused on a precise potential tension GLP-1 users experience: they arrive at the end of the day tired and need to mentally and emotionally unwind. As non-users do, they still seek the comfort and joy of food, but may discover that, with the altered sensory experience, traditional indulgent foods no longer provide the feeling they need. The IFF team prioritised this as a potential unmet need for co-creation and future validation. With the opportunity defined, Claira generated hundreds of product ideas grounded in that behavioural context, each connected to the consumer moments and motivators that made it viable. From this, the IFF team identified a standout concept: "Liquid Comfort Pearls. Gentle Comfort, Clean Finish" (see Figure 2).

## Liquid Comfort Pearls
## Tagline: Gentle Comfort, Clean Finish.

**The Moment & The Tension**

It's the end of a long day. You're tired, and you just want to unwind. You crave that feeling of comfort, but your usual go-to snacks don't taste right anymore. They're too sweet, too rich, or leave a strange aftertaste. You're stuck between wanting something soothing and fearing the discomfort. You need a moment of true comfort that works with your new senses, not against them.

**The Product**

Liquid Comfort Pearls are a small, teardrop-shaped treat that dissolves slowly, releasing a gentle, pure flavor. Designed for your altered senses, they offer a mindful moment of comfort without overwhelming sweetness or a lingering aftertaste. The unique pearl shape evokes a liquid, providing a soothing, flowing sensation that helps you truly relax and reset, with a clean finish that lets you move on.

INSIGHTSNOW.COM 23

*Figure 2*

Designed for the moment when consumers want to unwind but can no longer turn to familiar snacks, the product takes the form of small, teardrop-shaped treats that dissolve slowly and release gentle flavours, such as ginger and honey, lemon and chamomile, as well as sea moss

and kelp (see Figure 3). They are designed to soothe rather than overwhelm, so that the comfort of eating survives without the sensory intensity GLP-1 users may no longer tolerate.

*Figure 3*

Once the concept was confirmed, Claira produced a playbook: a single, behaviourally-grounded reference document giving every team involved a unified direction. For Liquid Comfort Pearls, this included:

- Moment of opportunity;
- Consumer-facing concept;
- Product guardrails;
- Key attributes and cues in target-moments of use;
- Key ingredients;
- Target audience for screener development;
- Moment of definition and context for priming.

Recommended behavioural key performance indicators and methods are:

- Tension resolution and benefit realisation;
- Product performance and brand perception;
- Behavioural metrics of success.

The playbook is what keeps product development, marketing, insights and packaging working from the same consumer understanding, rather than diverging as briefs get written by different people and handed out. In this engagement, work that would typically take months of cross-functional effort was completed in under 48 hours of elapsed time, requiring seven hours of IFF's time and seven hours from InsightsNow. These numbers are case-specific and reflect a project with a pre-existing evidence base; they do not represent generalised benchmarks. Still, they show the practical compression of the front-end innovation cycle when evidence, behavioural interpretation and co-creation are organised around a shared intelligence layer. Concept development moved just as quickly. From the selected moment of opportunity, the IFF and InsightsNow teams evolved the Liquid Comfort Pearls concept across two focused sessions in a single day, demonstrating that speed and rigour do not have to be a trade-off when innovation is grounded in a shared behavioural foundation from the outset.

In three 90- to 120-minute sessions, IFF was able to review the identified behavioural opportunity map, align on the most promising direction and commit to the "end of day emotional reset moment" as its focus. This is the kind of decision that would previously have required weeks of research consolidation, multiple cross-functional workshops and months of internal debate. More importantly, that decision was not a hasty gut call. It was grounded in 17 proprietary research reports and a modelled synthesis of GLP-1 social media posts, organised through behavioural intelligence and presented in a form the team could immediately discuss and act on. The agentic panel discussion, structured around synthetic persona agents interrogating the evidence base, demonstrated the workflow's value as a decision-support mechanism. Because the agents were grounded in proprietary research and social-listening data rather than representing live consumer samples, their outputs were hypothesis-generation tools requiring downstream human validation.

### *What was Validated*

The workflow was run as decision support and not as a substitute for primary research, so it is worth stating plainly which claims rest on what. The synthesis rests on the ingested evidence, and the agents acted as conversational interfaces to that evidence rather than as sources of new empirical data. The behavioural interpretation built on it was reviewed by the human team before it was carried forward. The persona agents produced no primary data and no prevalence estimates, and their contributions were treated throughout as hypotheses for testing. The concepts, the moment prioritisation and the playbook therefore remain subject to primary consumer validation, which is the condition under which they were produced (ICC/ESOMAR, 2025).

### *Strengths and Limits*

The strengths of this way of working follow from its structure. Because the agents reason over an ingested evidence base rather than generating freely, their claims sit against material that can be inspected, and the gap analysis presents the weak parts of that base alongside the strong ones rather than leaving them implicit. The limits follow from the same structure. An agent grounded in a corpus cannot report a behaviour that no source in the corpus recorded, so the evidence base sets a ceiling on what the workflow can produce. This is a property of the method rather than of any particular platform, and it is the central caution in the literature on synthetic samples (Nuremberg Institute for Market Decisions, 2026; Sarstedt et al., 2024). Where the evidence is thin, or where part of a population is poorly represented in it, output stays fluent while becoming less reliable, which is why responsibility for interpretation and for deciding when new consumer research is required stayed with the human moderator and project leaders.

### *Comparison with Consumer Co-creation*

The co-creation described above was expert co-creation: an IFF team working with an AI agent, informed by synthetic personas grounded in prior consumer evidence. No consumers took part, so the comparison with consumer co-creation is a comparison of what each route can contribute rather than a head-to-head result. Consumer co-creation supplies what a grounded agent cannot: information that no existing source has recorded, language in the consumer's own register and the standing that a concept acquires inside an organisation when consumers helped build it (Hoyer et al., 2010; Sanders & Stappers, 2008). Agentic work supplies breadth and speed and holds an entire evidence base in one conversation, but its outputs are bounded by that evidence and are not measurements of a population (Li et al., 2024; Nuremberg Institute for Market Decisions, 2026; Sarstedt et al., 2024). Our view is that the two are sequential rather than alternative: agentic work to decide which problem is worth solving, consumer research to establish whether the problem and the proposed answer are real. Running both legs on the same moment, with the consumer leg blind to the agentic output, would test that view directly and remains open work.

## Conclusions

Conversing with agents changed the output as much as the pace. Every brief, concept and recommendation is tied back to a shared behavioural evidence base, giving teams more confidence to pursue bolder innovation while still knowing where evidence is strong, weak or missing. The

outcome of this end-to-end process is a shift in how innovation is oriented: away from product quality as the only organising principle, and toward the consumer behaviours, motivations and unmet needs that actually determine market success. Impactful innovation cannot happen in a silo. A conversational behavioural-intelligence workflow helps ensure that it does not. The IFF engagement shows what behavioural intelligence makes possible when it is embedded directly into the innovation workflow: decisions made faster and made with a clearer view of what supports them. Conversations with agents changes how insight suppliers and client-side insight functions deliver research project-by-project, and it changes what is possible for the internal and external clients who use it.

For a business like IFF, whose success depends on understanding where consumer demand is heading and helping its clients get there first, this new paradigm can shift how it operates. Instead of waiting for a consumer problem to become news and then doing research into that problem to find an opportunity, IFF can arrive at clients with behavioural intelligence already synthesised, specific opportunities identified, moments mapped and ideas grounded in evidence. It creates a path from reactive supplier toward proactive innovation partner. This matters commercially as much as strategically. When IFF can identify a plausible unmet need rapidly, such as the “end of day emotional reset moment” for GLP-1 users, and prototype a product concept built around its own ingredients, it is creating the conditions for a long-term commercial relationship. Responsiveness builds trust, trust supports product adoption and every product on the shelf is an ongoing ingredient relationship. The industry implication follows from the same mechanism. Insights teams have long operated in a world where the gap between consumer understanding and business action was measured in months. Conversations with agents compresses that gap by ensuring that research, social signals and behavioural patterns are organised, accessible and available for immediate interrogation. Teams no longer start from scratch or a hunch; they start from the knowledge they already have.

Across engagements, this kind of workflow can increase innovation throughput by replacing the slow, costly process of building consensus across R&D, marketing, insights and packaging, with a shared behavioural language that teams carry from the first session through to execution. The precise lift will vary by organisation and should be measured engagement by engagement, but the mechanism is clear: less time reconciling fragmented evidence, more time making informed innovation choices. For businesses operating in fast-moving categories, where consumer trends, tastes and behaviours shift continuously, this responsiveness is what separates “Big I” innovation from “little r” renovation that does not make true market impact. Conversing with agents offers a standard the industry can reasonably hold itself to: insights that can be questioned rather than only reported; stress tested rather than only summarised; and carried forward as a compounding knowledge base rather than archived.

## References

113 Industries. (n.d.). *Areas of Application.* Available: https://113industries.com/applications/

Aigora. (2026). *THEUS: Multimodal Research Intelligence Platform*. Available: https://theus.aigora.ai/

Eberle, B. (1972). Developing Imagination Through SCAMPER. *The Journal of Creative Behaviour, 6*(3), 199-203.

Hoyer, W. D., Chandy, R., Dorotic, M., Krafft, M., and Singh, S. S. (2010). Consumer cocreation in new product development. *Journal of Service Research, 13*(3), 283-296. https://doi.org/10.1177/1094670510375604

ICC/ESOMAR. (2025). *International Code on Market, Opinion and Social Research and Data Analytics*. Available: https://standards.esomar.org/assets/documents/icc-esomar-code-2025.pdf

InsightsNow. (n.d., a). *Claira Behavioral Intelligence.* Available: https://insightsnow.com/claira-behavioral-intelligence/

InsightsNow. (n.d., b). *Behavioral Frameworks.* Available: https://f.hubspotusercontent00.net/hubfs/6206054/_New%20Website%20Resources/Sales%20Sheets/INI%20Behavioral%20Frameworks%20Paper.pdf

Li, P., Castelo, N., Katona, Z., and Sarvary, M. (2024). Frontiers: Determining the Validity of Large Language Models for Automated Perceptual Analysis. *Marketing Science, 43*(2), 254-266. https://doi.org/10.1287/mksc.2023.0454

Copyright © **Esomar** 2026


KPMG. (2024). *Getting to know GLP-1 users, a new kind of consumer*. Available: https://kpmg.com/kpmg-us/content/dam/kpmg/pdf/2024/glp-1-meds-impact-on-food-and-bev-ind.pdf

Nuremberg Institute for Market Decisions. (2026). *Leaving Insight to Digital Twins? Promise, Progress and Limits of Synthetic Respondents*. Available: https://www.nim.org/en/publications/detail/leaving-insight-to-digital-twins

Sanders, E. B.-N., and Stappers, P. J. (2008). Co-creation and the new landscapes of design. *CoDesign, 4*(1), 5-18. https://doi.org/10.1080/15710880701875068

Sarstedt, M., Adler, S. J., Rau, L., and Schmitt, B. (2024). Using large language models to generate silicon samples in consumer and marketing research: challenges, opportunities, and guidelines. *Psychology & Marketing, 41*(6), 1254-1270. https://doi.org/10.1002/mar.21982

White, A. M., et al. (2024). Changes in food preferences and ingestive behaviors after glucagon-like peptide-1 analog treatment: techniques and opportunities. *Nature*. Available: https://www.nature.com/articles/s41366-024-01500-y

## About the Authors

Dave Lundahl, Chief Executive Officer, InsightsNow, US.

John Ennis, Chief Executive Officer, Aigora, US.

Mimi Sherlock, VP Consumer Intelligence, IFF, US.

Greg Stucky, Chief Research Officer, InsightsNow, US.

# Metamorphosis of Expertise

## How AI is Redefining Value in the Insights Industry

*Kamila Zahradníčková, MSc*

## Esomar

Office address:

Burgemeester Stramanweg 105

1101 AA, Amsterdam

Phone: +31 20 664 21 41

Email: info@Esomar.org

Website: www.Esomar.org

Publication Date: September 2026

ESOMAR Publication Series Volume S424 Congress 2026

ISBN 92-831-0337-8

Congress 2026
Copyright © **ESOMAR** 2026

Copyright © **ESOMAR** 2026


# Metamorphosis of Expertise

## How AI is Redefining Value in the Insights Industry

*Kamila Zahradníčková, MSc*

### Market Research Is Not in Crisis: It Is in Metamorphosis

The distinction matters. Crisis implies breakdown. Metamorphosis implies that something new is being formed, even when the process feels disorienting from the inside. And disorienting it is.

Talk to enough people in the industry and a strange pattern emerges. Teams are delivering projects faster than ever, yet revenues stay flat. Methodological training, the foundation careers were built on, is still respected in principle but increasingly absent from daily practice. Junior researchers struggle to enter a field that is hiring at lower volumes, while the juniors already inside, the ones fluent in AI tools, find themselves teaching senior colleagues. The career paths that once felt reliable no longer read as legible.

These are not random symptoms. This paper argues they reflect a single, coherent structural condition. Once that condition is named, the path forward becomes clearer; not easy, but legible. The argument rests on a qualitative study of how senior and junior practitioners across European agencies, corporate insight teams and industry associations are experiencing AI disruption, interpreted through Pierre Bourdieu's theory of professional fields.

The core finding is this: AI is not disrupting one part of the insights profession at a time. It is disrupting all of them at once. And that simultaneity is exactly why this moment feels different from every disruption that came before it.

### Why Everything Feels Uncertain at Once

Studies of AI at work keep returning the same pattern across domains: generative tools lift productivity in writing and analysis (Noy & Zhang, 2023), reshape consulting performance (Dell'Acqua et al., 2024), and change who commands authority inside organisations (Brynjolfsson et al., 2023). Institutional scholars go further, arguing that such shifts do not merely change how work is done but alter the rules of the game within entire fields (Hinings et al., 2018; Kellogg et al., 2020). What is less well specified is the mechanism by which disruption operates at the level of the field rather than the individual. Market research is a useful place to study it, because here the disruption is unusually visible: AI now performs tasks that used to define professional expertise, from sampling and segmentation to the synthesis of insight (Chintalapati & Pandey, 2021; Mariani et al., 2022; Hermann & Puntoni, 2024).

To explain the pattern, this study draws on Bourdieu's field theory, a framework for understanding how a profession reproduces itself over time (Bourdieu, 1977; Bourdieu & Wacquant, 1992). Strip away the jargon and the idea is intuitive. Every professional field runs on a set of conversion rules: expertise becomes reputation, reputation becomes pricing power, seniority becomes authority. Bourdieu called the things being converted "capital," and identified four forms of it (Bourdieu, 1986):

- Cultural capital – your training, your methods, your craft.
- Economic capital – what your work earns.
- Social capital – your standing and authority among peers.
- Symbolic capital – the shared sense of what counts as legitimate, rigorous work.

When these conversion rules hold, the field is stable. People can invest in skills, credentials, and relationships with a reasonable expectation of return. A junior learns proper methodology because methodology converts into recognition, recognition converts into seniority and seniority converts into income and authority. The machine runs.

Fields have always had to adapt their conversion rules. What is unusual about this moment is not that one rule is changing. It is that all four appear to be changing at the same time. This study names that condition *field hysteresis*: a state in which the conversion mechanisms across all four forms of capital become unreliable simultaneously, leaving practitioners without the stable footholds that normally guide adjustment.

This is an important reframe. Hysteresis, in Bourdieu's original sense, describes a lag inside a person – an individual whose instincts, formed under old conditions, no longer fit new ones (Strand & Lizardo, 2017). The condition described here is not in the person. It is in the field. The rules themselves have gone quiet. People are not failing to adapt because they are resistant or complacent. They are operating in a field whose new rules have not yet been written.

## Why AI Is Different from the Last Disruption

The industry has absorbed technology shocks before. The shift to online panels in the early 2000s was disruptive, but it was absorbed gradually and conversion rules were recalibrated one at a time as the field adjusted.

AI is different, and the study identifies the mechanism: *heteronomous entry*. The term sounds academic; the idea is sharp. AI did not emerge from inside the insights field, vetted by its standards, taught through its training pathways and blessed by its professional bodies. It arrived from outside, from industries whose scale, data and computing infrastructure let them define what "good research" looks like without ever asking the field's permission. As a senior association representative put the asymmetry: the whole market research industry is small next to the data industry, and the firms that already hold the data and the infrastructure increasingly hold the tools to do research, and so increasingly get to define it.

Previous disruptions came from within the field's own logic. This one comes from outside it, at a speed and scale that compresses what would normally be a sequence of adjustments into a single simultaneous shock. That is why all four conversions are destabilised at once: they were all calibrated to a logic that no longer holds exclusive authority.

## How the Study was Done

This is a qualitative, interpretivist study, designed to surface how professionals experience and narrate disruption rather than to measure it statistically (Arsel, 2017; Dolbec et al., 2021). Bourdieu's concepts, "field", "capital" and "hysteresis", were used as sensitising lenses to guide the interviews and an abductive analysis, not as a rigid grid imposed on the data.

Ten in-depth interviews were conducted in Q4 2025, in Czech and English, lasting 40 to 60 minutes each. The sample was chosen for maximum variation in field position across roles, seniority and organisational context rather than for size. Participants span agency, corporate and association roles; junior to senior levels; and CEE, DACH, UK and global contexts.

Names and identifying details have been changed. That practitioners in structurally different positions report encountering the same disruptions is itself meaningful: it is consistent with a field-level condition rather than a set of isolated personal experiences. The small sample means the findings are theoretically transferable rather than statistically representative: they illuminate a process, not a population.

On positionality: the interviewer occupies a dual role as academic researcher and founder of an AI company operating in this same domain. That position can shape how participants respond and how findings are read. To mitigate this risk, a university email was used for recruitment, interviews were scheduled through university channels where possible and the academic and independent nature of the work was made explicit at the start of each conversation.

## What Practitioners Are Experiencing

The findings organise around three dynamics: AI arriving as an outsider that reorders the field; awareness that outruns the ability to act; and the breakdown of all four capital conversions at once.

### *1. AI Arrives as an Outsider and the Work Gets Reclassified*

Respondents describe AI entering from outside the field, importing logics of scale and speed that sit awkwardly against the craft-based norms the profession was built on. The most visible symptom is quiet but telling: research work is being reclassified as technical work. A junior agency employee finishing a master's in applied research methodology described knowing the methods but rarely applying them: there is not much research anymore, they said, more IT. When interpretation and methodological reasoning recede and technical execution moves to the centre, the field's evaluative criteria are no longer being generated from within.

### *2. The Say-Do Gap: Awareness Without the Means to Act*

Almost everyone recognises that a leap is required. Almost no one finds it easy to make. As one senior respondent put it, everyone knows the leap has to happen; it is hard precisely because it is a technology leap and most people did not come from technology, and there is little time to adapt.

This is not resistance. It is a mismatch between the skills people have and the skills now rewarded, made worse by clients who continue to demand the same comprehensive outputs while budgets stay flat. As a supplier-side respondent observed, research budgets have stopped growing even as demand for research has not fallen – so the pressure lands on researchers to become more efficient with less. The result is a practical lag: people know what is required but lack the time, tools or organisational support to do it. That is hysteresis as a lived experience, not a strategic choice.

## 3. The Four Conversions Break at Once

This is the analytical core. Across the interviews, all four capital conversions show signs of breaking, concurrently.

Cultural capital decouples from rewarded practice. A junior researcher described teaching a colleague to analyse interview transcripts with an AI assistant and using none of the traditional methodological skills they had been trained in. It was, they said, all just a matter of asking the right question in a prompt. The training is still respected; it is simply less frequently enacted. The skill is not disappearing so much as relocating: another junior described the shift in thinking

as making sure you instruct the system to do the work well. Expertise is migrating from executing analysis to specifying, supervising and translating it, with orchestration as the new basis for distinction.

Economic capital conversion stalls. Efficiency gains are real but not yet priced. As one respondent put it, projects move through far faster, but the cost has not changed. Speed has not translated into value capture.

Social capital hierarchies invert. Uncertainty has slowed hiring, especially at junior levels: one junior called it hell for newcomers. Yet the juniors who are inside and AI-fluent find colleagues coming to them for help, converting technical familiarity into standing. Entry gets harder; internal mobility accelerates for the well-positioned. The apprenticeship model, where juniors learn craft from seniors, runs in reverse.

Symbolic capital is contested. Industry bodies are trying to re-establish what counts as legitimate practice, but without consensus. One representative described setting standards and self-regulation about what the field is allowed to do; another captured the difficulty more honestly – a crowd of people with limited knowledge debating things they do not fully understand, while the actual AI experts stay silent.

| Capital form | The conversion it powered | What is breaking |
|---|---|---|
| **Cultural** | Training and method become recognised expertise | Methodological training is still respected but rarely enacted; value migrates from executing analysis to orchestrating it |
| **Economic** | Recognition becomes pricing power and income | Efficiency gains are real but unpriced – faster delivery, flat fees |
| **Social** | Seniority becomes authority and mentorship | Junior hiring slows; AI-fluent juniors gain standing and teach seniors; the apprenticeship model runs in reverse |
| **Symbolic** | Shared standards define legitimate practice | Legitimacy is contested; bodies debate standards without consensus |

*Table 1. The four capital conversions and how each is breaking under AI.*

No single one of these would be remarkable on its own. Fields lose and rebuild conversion rules all the time. What distinguishes this moment is that all four are reported as unreliable at once, by practitioners in different roles and organisations. There is no stable form of capital left to anchor the adjustment. A researcher whose pricing power erodes would normally fall back on credentialled expertise or reputation. Here, those footholds are shifting too.

## How Practitioners Are Coping (Provisionally)

Adaptation is happening, but it is best read not as the field restabilising. It is people navigating a field that has not yet settled on new rules. The responses are provisional and uncoordinated – which is itself the point.

At the field level, associations attempt to stabilise legitimacy through standards and self-regulation. These are genuine efforts to rebuild the rules for symbolic capital, but their contested, unfinished character shows that no consensus has formed.

At the organisational level, firms experiment and invest selectively. Some adopt AI provisionally, testing whether it makes sense before committing. Others go further and build proprietary tools around their own expertise. One respondent's company hired an entire engineering team to build AI tooling internally rather than depend on outside vendors, protecting their advantage by owning the capability.

At the individual level, the emphasis is on disposition. As one participant said, those who benefit are the ones not shy about trying. Others were more sober about the consequences, rejecting the comforting line that AI will not cost jobs: it will, one argued, because being more efficient means needing fewer people. This is not resistance; it is a pragmatic reckoning with displacement as part of the transition.

Taken together, the findings depict a field whose mechanisms for reproducing professional value are reported as unreliable across multiple dimensions at once. Standards are contested without resolution. Pricing has not caught up with production. Entry and mentorship pathways are disrupted. The criteria for recognising expertise are in flux. This is not a field that has changed. It is a field whose rules for changing have not yet been written.

## What This Means

Existing research shows how professionals respond to AI, defending and redrawing their jurisdictions through boundary work, reconstructing collective identity (Faulconbridge et al., 2025; Goto, 2021; Bankins et al., 2024). What this study adds is an account of the structural conditions that those professionals are responding to. Professionals in market research are not simply adapting to a changed field. They are operating where the rules governing what counts as expertise, how outputs are priced, how authority is distributed, and what constitutes legitimate practice have all become uncertain at the same time. That is a categorically different situation from ordinary field change, in which some conversion mechanisms break while others hold and provide a foothold for adjustment.

The contribution is to move hysteresis from the individual to the field. Existing accounts locate the lag in the actor, one person's instincts persisting after conditions change (Strand & Lizardo, 2017). The pattern here locates the lag in the field itself: the reproduction system is, for now, suspended rather than gradually recalibrated. And the mechanism is heteronomous entry. This refines the institutional account of digital transformation (Hinings et al., 2018): when evaluative criteria are imported from outside rather than generated within, all capital forms destabilise together, because all were calibrated to a logic that no longer holds exclusive authority. That is precisely why the adaptation responses in the data stay provisional and uncoordinated rather than converging on a new professional model – and why so many practitioners can describe what is required without yet being able to act on it.

## Reasons for Optimism, and What to Do Next

Field hysteresis is a transitional condition, not a permanent one. Fields do restabilise. New conversion rules do emerge. And the insights industry has real assets to draw on as it navigates the transition.

The most important is this: human expertise in interpretation, judgement, and contextual sense-making is not being automated, it is being repositioned. As production moves to AI, the distinctively human contributions of the researcher become more visible, not less: framing the right question, recognising the finding that does not fit, translating insight into a decision. The practitioners thriving right now are the ones who have understood that their value lies in

orchestrating research, not executing it. That points to a practical agenda, and it is largely a collective one, because conversion rules are agreed at the level of the field, not invented alone.

For agencies and corporate teams: revisit the business model. Models built on billable, time-intensive production face structural pressure when speed collapses and clients expect the saving passed on. The teams that win will price judgement, governance and decision quality not hours.

For the profession: write the new rules deliberately, before the uncertainty calcifies into bad defaults. What counts as expertise in an AI-enabled field? How is rigour defined and demonstrated when AI handles production? These are not rhetorical questions; they are the standards-setting work of the next two years, and the field is far better off authoring those rules than inheriting them from outside.

For talent and careers: rebuild the path. The apprenticeship model is breaking, and pretending otherwise abandons the next generation. New credentialling and progression frameworks need to be built around AI-enabled practice defining what a junior learns, how mobility works, and what mastery looks like when the craft has moved from execution to orchestration.

None of this is automatic. But it is doable, and it is the work that turns a disorienting transition into a stronger field on the other side.

## Conclusion

This study set out to understand how AI disrupts the way professional value is built and recognised in market research. The answer is that AI does not disrupt one skill or one practice at a time. It suspends, concurrently, the conversion rules across all four forms of capital, cultural, economic, social, and symbolic, leaving the field temporarily unreadable rather than simply changed. The reason is heteronomous entry: AI arriving from outside the field's own logic, carrying evaluative criteria calibrated to a different industry, at a speed that compresses years of sequential adjustment into a single shock. That is why everything feels uncertain at once.

Naming the condition is not a counsel of despair. It is the prerequisite for acting well. Understanding that the lag is in the field, not in any one person, reframes what the moment requires: not just individual reskilling, but the collective rewriting of what expertise, rigour and value mean in an AI-enabled craft. Metamorphosis, navigated with intention, produces something stronger than what came before.

## References

Arsel, Z. (2017). Asking questions with reflexive focus: a tutorial on designing and conducting interpretive interviews. *Journal of Consumer Research*, 44(4), 909–933.

Bankins, S., Hu, X., & Yuan, Y. (2024). Artificial intelligence, workers, and the future of work skills. *Current Opinion in Psychology*, 58, 101828.

Bourdieu, P. (1977). *Outline of a Theory of Practice*. Cambridge University Press.

Bourdieu, P. (1986). The forms of capital. In J. Richardson (Ed.), *Handbook of Theory and Research for the Sociology of Education* (pp. 241–258). Greenwood.

Bourdieu, P. (1993). *The Field of Cultural Production*. Polity Press.

Bourdieu, P. (1996). *The State Nobility: Elite Schools in the Field of Power*. Stanford University Press.

Bourdieu, P., & Wacquant, L. J. D. (1992). *An Invitation to Reflexive Sociology*. University of Chicago Press.

Brynjolfsson, E., Li, D., & Raymond, L. (2023). Generative AI at work (Working Paper No. 31161). National Bureau of Economic Research.

Chintalapati, S., & Pandey, S. K. (2021). Artificial intelligence in marketing: a systematic literature review. *International Journal of Market Research*, 64(1), 38–68.

Dell'Acqua, F., Egidi, A., Gans, J. S., Goldfarb, A., & Tadelis, S. (2024). Consulting with generative AI: experimental evidence (Working Paper No. 24-013). Harvard Business School.

Dolbec, P., Fischer, E., & Canniford, R. (2021). Something old, something new: enabled theory building in qualitative marketing research. *Marketing Theory*, 21(4), 451–456.

Faulconbridge, J., Sarwar, A., & Spring, M. (2025). How professionals adapt to artificial intelligence: the role of intertwined boundary work. *Journal of Management Studies, 62*(5), 1991–2024.

Goto, M. (2021). Collective professional role identity in the age of artificial intelligence. *Journal of Professions and Organization*, 8(1), 86–107.

Hermann, E., & Puntoni, S. (2024). Artificial intelligence and consumer behavior: From predictive to generative AI. *Journal of Business Research, 180,* 114720.

Hinings, B., Gegenhuber, T., & Greenwood, R. (2018). Digital innovation and transformation: an institutional perspective. *Information and Organization*, 28(1), 52–61.

Kellogg, K. C., Valentine, M. A., & Christin, A. (2020). Algorithms at work: the new contested terrain of control. *Academy of Management Annals*, 14(1), 366–410.

Mariani, M. M., Perez-Vega, R., & Wirtz, J. (2022). AI in marketing, consumer research and psychology: a systematic literature review and research agenda. *Psychology & Marketing*, 39(4), 755–776.

Noy, S., & Zhang, W. (2023). Experimental evidence on the productivity effects of generative artificial intelligence. *Science, 381*(6654), 187–192.

Strand, M., & Lizardo, O. (2017). The hysteresis effect: theorizing mismatch in action. *Journal for the Theory of Social Behaviour*, 47(2), 164–194.

## About the Author

Kamila Zahradníčková, MSc, CEO, Lakmoos AI; PhD Researcher, Prague University of Economics and Business, Czech Republic.

# Agentic AI Application for Rapid Creative Production Rooted in Consumer Insights

*Minh Nguyen, PhD*
*Takumi Akuzawa*
*Togo Kida*
*Toshiaki Kamiya*
*Yuki Tanaka*
*Takamata Shoshi*
*Otsuki Erika*
*Takeuchi Ami*
*Munish Bhamoriya*

## Esomar

Office address:
Burgemeester Stramanweg 105
1101 AA, Amsterdam
Phone: +31 20 664 21 41

Email: info@Esomar.org
Website: www.Esomar.org

Publication Date: September 2026
ESOMAR Publication Series Volume S424 Congress 2026
ISBN 92-831-0337-8

Congress 2026
Copyright © **ESOMAR** 2026

Copyright © **ESOMAR** 2026
Congress 2026

# Agentic AI Application for Rapid Creative Production Rooted in Consumer Insights

*Minh Nguyen, PhD*
*Takumi Akuzawa*
*Togo Kida*
*Toshiaki Kamiya*
*Yuki Tanaka*
*Takamata Shoshi*
*Otsuki Erika*
*Takeuchi Ami*
*Munish Bhamoriya*

## 1. Introduction

Agentic AI is a rapidly evolving field with deep implications for marketing and consumer insights. Traditional generative AI (GenAI) applications often generate "average", generic outputs because they lack systematic grounding in authentic customer psychographics. In this paper, we present a novel stateful Multi-Agent AI System consisting of seven specialised agents (Researcher, Planner, Copywriter, Designer, Screenwriter, Director, Reviewer) collaborating under human oversight via an interactive wizard interface. We document a live empirical pilot campaign conducted in Japan in partnership with Rakuten Insight, Inc. for the *Rakuten Household Account Book (Rakuten Kakeibo)* mobile application. The pilot campaign demonstrated that human-in-the-loop AI achieved a 62.5% lift in app installs, a 29.8% relative conversion rate (CVR) lift, and a 16.8% reduction in CPA compared to pure autonomous AI, while executing the full agent pipeline in approximately 30 minutes end-to-end, enabling substantially faster iteration cycles than conventional creative production workflows. Furthermore, we present findings from a qualitative study conducted by Accenture and commissioned by Google with 10 C-suite marketing leaders in Japan. The study reveals that while organisations are achieving Level 4 (Autonomous) execution, marketers must pivot from Operators to Orchestrators, retaining critical "Judge" and "Decide" nodes. It also identifies friction points limiting scale, notably the reduction of the CFO value case strictly to cost savings, data ecosystem fragmentation, and a striking pattern where the most mature adopters are paradoxically the most sceptical about proving hard ROI. Finally, we establish the fundamental industry implication: "Insights are Gold". If the initial consumer insights are flawed, the downstream generative loop cascades in the wrong direction. Generative AI elevates market researchers to serve as critical system governors.

## 2. The Golden Rule of Insights

Large Language Models (LLMs) and text-to-multimodal diffusion models have promised to democratise and automate creative production. However, active brand managers frequently discover a significant hurdle: unsupervised generative systems produce "average" outputs that fail to resonate with consumers. This occurs because generative models optimise for consensus predictions, stripping creatives of the deep psychographic nuances required to drive conversion.

We establish the Golden Rule of Insights: "Insights are Gold". A creative generative pipeline is only as strong as its starting inputs. If the initial consumer insights are flawed or generic, every downstream agent will magnify those inaccuracies, culminating in an ineffective and off-brand campaign.

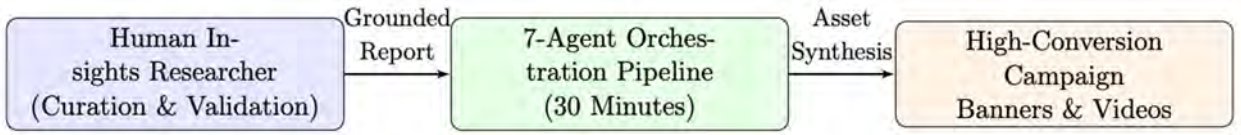



*Figure 1. The insight-grounded creative funnel architecture.*

## 3. The Multi-Agent AI Creative Agency

The platform decomposes the traditional creative agency's campaign production lifecycle into seven specialised cognitive agents built on the Google Agent Development Kit (ADK) and Google GenAI SDK. Operating under continuous human oversight via a sequential wizard, the agents maintain context record all intermediate outputs in a persistent state layer.

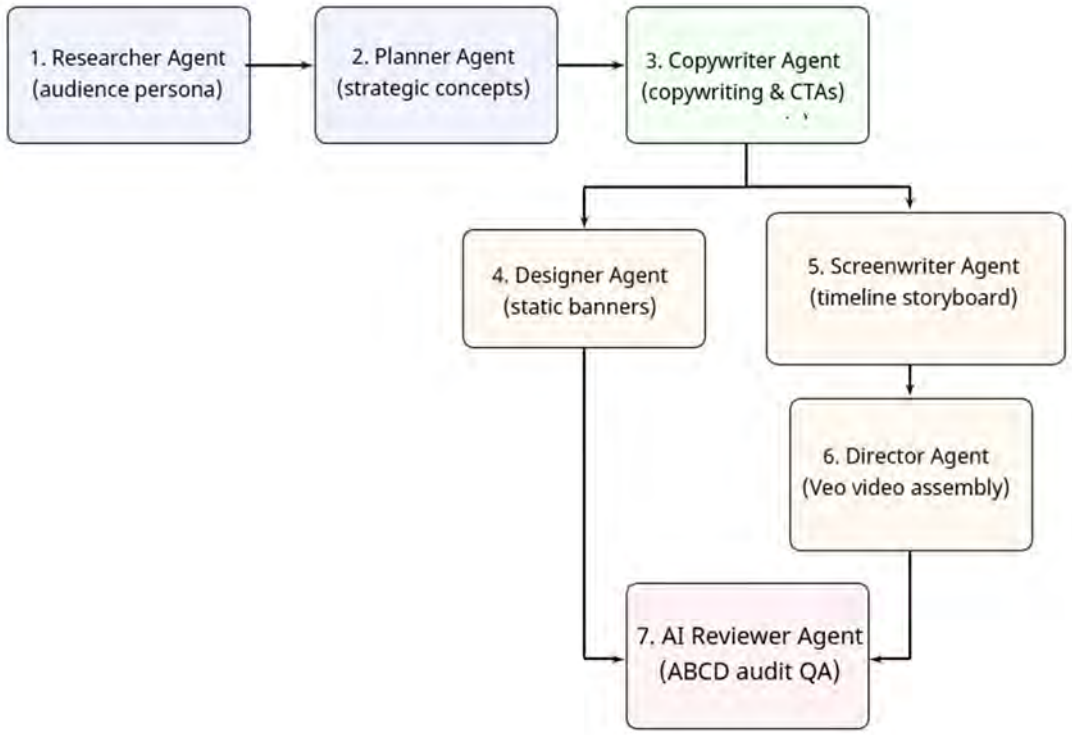



*Figure 2. Seven-agent collaborative creative funnel flowchart.*

## *3.1 Factual Persona Generation and Marketing Concepts*

### *Researcher Agent*

The creative generation process begins at Step 1 with the Researcher Agent, whose core function is data aggregation and factual synthesis.

- Input ingestion: Ingests unstructured data sources, including active product landing pages, competitor advertising URLs, competitive YouTube videos and customer insights files (supports .docx, .pdf, .xlsx, .txt etc.).
- Synthesis and grounding: The agent synthesises this information into a structured *Ground Truth Research Report*. Drawing on retrieval-augmented generation (RAG) methods (Lewis et al., 2020) to ground LLM responses, this report maps verified customer frustrations, competitors' messaging strategies, and behavioural patterns (Action Data), establishing a strict, unhallucinated base context (Ji et al., 2023) for all downstream agents.

### *Planner Agent*

The Planner Agent acts as the core strategist, translating the Researcher's *Ground Truth Report* into rich, multi-dimensional buyer personas and conceptual campaigns.

- OCEAN psychographic modelling: To move beyond simplistic demographics (e.g., age or gender), the Planner translates customer behavioural traits into the *OCEAN Five-Factor Model* (Openness, Conscientiousness, Extraversion, Agreeableness, Neuroticism) to define psychological profiles. This grounds the system in established psychological targeting and digital persuasion principles (Kosinski et al., 2013; Matz et al., 2017).
- Jobs-to-be-done (JTBD) mapping: Integrates Christensen's JTBD framework (Christensen et al., 2016), detailing motivations, triggers and friction points for each persona (Mind Data).
- Strategic campaign matrix: Based on these generated personas, the agent outputs up to three distinct campaign concepts. Each concept specifies a core advertising angle, visual direction and strategic rationale. Users can critique, merge or refine these concepts through a co-pilot chat interface.

## *3.2 Creative Asset Generation*

### *Copywriter Agent*

The Copywriter Agent receives the approved campaign concept and target persona, translating them into compelling messaging suites.

- Dynamic tone matching: Crafts tailored headlines, body copy and Calls-to-Action (CTAs) mapped to specific emotional tones (e.g., reassuring for high-neuroticism profiles, energetic for high-extraversion profiles).
- Structure variations: Generates multiple copywriting alternatives (e.g., question-based hooks, benefit-focused headers or urgency-driven CTAs), giving marketers diverse options aligned with core insights.

### *Designer Agent*

The Designer Agent compiles static banner layouts utilising the Gemini 3 Pro Image Preview (Nano Banana Pro) model.

- Style reference image mapping: To maintain strict brand style guidelines, the agent processes user-uploaded reference images (logos, layouts, colours, app screens). Using multimodal vision capabilities, it aligns the generated banner's colour scheme, typography, and composition with these references.
- Aspect ratio adaptation: Automatically adjusts compositions for different aspect ratios (16:9 landscape, 9:16 portrait or 1:1 square) to fit varied programmatic platforms without visual distortion.

### *Screenwriter Agent*

The Screenwriter Agent designs the sequential narrative blueprint for cinematic video ads.

- Structured timeline output: Translates the campaign concept and copy hooks into an editable scene-by-scene script table. For each scene, the agent defines:
    - Visual prompt: Highly descriptive cinematic prompts optimised for video generation models.
    - Dialogue and voiceover: Natural spoken scripts for narration.
    - Text overlay parameters: Exact text overlay strings, font weights, overlay colours and positioning coordinates.
    - Scene duration: Precise timing offsets in seconds.
- Interactive UI integration: Feeds this data directly into a visual frontend timeline editor where humans can dynamically drag, reorder, edit overlay scripts or adjust durations.

### *Director Agent*

The Director Agent functions as the media compiler, executing video synthesis and vocal narration coordination.

- Google Veo video generation: Directs Google Veo (*Model ID: veo-3.1-generate-001*) to synthesise high-definition video clips matching the Screenwriter's visual prompts.
- Voiceover synthesis and timing auto-alignment: Calls the Google TTS API to synthesise voiceovers. If a track's duration exceeds the scene length, the agent dynamically stretches the video or adjusts script segments to maintain perfect AV synchronisation.
- FFmpeg stitching engine: Assembles the assets through a non-blocking backend compilation pipeline. It probes clips for audio presence via ffprobe, compiles smooth visual crossfades (xfade) and audio crossfades (acrossfade), and runs silent fallback routines to guarantee 100% build success.

### *AI Reviewer Agent*

The AI Reviewer Agent operates as the final quality gate, running multimodal audits via Gemini 3.5 Flash's vision capabilities before human validation.

- ABCD marketing audits: Audits video drafts frame-by-frame against Google's ABCD framework:
    - Attract: Does the video present an immediate hook and engaging visual within the first two seconds?
    - Brand: Is the brand logo clearly visible early in the scene?

- Connect: Does the pacing and emotional tone align with the psychographic persona?
- Direct: Is there a highly visible call-to-action overlay?

- Custom compliance auditing: Audits banners and videos against institutional legal rules and brand style rules, generating an audit log that highlights margin clearance errors, missing regulatory disclosures or off-palette styling elements.

## 4. Methodology and Validation

Our methodology integrates a hybrid approach: prompt/context engineering using Google ADK, a stateful multi-agent pipeline with session database mapping and continuous human-in-the-loop feedback loops.

### *4.1 Quantitative Pilot Results: Rakuten Case Study*

We executed a live pilot campaign in Japan for the Rakuten Household Account Book (Rakuten Kakeibo) mobile app, targeting two distinct psychographic archetypes: an "Effortless Household Manager" persona (women seeking low-effort budget management to create room for small indulgences) and a "FIRE-Oriented Professional" persona (career professionals pursuing early financial independence through long-horizon wealth building). The Step 1 intake phase was grounded in publicly available product context (Rakuten Kakeibo app landing pages and related creative references) together with consumer insights curated by the Rakuten Insight team.[1]

To evaluate the role of human curation, a live programmatic split campaign was conducted concurrently for 15 active delivery days between 13 February and 5 March 2026 on the Moloco DSP network. The campaign directly compared pure autonomous AI creative assets (Group A: Autonomous AI) to human-revised agentic-AI assets (Group B: Human-in-the-Loop AI). Aggregated performance metrics are detailed in Table 1.

| Campaign Metric | Autonomous AI (Group A) | Human-in-the-Loop (Group B) | Relative Lift (Group B vs. Group A) |
|---|---|---|---|
| Impressions | 2,587,738 | 3,275,364 | +26.6% |
| Clicks | 25,910 | 32,433 | +25.2% |
| App Installs | 40 | 65 | +62.5% |
| CTR | 1.001% | 0.990% | -1.1% |
| Install CVR | 0.154% | 0.200% | +29.8% |
| CPA (CPI, JPY) | 1,464.87 | 1,218.98 | -16.8% |

*Table 1. Campaign Performance Metrics (Autonomous AI vs. Human-in-the-Loop).*

These live campaign metrics highlight the limits of unsupervised autonomous AI and prove the core thesis of our hybrid architecture:

[1] All consumer insights and behavioural data used to inform the audience personas and creative development in this pilot were obtained after gaining the necessary permissions from users and in thorough adherence to Rakuten Group's Privacy Policy.

- The downstream conversion gap in autonomous AI: Under unsupervised autonomous loops (Group A), models tend to optimise for generic visual aesthetics and standard copywriting copy hooks. While these creative assets successfully generated exploratory interest (matching Group B in CTR), they failed downstream, converting fewer clicks into actual installs. This disconnect is evident in the lower Click-to-Install conversion rate (0.154%).
- The human curation performance boost: In hybrid Human-in-the-Loop (HITL) workflows, insights professionals act as vital governors. By validating, adjusting, and steering creative assets in real-time to align with verified local nuances (such as low-effort household budgeting for the household-manager persona or long-horizon wealth accumulation for the FIRE-oriented persona), they successfully bridged the conversion gap. Group B achieved a +62.5% lift in absolute app installs and a +29.8% relative CVR increase over the autonomous model.
- Direct marketer cost efficiencies: This conversion lift resulted in significant media efficiency. The CPA dropped from JPY 1,464.87 to JPY 1,218.98, representing a 16.8% cost reduction in user acquisition.

A two-sample z-test for proportions evaluated the statistical significance of the click-to-install conversion rates ($\hat{p}$), where the sample size comprised the total clicks per group ($n_A$ = 25,910; $n_B$ = 32,433):

$$z = \frac{(\hat{p}_B - \hat{p}_A)}{\sqrt{\hat{p}(1-\hat{p})(\frac{1}{n_B} + \frac{1}{n_A})}}$$

The z-test comparing Group B to Group A yields a directional positive lift ($z$ = 1.30, $p \approx 0.192$). While this does not cross the traditional statistical significance threshold ($p < 0.05$) due to the relatively short 16-day campaign window and low absolute conversion counts, the substantial +29.8% CVR lift and 16.8% CPA reduction present a compelling business case for organisations. Human curation effectively mitigates the generic, low-converting failures of unsupervised generative loops.

Beyond the quantitative performance results above, in-session feedback from the Rakuten Insight team surfaced two qualitative findings that we consider indicative of the collaborative value of an insight-grounded multi-agent pipeline:

- Accelerated execution: Creators reported that having AI-generated drafts as a concrete visual and copy starting point enabled them to move directly into refinement and stakeholder alignment, bypassing the exploratory ideation phase typically required when briefing from scratch.
- Uncovered new strategy: The Planner Agent surfaced persona archetypes that the internal team had not previously considered – most notably the FIRE-oriented professional persona for a household-budgeting app. Rakuten Insight described this as a genuinely new strategic direction, illustrating the exploratory value of AI-generated psychographic hypotheses grounded in insight data

### *4.2 Qualitative Research Results: The Consolidated Enterprise Adopter Framework*

To capture the operational, organisational, and strategic shifts among adopting organisations, we concluded a qualitative research protocol in June 2026. The study was conducted by Accenture and commissioned by Google. The study consisted of in-depth 1:1 interviews with N=10 C-suite marketing and MarTech stakeholders in Japan (seven brand-side buyers, three vendor/agency sides).

The qualitative research is structured around a finalised, consolidated discussion guide comprising four core methodological pillars:

1. Agentic AI maturity screening: Participants are screened and classified using a specialised Five-Level Agentic AI Maturity Grid:
    - Level 0 (Generative Interface): Ad-hoc content drafting and prompt-based ideation; user takes all manual actions.
    - Level 1 (Virtual Assistant): Workflow recommendation (e.g., A/B test setups) requiring explicit user confirmation.
    - Level 2 (Assistive Agent): Repetitive task automation (e.g., sentiment tagging in CRM, rule-based bid optimisation) under passive human oversight.
    - Level 3 (Semi-Auto Agent): Complex campaign plans proposed as a multi-channel calendar, executing immediately upon user approval.
    - Level 4 (Autonomous Agent): Goal-directed execution where the system independently creates assets, allocates budgets across channels (Search, Social, Display), and optimises bids in real-time to hit a target CPA.

We found that 50% of the sample already operates at Level 4 (Autonomous Agent), where AI manages campaigns within guardrails. However, the marketer's role has shifted from Operator to Orchestrator. While AI handles execution, human experts retain the critical "Judge" and "Decide" nodes due to the high risks associated with brand reputation and legal compliance.

> *"AI can make judgments and take actions, but only humans can take responsibility."*
>
> – Marketing Lead, Technology Company

> *"I still make the final judgment…determining the overall strategy. I consciously make sure to perform final checks and think things through myself."*
>
> – Head of Marketing, Technology Company

2. The CFO "Investor-Mode" Value Case: When forcing respondents to defend their investments, the CFO test ultimately collapses to Cost. *(CFO, Investor Mode test, was a mock test to push Marketers to defend or pitch to CFO or an Investor for Investment in Agentic AI across three dimensions of Cost, Time & ROI, differentiating between what is realised today vs. what is expected in next 12 months)*
    - Cost: Savings are real but finite, primarily driven by agency fee substitution and in-sourcing. As one respondent noted: *"You cut one agency and saved 50 million yen, okay. But what do you do after that?"* (Digital Strategy Lead, FMCG).
    - Time: Marketers reported massive time savings (20–40%), but these savings rarely reach the P&L. Freed staff often absorb the saved time rather than redeploying it: *"Work gets done much faster, so the team goes home early, instead of doing more in the time they were initially putting in."* (CMO, Technology Company).

- ROI: ROI lift is universally claimed, but almost no organisation successfully isolates Agentic AI's true contribution from broader media spend. "*It's difficult to translate the investment in agentic AI into revenue…there are several logical leaps. That makes it hard to get senior stakeholders' buy-in."* (Consultant, Services Firm).

3. Reality Check & Ecosystem Walkthrough: The true bottleneck to scale is no longer the AI model, but rather the data foundation. The industry is shifting from Content Creation to Decision Intelligence. However, because data remains heavily fragmented across silos (the shift from Tools to Ecosystems), the system cannot optimise the full value chain.

   *"If we had a system to properly collect advertising data into BigQuery for optimisation, that would've been ideal. But building it is extremely time-consuming and expensive, so we couldn't do it."*

   – Senior Marketer, Technology Company

4. Strategic Conviction & CFO Tripling Test: The study revealed that conviction does not scale linearly with maturity. Instead, a striking pattern emerged: the only true skeptics were the most mature, Level-4 brand-side deployers. Having pushed the technology into autonomous execution, they encountered the friction of proving hard ROI and the reality of hallucination risks. *("If something like a hallucination occurs, it is an instant failure,"* noted an FMCG Digital Strategy Lead). Conversely, mid-stage brand adopters (L2–L3) and sell-side vendors maintained highly optimistic, often unverified convictions about the technology's transformative potential.

## 5. Implications for the Insights Community

The emergence of Agentic AI elevates the role of the insights professional to the most critical component of the marketing pipeline. Traditionally, the researcher's job ended with the delivery of a slides report. Today, the researcher's output serves as the direct fuel for the entire marketing automation process.

Based on our findings, this shift has four core practical implications for the insights community:

- Curating the Fuel for Automation: The primary role of the insights manager is to curate authentic, deep-seated consumer truths and inject this real data directly into the AI's context window.
- Ensuring Zero-Hallucination: As highlighted in our executive interviews, the risk of hallucination is the biggest barrier to autonomous scale. By supplying robust, verified insights upfront and retaining the "Judge" node in the workflow, insights professionals act as the definitive guardrails that ensure brand safety and zero-hallucination in creative outputs.
- Preventing the Negative Cascade ("Insights are Gold"): As introduced earlier, a creative generative pipeline is only as strong as its starting inputs. If the initial consumer insights are flawed, generic or inaccurate, the entire multi-agent system will cascade in the wrong direction, magnifying the error at every step. This inevitably results in generic assets and low conversion performance. High-performance AI requires high-quality human insights.
- Co-creation vs. full automation: High-performance creative production is collaborative. Human oversight ensures that AI velocity is bounded by ethical standards, brand compliance, and cultural nuances.

## 6. "Planning with Gemini" and Agile Planning Innovation

The seven-agent system demonstrates a powerful sequential pipeline for execution. However, its success depends heavily on the quality of the initial insights. To address this, our research evolved into an elevated co-creation process for upstream strategic planning. "Planning with Gemini" – currently advanced by Google Japan G.K. – is a co-creation methodology wherein human marketers and AI collaborate to streamline the entire marketing planning lifecycle, moving directly into an agile planning framework.

The "Planning with Gemini" methodology redefines AI – not as a mere processing automator or an answer-generating engine – but as the marketer's "Thought Partner." Consequently, the three planning domains that previously operated in isolation are transformed into a non-linear, synchronous, and agile process:

- Insights: The process of visualising and structuring deep consumer psychology and emotional friction points from diverse market datasets.
- Creative: The process of objectively validating messaging angles and narrative structures derived from those insights, enabling rapid prototyping.
- Media: The process of simulating the user journey and holistically optimising budget allocations and distribution channels.

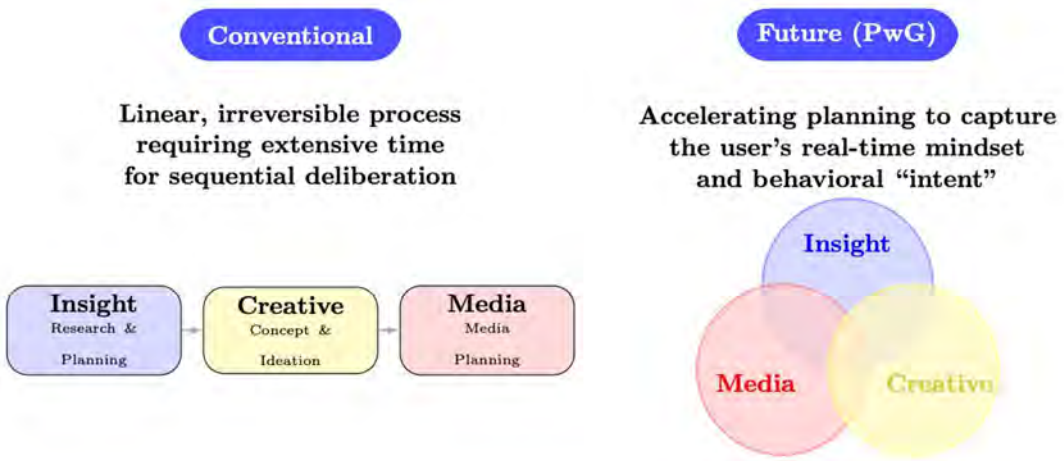



*Figure 3. Evolution from sequential process to co-creation cycle with AI.*

Under this methodology, these three domains are executed concurrently within the same shared context. For instance, updating a hypothesis regarding prospective consumer insights through dialogue with Gemini instantly triggers a coordinated re-evaluation of the creative messaging context and a re-simulation of optimal media touchpoints. Marketers retain full ownership of the critical "Judge" and "Decide" nodes, enabling them to iterate through hypothesis-generation and pre-validation PDCA (Plan-Do-Check-Act) cycles instantaneously, thereby securing the velocity and precision required to anticipate and lead market changes. The "Planning with Gemini" methodology has already been deployed across numerous enterprise customers, generating substantial business impact through "Human-AI Co-creation".

## 7. AI Disclosure and Disclaimer

*This is not an official Google product. It is a working prototype resulting from a cross-functional collaboration of Google's Senior Marketing Research Manager, Principal Analytical Consultant, Performance Specialist (CX), Creative Technologist and Creative Director, alongside Rakuten product developers and AI engineers, to demonstrate the strategic power of insights-driven creative production.*

*Disclosure: The underlying multi-agent system was developed with the assistance of the Gemini coding agent, and this paper was drafted with the assistance of the same agent. However, all system configurations, empirical campaign datasets, mathematical formulations and written manuscript content were highly revised and carefully validated by the human engineering team to ensure empirical accuracy, technical soundness and academic rigor.*

## References

Christensen, C. M., Hall, T., Dillon, K., & Duncan, D. S. (2016). *Competing Against Luck: The Story of Innovation and Customer Choice*. Harper Business.

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., . & Fung, P. (2023). Survey of hallucination in large language models. *ACM Computing Surveys*, 55(12), 1–38.

Kosinski, M., Stillwell, D., & Graepel, T. (2013). Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences*, 110(15), 5802–5805.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., ... & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems*, 33, 9459–9474.

Matz, S. C., Kosinski, M., Nave, G., & Stillwell, D. J. (2017). Psychological targeting as an effective approach to digital mass persuasion. *Proceedings of the National Academy of Sciences*, 114(48), 12714–12719.

## About the Authors

Minh Nguyen, PhD (AI), Senior Marketing Research Manager, Google, Japan.
Takumi Akuzawa, Creative Technologist, Google, Japan.
Togo Kida, Creative Director / Creative Development Lead, Google, Japan.
Toshiaki Kamiya, Performance Specialist (CX), Google, Japan.
Yuki Tanaka, Principal Analytical Consultant, Google, Japan.
Takamata Shoshi, General Manager, Rakuten Insight, Inc., Japan.
Otsuki Erika, Rakuten Insight, Inc., Japan.
Takeuchi Ami, Rakuten Insight, Inc., Japan.
Munish Bhamoriya, Growth Marketing & Commercial Leader, Accenture, Singapore.

# Gen Alpha Outsmarts Research

## Why They Won't Tell Us The Truth

*Jaime Solorzano*

*Rebecca Boden*

Esomar

Office address:

Burgemeester Stramanweg 105

1101 AA, Amsterdam

Phone: +31 20 664 21 41

Email: info@Esomar.org

Website: www.Esomar.org

Publication Date: September 2026

Esomar Publication Series Volume S424 Congress 2026

ISBN 92-831-0337-8

Copyright © **Esomar** 2026

Copyright © **Esomar** 2026

Congress 2026

# Gen Alpha Outsmarts Research

## Why They Won't Tell Us The Truth

*Jaime Solorzano*

*Rebecca Boden*

*"We do not misunderstand Gen Alpha. We misunderstand what we assume they will say".*

### The Old Assumption No Longer Holds

Most youth insight work is still built around a comforting idea: children are direct, transparent and comparatively unfiltered. In practice, this assumption produces clean decks, tidy narratives and a dangerous sense of certainty. It invites us to believe that stated values naturally predict behaviour, that interviews capture stable attitudes and that observation merely supports what verbalisation has already told us. That logic once felt efficient; it now feels increasingly fragile. Gen Alpha are not only media-savvy—they are context-savvy. They know when they are being watched, when a question is leading, when a brand is present behind the curtain and when an adult expects a particular kind of answer. This does not make them deceptive in a simplistic sense. It makes them socially intelligent participants in the research encounter. Their answers may be sincere, but sincerity is now braided with impression management, digital fluency, peer awareness and moral performance. The result is that the data often look mature, coherent and values-rich while still failing to predict what actually captures attention, drives imitation or shapes commercial response in everyday life.

That is the central tension of this paper: the better Gen Alpha become at understanding the systems around them, the less sufficient it is for research to rely on declared response alone. If the industry continues to treat polished articulation as the highest form of evidence, it will keep producing strategically elegant misunderstandings. In that sense, the challenge is methodological before it is generational. The problem is not that Gen Alpha refuse to tell us the truth. The problem is that the conditions under which we ask for truth increasingly invite performance. Classic qualitative logic often stops here. This paper argues that insight quality improves when observation re-enters the frame.

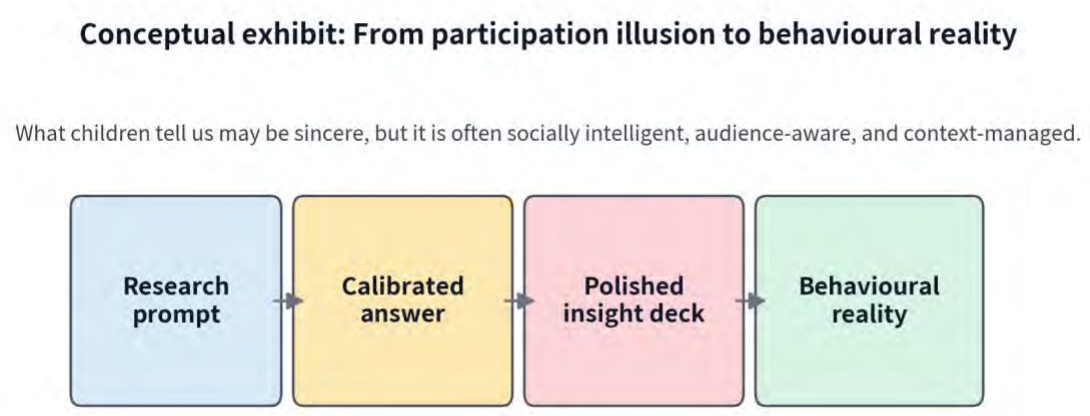



*Figure 1. The paper's central argument: classic qualitative logic often travels from prompt to polished answer too quickly, while behavioural reality remains under-observed.*

### Working Hypothesis

Copyright © **Esomar** 2026


This study begins from a simple but consequential proposition: performance is not the opposite of truth, but neither is it identical to behaviour. When Gen Alpha enter research, they frequently produce coherent narratives that are socially intelligent, method-aware and strategically adapted to adult interpretation. The task of research is therefore not to collect cleaner answers alone, but to understand how answers are shaped, what they conceal and what emerges when the research frame is temporarily loosened.

- Performance is not behaviour: Gen Alpha often calibrate responses, producing polished and convincing narratives that should not automatically be mistaken for conviction;
- Research literacy shapes data: awareness of the interview's purpose, of brand motive and of adult expectation changes what becomes sayable;
- Spontaneous behaviour reveals real drivers: unsignalled stimuli and in-the-moment exposure surface reactions that structured questioning often domesticates;
- Authority is negotiated: the adult-researcher dynamic is no longer fixed as children test, manage and tactically respond to it;
- Participation itself distorts understanding: research settings trigger performance mode and can over-index coherence at the expense of lived behaviour;
- A reverse-awareness approach is actionable: comparing calibrated and spontaneous response allows researchers to improve validity immediately.

## Methodological Approach

The study design is an experimental qualitative inquiry across Spain and Mexico with participants aged right to 15, 12 in-home and remote ethnographic sessions and experimental observations, six per market. The goal is not statistical representativeness, but rather methodological contrast to examine how the same participants appear under different research conditions and what becomes visible when conventional elicitation is placed alongside more behaviourally sensitive techniques. Each session combines several layers. First, a classic structured depth interview establishes the surface narrative a conventional study would likely collect. Second, device walkthroughs open a view onto actual digital environments: apps, games, short-form video feeds, browsing habits and the rhythms of attention across screens. Third, role-reversal exercises invite participants to design the research for adults, exposing what they believe research is, how they think adults judge them and what they would ask if the authority structure were reversed. Fourth, unsignalled stimulus exposure introduce ads, influencer content, game mechanics and gamified prompts without announcing their analytical purpose. Finally, reflexive researcher journalling documents shifts in tone, timing, compliance, resistance and moments in which observation changes the social atmosphere of the session itself.

This architecture matters because the research question is not only substantive; it is methodological. We want to compare articulated response with spontaneous reaction; polished explanation with observed behaviour; declared self with enacted self; and adult assumptions about authority with the more fluid authority actually operating in the encounter. The result has two outputs from the same participants. One is familiar: the kind of insight deck the industry often produces. The other is more difficult, but more revealing: a behavioural reality capture that preserves tension instead of smoothing it out. The cross-market design also matters. Spain and Mexico are not treated as decorative settings for a universal generational story. They function as two contexts in which digitally fluent childhood unfolds under different social rhythms, family negotiations and cultural expectations. That distinction is important because Gen Alpha may share platform conditions globally, but behavioural reality is always locally expressed.

## The Three Distortions Hidden Inside Conventional Youth Insight

| Distortion | Definition | Why it misleads strategy |
|---|---|---|
| **Calibration bias** | Participants adjust answers to align with perceived adult, social or brand expectations. | Brands may confuse well-formed narratives with stable motivation and |

| | | |
|---|---|---|
| | | design strategies around what sounds right rather than what actually drives response. |
| **Research literacy effect** | Awareness of commercial, social and methodological motives changes how participants frame what they say. | Insight teams may overestimate the innocence or transparency of the interview, and underestimate the degree of participant self-editing. |
| **Authority reversal** | Children subtly manage the interviewer dynamic, test control and shape the frame from within. | Researchers can misread compliance as openness and miss the social negotiation structuring the session. |

*Table 1. Three distortions.*

These three distortions (see Table 1) help explain why Gen Alpha insights can sound coherent, mature and purpose-driven while still failing in application. Conventional synthesis is not useless; it is incomplete. It translates the most legible parts of the session and often removes the very signals that would make the interpretation more valid.

This is where the phrase “participation illusion” becomes analytically important. Researchers may leave the session feeling they have seen authentic attitude. In reality, they may have witnessed a strategically competent performance shaped by the social fact of being researched.

## Findings

*“Clean. Coherent. Strategically usable. But is this what Gen Alpha actually give us—or what survives our methods?"*

Our findings would not suggest that Gen Alpha are impossible to understand. They will suggest that the default tools of understanding are increasingly vulnerable to the conditions they themselves create. Once children recognise the structure of the encounter, the encounter becomes part of the data. The question is no longer whether they are being honest. The more relevant question is what kind of social and methodological work their honesty is doing. To make this visible, we will present two outputs from the same sessions. Output A represents the kind of deck most experienced teams could credibly produce from these participants. Output B represents what emerges when contradiction, spontaneity, participant intelligence and the research frame itself are treated as evidence rather than noise.

## Output A: The Conventional Insight Deck

If this research is processed through a standard qualitative lens alone, the result could likely be strategically attractive. We could have three compelling personas and five convincing tensions. The work could sound contemporary, nuanced and commercially useful. It could also quietly flattened the very dynamics that make the sessions analytically interesting. That is precisely why Output A matters. It is not a straw man. It is a realistic account of what the industry often delivers when coherent articulation is treated as the primary evidence stream. Output A captures what survives classic synthesis. Output B captures what classic synthesis tends to remove.

| Persona | Surface narrative | Observed reality that Output A tends to soften |
|---|---|---|
| **The strategic performer** | "I answer what is correct, not what I do". | Calibrates answers to perceived adult or brand expectations, recites values fluently, avoids social contradiction, then shifts behaviour rapidly when observation loosens. |

Copyright © **Esomar** 2026


| | | |
|---|---|---|
| **The socially conscious informant** | "I care about fairness and the planet". | Recognises commercial motive, critiques brands indirectly and behaves far more adaptively than their values-led narrative suggests. |
| **The hidden spontaneist** | "I like games, memes, and fun". | Engages impulsively outside the interview frame, contradicts structured claims and shows strong awareness of surveillance and audience. |

*Table 2. Persona structure likely to emerge from a classic synthesis.*

### *The Five Tensions Output A Can Name but Not Fully Resolve*

1. Performance versus reality
2. Awareness versus impulse
3. Digital fluency versus fragmented attention
4. Transparency versus concealment
5. Play versus responsibility

These tensions are not wrong; they are simply too tidy when packaged as persona logic. They imply that contradiction can be stabilised into segments when, in reality, contradiction often operates within the same child, within the same session and even within the same minute. When auditing the raw material behind Output A, we find that classic synthesis systematically flattens three kinds of evidence: polished answers that masked impulsive choices; meta-commentary that reveal awareness of research and brand intent; and subtle acts of interviewer management that alter disclosure. The outcome is familiar: traditional insights capture performance, but not the deeper structure of what is driving that performance.

## The Behavioural Reality Approach

Output A's synthesis tends to remove contradiction and spontaneity, while participant intelligence and the research frame are treated as noise, rather than evidence. What is removed from Output A is essential to understand the underlying purpose and foundational context of this paper; Output B will later show the methodological move where observation is turned into insight rather than a mere illustration. The behavioural reality approach emerges from a practical recognition: when researching Gen Alpha, declared response is valuable but insufficient. It is one layer of reality, not the whole of it. Children do not enter the session as transparent carriers of stable attitudes; they enter as actors in a social situation they can read, interpret and influence. Behavioural reality therefore widens the evidentiary frame. It treats interviews, device behaviour, hesitation, switching, contradiction, peer references, jokes, self-corrections, glances, silences and adult-calibrated performance as parts of the same analytic object. Its aim is not to discredit what participants say, but rather to prevent researchers from granting stated content a monopoly on truth. In this approach, articulation is still data, but so is the choreography around articulation: the speed of response; the confidence of a claim; the moments of over-coherence; the shift that happens when attention is split; the micro-signs of compliance; the tactical display of moral intelligence; the instinctive move toward or away from certain content when it appears without warning.

This matters specifically with Gen Alpha because they have been socialised inside environments where visibility, audience and optimisation are everyday conditions. They learn early that different spaces invite different selves. A teacher, a parent, a creator, a game, a brand and a researcher each activate distinct codes of expression. This is why polished articulation should not automatically be treated as superior evidence. Often, it is the most filtered evidence. A child who says advertising does not matter may still demonstrate immediate fluency in creator aesthetics, recognition of sponsorship cues and a practiced instinct for platform-native branding. Another may articulate strong values but make behavioural choices in the moment according to convenience, play, peer energy or status management. These are not flaws in the participant. They are features of contemporary social life. The behavioural reality approach asks researchers to analyse this tension rather than erase it. Instead of deciding whether to trust speech or behaviour, it studies the relationship between them. Where do they align? Where do they diverge? Under

what conditions does coherence increase? Under what conditions does impulse break through? When does the adult frame tighten, and when does it loosen? The analytical payoff is significant because the gap between declaration and enactment is often where the real drivers become visible.

Three principles define the approach. First, observation is not secondary support for interview data; it is core evidence. Second, the research encounter is not neutral; adult presence, task structure and participant awareness all shape what emerges. Third, contradiction is not a defect to be edited out during synthesis; it is often the most commercially meaningful signal available. For marketers, this means a better understanding of how brands are actually encountered in life as lived. For researchers, it means a more serious treatment of participant intelligence. For brand managers, it means reducing the risk of strategies built on elegant articulation that collapses at the point of behaviour. In that sense, behavioural reality is not an anti-qualitative provocation. It is a defence of qualitative seriousness in the age of algorithmically literate participants. It protects the depth of the discipline by insisting that insight should be earned through attention to context, performance, environment and the social conditions of disclosure. It replaces the false comfort of neat answers with the more demanding but more useful task of reading reality in motion. "The researched now understand the research. Our methods must become intelligent enough to understand what that changes".

## Output B: The Behavioural Reality Capture

Output B is designed as a direct alternative to tidy synthesis. Instead of smoothing tension into stable personas or forcing contradiction into a narrative of developmental inconsistency, we build a structured capture of what becomes observable when children interact, react, recalibrate and negotiate the research setting in real time. No personas. No tidy resolutions. Only structured, observable realities. This is not an argument for chaos. The behavioural reality capture is highly structured. It organises the evidence differently. Rather than grouping children by stable identity stories alone, it maps the mechanisms through which behaviour becomes visible, concealed, staged or socially managed. That shift matters because it creates a more useful bridge between ethnographic reality and commercial interpretation. Instead of asking, “Which persona is this child?” the more strategic question becomes, “What conditions produce this behaviour, this calibration, this contradiction and this response to branded stimuli?”

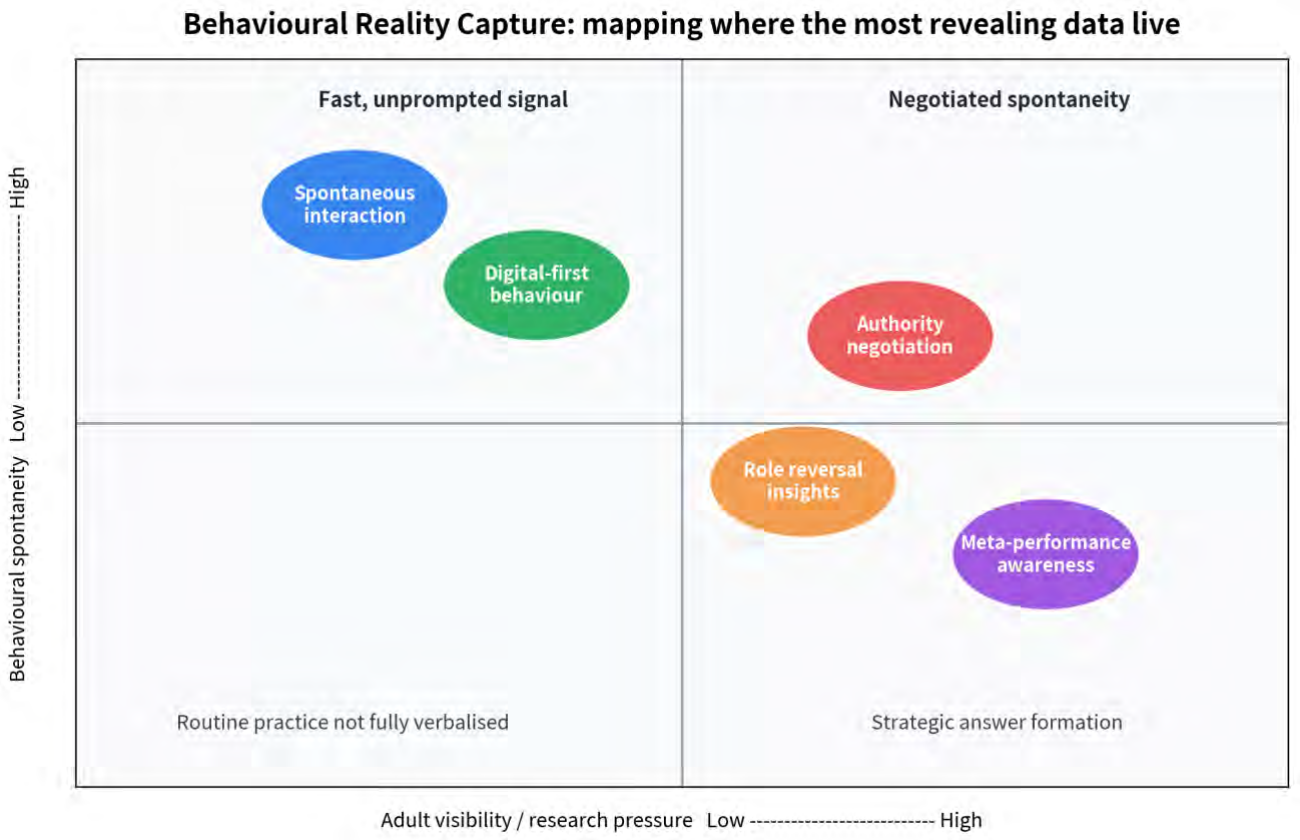



*Figure 2. Output B maps the most revealing evidence according to spontaneity and research pressure, rather than reducing it to stable identity categories.*

| Mapping layer | What becomes visible | Why it matters to marketers, researchers and brand managers |
|---|---|---|
| **Spontaneous interaction** | Real-time reactions to ads, influencers, games, aesthetics and references before participants could turn them into polished explanation. | This is where salience, familiarity, excitement, rejection and symbolic charge become visible before rationalisation reorganises them. |
| **Meta-performance awareness** | Children monitoring their own answers, managing how they appear to adults and strategically shaping what counts as a good response. | This reveals that reflexivity is part of the phenomenon. Insight teams must analyse not only content but also calibration. |
| **Authority negotiation** | Subtle testing of boundaries, selective compliance, joking resistance, permission checking, tactical silence and micro-acts of control. | Adult power is fluid, not fixed. The structure of authority directly affects disclosure and apparent confidence. |
| **Digital-first behaviour** | Multi-tasking, rapid device switching, creator literacy, algorithmic awareness and sensitivity to platform-native codes of relevance. | Brands compete inside this environment, not outside it. Exposure is easy; meaningful cultural fit is not. |
| **Role reversal insights** | Moments in which children redesign the research, question the frame or show how they would study adults instead. | This exposes the distance between adult method design and child reality, making the limits of conventional research newly visible. |

Copyright © **Esomar** 2026


*Table 3. What we map.*

## *Analytical Reading off the Capture*

The first mapping layer, spontaneous interaction, proves critical because it surfaces attentional truth before narrative repair. When a stimulus arrives without warning, children often reveal immediate recognition, indifference, delight, suspicion, imitation or dismissal in ways that are more behaviourally informative than later explanation. A visual style, a creator voice, a game mechanic, a meme format or a branded cue could carry instant meaning long before the participant explains why. That first movement matters. It tells us what is already alive in memory, what is socially coded as relevant and what the child's system treats as noise. The second layer, meta-performance awareness, shows that many participants are not merely answering questions. They are actively managing the answer as a social object. Some optimise for maturity. Some optimise for morality. Some optimise for smartness, humour or self-control. In several moments, the answer itself appears less revealing than the sophistication with which the child adjusts it. This is a crucial point for qualitative practice. When a participant is highly aware of how they are being read, the resulting articulation cannot be interpreted as a pure statement of preference. It must be read as both expression and self-presentation.

Authority negotiation adds a third layer of insight. Adult power does not disappear, but it becomes unstable and negotiated. Children test limits through play, delay, strategic vagueness, selective openness and subtle boundary management. Some check what adults would tolerate. Others perform compliance while preserving a private zone of control. These are not side notes. They change what can be observed and what remains withheld. They also warn us against a common professional mistake: assuming that the formal power of the researcher guarantees access to authentic disclosure. The digital-first layer clarifies the commercial stakes. Across the sessions, digital life does not appear as a separate topic. It functions as the atmosphere of decision-making. Children switch quickly, read interfaces fluently, recognise monetisation cues and differentiate between what feels native to their world and what feels inserted by adults trying to imitate it. This matters because many brand failures begin with a category error: treating recognition as engagement or engagement as conviction. Gen Alpha may understand a brand perfectly and still decline to grant it relevance. Role reversal, finally, is methodologically revelatory. When children are invited to design research for adults, they expose what they believed adults hide, fake or perform. In doing so, they also expose what they themselves understand about research as a system. The role-reversal task becomes more than a playful exercise. It becomes a mirror. It shows that participants are not passive subjects awaiting extraction. They are active interpreters of the institutional script. Once that becomes clear, the ethics and craft of youth research both need to evolve.

Taken together, Output B produces a more demanding but more honest picture of Gen Alpha. It suggests that coherence is often an artifact of method, that contradiction is structurally informative and that insight improves when the session is analysed as a social event rather than as a neutral pipeline for attitude collection.

| Dimension | Output A: Conventional deck | Output B: Behavioural reality capture |
|---|---|---|
| **Primary unit of analysis** | Persona and articulated tension | Behavioural mechanism and observable condition |
| **Typical evidence** | Structured answers and coherent narratives | Spontaneous reactions, calibration, timing, switching, negotiation |
| **View of contradiction** | Something to reconcile | Something to analyse |
| **Role of the researcher** | Collector and interpreter | Active condition inside the data |
| **Commercial outcome** | Clear but potentially over-smoothed strategy | More valid, context-sensitive strategy anchored in behaviour |

*Table 4. Comparative exhibit: Output A versus Output B.*

Copyright © **Esomar** 2026


### *Key Insights From Output B*

Performance does not cancel behaviour; it is one form of behaviour. What children say should not be dismissed, but it must be interpreted alongside the audience they imagine, the social work the answer performs and the conditions under which the answer is produced. Meta-awareness shapes the data before the researcher begins to analyse it. When participants already understand research logic, the interview is never a neutral mirror. It is a stage with consequences. Authority is negotiated and methodologically consequential. Adult presence affects candour, humour, dissent and the willingness to expose contradictory impulses. Spontaneous behaviour is critical because it captures drivers polished discourse often conceals. Fast reactions, switching behaviour and unprompted recognitions reveal what is culturally alive before it been cleaned into strategy language. Participation itself distorts understanding. Classic qualitative methods often underestimate how strongly the act of being researched activates calibration. The distortion is not a nuisance variable alone; it is evidence of social intelligence. Digital fluency is social fluency. Gen Alpha read platform codes, authenticity cues, monetisation signals and reputational risk with striking sophistication. Brand teams who ignore this will misread both enthusiasm and rejection.

## The Reverse Awareness Protocol

If Output B provides the diagnosis, the reverse awareness protocol provides the operational response. It is designed to be useful tomorrow, not only admirable in theory. Its purpose is to help insight teams redesign youth research so that meta-awareness is not ignored, spontaneous reaction is not left to chance and behavioural contradiction can be turned into analytical value instead of cleaned away in debrief. The protocol does not require abandoning depth interviews or rebuilding every research system from zero. It requires a shift in sequencing and interpretation. It asks teams to create moments in which participants are not only asked to explain themselves but are also observed navigating stimuli, reframing tasks and revealing how they manage the adult gaze.

| Protocol element | What to do | Why it improves validity |
|---|---|---|
| **Unsignalled stimulus insertion** | Introduce ads, creators, visuals, game mechanics or prompts without over-framing their purpose. | Captures instinctive attention and emotional reaction before socially desirable explanation takes over. |
| **Role-reversal exercises** | Ask participants to co-design the survey, interview or research task for adults. | Reveals participant theories of research, exposes hidden assumptions and surfaces social intelligence. |
| **Articulated versus spontaneous comparison** | Contrast what participants declare with how they respond under looser observation. | Makes calibration visible and prevents over-reliance on polished verbalisation. |
| **Audit the "too perfect" narrative** | Interrogate highly coherent, values-rich, or suspiciously finished responses instead of rewarding them automatically. | Protects teams from mistaking a well-performed answer for a stable behavioural truth. |
| **Iterative observation** | Blend structured prompts with unstructured moments, device behaviour and reflexive field notes. | Keeps insight anchored in lived behaviour rather than single-mode testimony. |

*Table 5*

Used well, the reverse awareness protocol does more than improve methodological validity. It changes the quality of strategic recommendation. Teams become less likely to build innovation pipelines around articulated ideals that never survive context, and more likely to identify the real cues, frictions, rituals and moments of permission that shape behavioural uptake. Its importance will only increase. As younger participants become more capable of reading institutional systems, all research that depends on declared transparency will face the same challenge now visible in youth work. Gen Alpha are simply the clearest early warning signal.

Copyright © **Esomar** 2026


## Implications For Marketers, Researchers and Brand Managers

### *For Marketers*

The paper warns against over-trusting explicit preference statements in categories where social signalling, peer legitimacy and platform-native behaviour shape outcomes more strongly than stable verbalised attitudes do. If brands build strategy only around what children can explain to adults, they risk producing communications that test well in language, but fail in culture. The more useful question is not simply what Gen Alpha claim to value, but what they reward with attention, what they dismiss instantly, what they share performatively, what they privately keep using and what conditions allow a brand to move from recognition into relevance.

### *For Researchers*

The implication for qualitative practice is not methodological despair; it is methodological maturity. Interviews still matter. Co-creation still matters. Reflexive discussion still matters, but they must be interpreted inside a wider behavioural frame that includes adult pressure, participant awareness, digital environment and the social labour of answer-making. Researchers who continue to treat participant calibration as contamination will miss one of the richest developments in contemporary fieldwork: participants are telling us, through their performance, how they understand the systems that seek to understand them.

### *For Brand Managers*

For brand managers, the message is practical. Better understanding of behavioural reality improves how brands test ideas, read friction, identify actual moments of influence and judge whether a concept is merely acceptable in research or genuinely alive in use. The gap between articulated approval and behavioural adoption is often where money is lost. This paper offers a way to narrow that gap. It also suggests that better youth strategy will come less from asking children to speak the language of brand frameworks and more from observing how brand meaning is assembled in lived contexts of play, peer visibility, family negotiation and algorithmic exposure.

## Key Learnings and Takeaways

- Conventional interviews often capture performance before they capture behaviour;
- Research literacy changes what Gen Alpha are willing and able to say;
- Spontaneous reaction is not peripheral evidence—it is often the closest available route to the real driver;
- Authority is negotiated in-session and must be analysed, not assumed away;
- Participation itself can distort understanding, which means method design must become part of the interpretation;
- Small changes in research craft can dramatically improve validity and strategic usefulness;
- The evolution of human understanding now requires recognising that the researched understand the research.

*“If you do not evolve your methods, your insight is just performance observation".*

The broader significance of this paper extends beyond youth research. Gen Alpha simply make visible a condition that is spreading across contemporary society: participants are increasingly aware of how institutions categorise, evaluate and monetise them. The more research ignores that fact, the more it will confuse articulate self-management with transparent truth. That is why this paper is ultimately optimistic. It does not argue that understanding has become impossible. It argues that our craft must catch up with the people we study. When we design for meta-awareness, compare articulation with behaviour and treat contradiction as a clue rather than an inconvenience, the quality of insight improves. We move closer to reality as lived, not merely reality as performed for research.

Copyright © **Esomar** 2026



For ESOMAR, and for an industry committed to the evolution of human understanding, that is the challenge and the opportunity. We are no longer studying passive respondents. We are studying participants who can read the system. The future of insight depends on whether our methods can do the same.

## About the Authors

Jaime Solorzano, LATAM Regional Manager & Global Head of Insights, Q2Q Global, LATAM Office, Mexico City.

Rebecca Boden, Senior Vice President of Partner Solutions, Q2Q Global, UK Office.

# Creative Consequences

## Why the World's Strongest Brands Create the World's Best Advertising

*Vera Sidlova*

*Polly Wyn-Jones*

*Horacio Gonzalez*

*Felipe Thomaz*

*Neta Livneh*

Esomar

Office address:

Burgemeester Stramanweg 105

1101 AA, Amsterdam

Phone: +31 20 664 21 41

Email: info@Esomar.org

Website: www.Esomar.org

Publication Date: September 2026

Esomar Publication Series Volume S424 Congress 2026

ISBN 92-831-0337-8

Copyright © **Esomar** 2026

Copyright © **Esomar** 2026

Congress 2026

# Creative Consequences

## Why the World's Strongest Brands Create the World's Best Advertising

*Vera Sidlova*

*Polly Wyn-Jones*

*Horacio Gonzalez*

*Felipe Thomaz*

*Neta Livneh*

## Introduction

Creativity is the industry's most celebrated asset, but least understood growth driver. Does creative drive brand or is it the brand name that draws an audience? It is likely that both play their part in the high stakes of building brand growth. Brands now have more opportunities than ever to expand their role into a multitude of new and evolving touchpoints to broaden their reach, messaging and ability to build affinity. The current surge in what advanced creative techniques new AI tools are bringing to the industry, expanding mindsets and imaginations like never before, mean creatively reaching the consumer psyche and the algorithms behind LLM's is now an even more exciting yet complex proposition. The ability to create what were once multimillion-dollar film set creations with just a few clicks, leaves opportunity and imagination wide open as AI continues to lower production barriers and enable a new level of scale and experimentation. So, with everything to play for, the challenge for marketers lies in understanding which creative choices truly contribute towards building long-term brand equity. This paper brings together two complementary studies to show that creative quality can predict future brand equity, and that the cohesion of creative elements within ads (how elements work alongside each other) helps shape likely brand outcomes. Together, they demonstrate how creative systems translate into brand growth.

The first, conducted by Kantar, brings together the scale of the LINK+ and BrandZ databases to provide data-driven evidence of the relationship between creative quality and brand strength. Drawing on more than a decade of Kantar data, the analysis examines whether the world's strongest brands also create the world's most effective advertising, how these effects develop over time and how marketers can use this evidence to make more informed decisions. The second study, conducted by Saïd Business School at the University of Oxford, examines the relationship between brand and creative structure at scale. It extends the analysis beyond individual ads to explore how creative elements combine, how those combinations relate to brand outcomes and how moments of disruption, such as an agency change, can reshape the structure of both creative work and the brand system around it.

## Data and Methodology, Kantar

The data used for the Kantar analysis is based on a large global dataset originating from our LINK+ and BrandZ databases. It spans a 10-year period covering data points across 127 months from March 2015 to September 2025. The size of the data provides geographic diversity across 13 countries and spans multiple industries. A total of 110 brands were identified that appeared in both datasets over time. Quality assurance required a minimum threshold of five months of observed data per brand across the full time-window to ensure statistical reliability. This resulted in the matching of over 8,500 LINK+ researched ads, with the breadth of the data set providing a robust foundation for identifying generalised patterns while also enabling category-specific analysis. The BrandZ data underwent Kalman smoothing to convert sparse measurement points into consistent monthly series data suitable for time-aligned statistical comparisons. Date standardisation assigned each measurement to the

15th of the month, creating an anchor point for all brands and enabling synchronised cross-brand comparisons. This resulted in 14,000 data points. The Kalman smoother denoising algorithm filled gaps between measurements, producing complete monthly series data while preserving the integrity of the data and reducing noise within it. LINK+ ad-level results were aggregated to monthly brand averages and denoised using the same Kalman methodology to ensure direct comparability with the BrandZ dataset. Monthly brand aggregation averaged all ads per brand, category and country within each month, matching BrandZ's temporal granularity.

Finally, to account for the time between exposure to advertising and subsequent changes in brand equity, a lag analysis was conducted across a 0–24-month window. This approach identifies the time delay at which the relationship between creative performance and brand equity is strongest, allowing both the strength and timing of the relationship to be understood.

## Data and Methodology, Saïd Business School, University of Oxford

The Creativity Space study, undertaken by Saïd Business School at the University of Oxford, used the full historical LINK+ dataset to identify and connect creative elements identified via LINK AI. Over 130,000 ads were included, over the same 10-year period used in the Kantar dataset, spanning around 80 markets, from more than 18,000 brands across hundreds of categories. This scale enabled creative structure to be examined as a broad, representative system rather than as a set of isolated executions. The resulting learning was then applied to the 110 brands matched across the LINK+ and BrandZ Kantar dataset, allowing the relationship between creative structure and brand outcomes to be explored on a consistent analytical base. A second dataset supplied by Kantar identified agency-switch events, together with the relevant time periods and brands affected. This enabled the analysis to examine how creative structure responds to a defined point of disruption, and how changes such as an agency transition may influence both the composition of creative work and its subsequent relationship with brand equity.

## Linking Creative Quality and Brand Equity? - Kantar

The matched predictive and in-market data also provided further validated evidence that Kantar's LINK+ measures align with brand equity. LINK+'s long-term metric, Demand Power, was proven to predict how an ad will work to help build the brand into the future, by showing the strongest alignment at 0.48 correlation between the two data sets. Looking more closely at the components of Demand Power, we saw correlations of 0.43 for "meaningful", 0.40 for "salience/impact" and 0.30 for "difference". This confirmed a clear and systematic relationship between creative quality and brand equity, indicating that creative effectiveness measures, particularly those related to brand affinity and impact, can act as leading indicators of future brand strength. The slightly weaker relationship for differentiation suggests that not all aspects of brand equity are equally shaped by advertising alone.

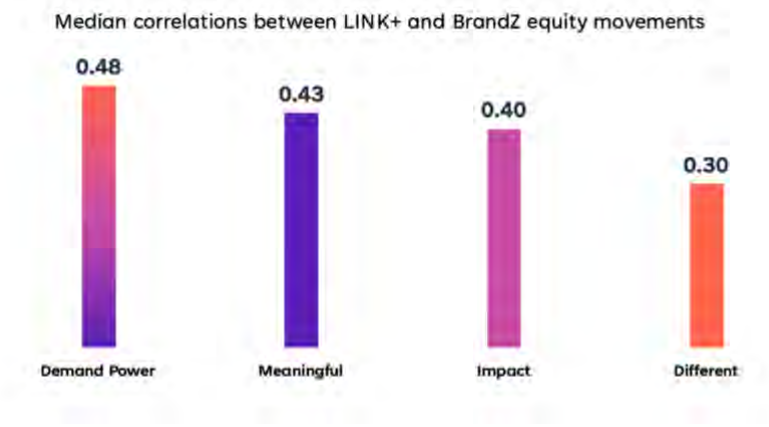



*Figure 1*

A key finding from the study showed us that creativity works on two clocks; some brands see effects almost immediately and others build equity much more gradually over time. The lag analysis conducted, to account for the time between when the creative asset is seen and its eventual impact on the brand, showed that on average, changes in brand equity follow changes in creative quality by around 10 to 11 months. The comprehensive lag search between 0 and 24 months identified the optimal predictive timing relationships between the LINK+ and the BrandZ data sets. Predictive strength and lag timing varied slightly across the different components of Demand Power and by category, requiring tailored approaches for different brand/category combinations to maximise forecasting accuracy. However, this provides a useful planning horizon for marketers, indicating that improvements in creative performance today are likely to influence brand equity within the following year. However, these averages mask important variations. The distribution of lag timings reveals a bimodal pattern, with two distinct peaks:

- Immediate movement, where creative performance and brand equity move in tandem, potentially reflecting underlying brand strength;
- Delayed cumulative effect, where creative builds brand equity more gradually over time.

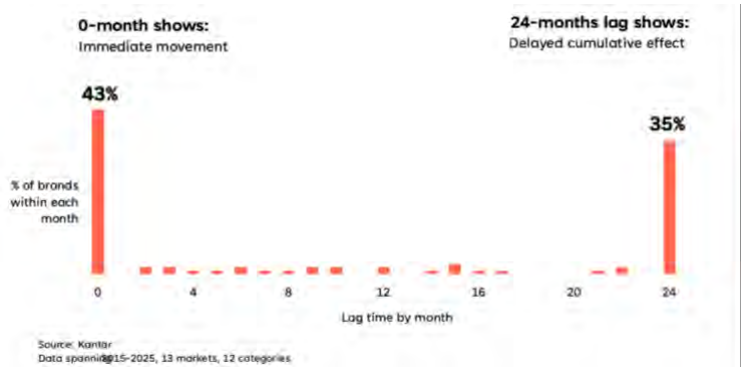



*Figure 2*

For marketers, this distinction is important. Brands that exhibit immediate movement can respond quickly to changes in creative quality, whereas those with longer lag effects require greater consistency and patience for effects to materialise. It is important that brands understand which pattern applies to their specific category and metric combination. A brand showing immediate co-movement needs a different planning approach than one with a two-year lag.

## *Starting Point and Direction Of Travel*

Beyond the two-clock timing, both the starting level of creative effectiveness and its direction of change are associated with brand outcomes. The quadrants below (see Figures 3 and 4), measure the initial start point of where a brand sits in terms of its predicted likelihood to build equity for the brand (Power), versus the average change in that equity measured in percentage points per year. The growth percentages on both quadrants represent two years. For brands with shorter lag times, stronger starting creative is associated with stronger subsequent growth. Short-lag brands steer quickly, but if ads start below average and do not improve, brands risk stagnation or decline. The quadrant below shows that the initial starting point, regardless of marketing that followed, was a crucial predictor of whether the brand was likely to grow over a two-year period. Starting with a lower brand equity prediction level resulted in lower brand growth in general, starting with a higher prediction showed strong positive brand growth. Brands that started with higher predictions, but did not move into the higher average change group, still showed overall higher growth than those starting with low brand equity predictions highlighting the importance of starting strong. The total brand set fell evenly across the quadrants.

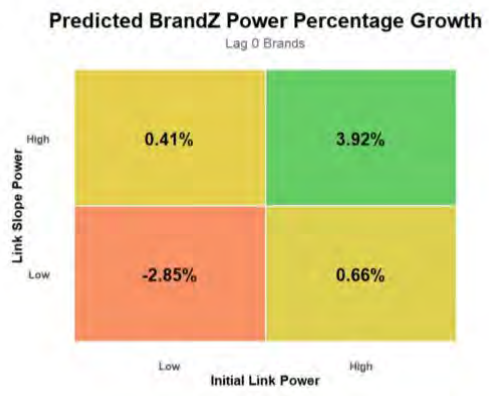



*Figure 3*

For brands with longer lag times, the relationship is more gradual but follows a similar principle, the starting point is again important: improving weak creative slows decline rather than reversing it. Brands with a longer lag are slower to change course. However, if your ads are good and continue to get better, future brand growth is much more likely. Brands in this group also fell out evenly across the quadrants, but we can see that the "no lag" group had a much stronger relationship between the two data sets than the brands with a longer lag time. This makes perfect sense as the longer an ad takes to have an effect, the more likely other elements will play a role, such as competitor activity, other promotional activity, spend levels, etc., all of which are not accounted for in the data sets.

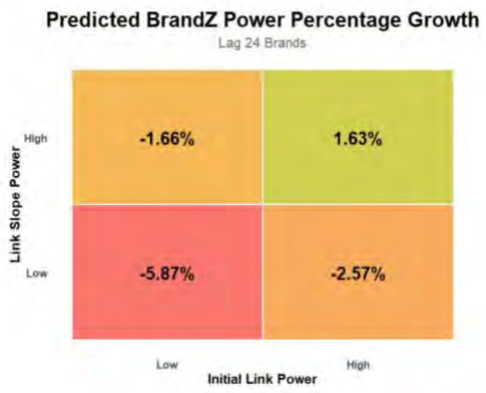



*Figure 4*

Across both groups, creative quality appears to function not only as a reflection of current brand strength, but also as an early signal of future performance. Overall, the analysis validates a systematic association between creative quality and subsequent brand equity, with Demand Power emerging as the most robust predictor of long-term equity development. Importantly, this relationship is not uniform in its temporal pattern: for some brands, the effects of creative quality are observable with little or no lag, while for others they accumulate more gradually over an extended period. Across both trajectories, the initial strength of the creative and its direction of change are critical factors, suggesting that ad quality functions not only as a current marker of brand strength, but also as an early indicator of future brand performance.

## Structural Analysis of Creative and Brand - Saïd Business School, University of Oxford

### *The Creative Space Study*

The second stage of the analysis, the Creativity Space structure, shifts from measuring creative effectiveness to understanding its internal structure via use of the extended data set. Each ad can be seen as a combination of elements: some fixed (such as brand, category, market, the

brand's status in that market, ad format), and others more fluid (such as tone, pacing, setting or creative devices). By modelling these elements and identifying where they co-occur, it becomes possible to map creative as a relational structured space, rather than assess it purely at an overall level or via the individual creative elements in each ad.

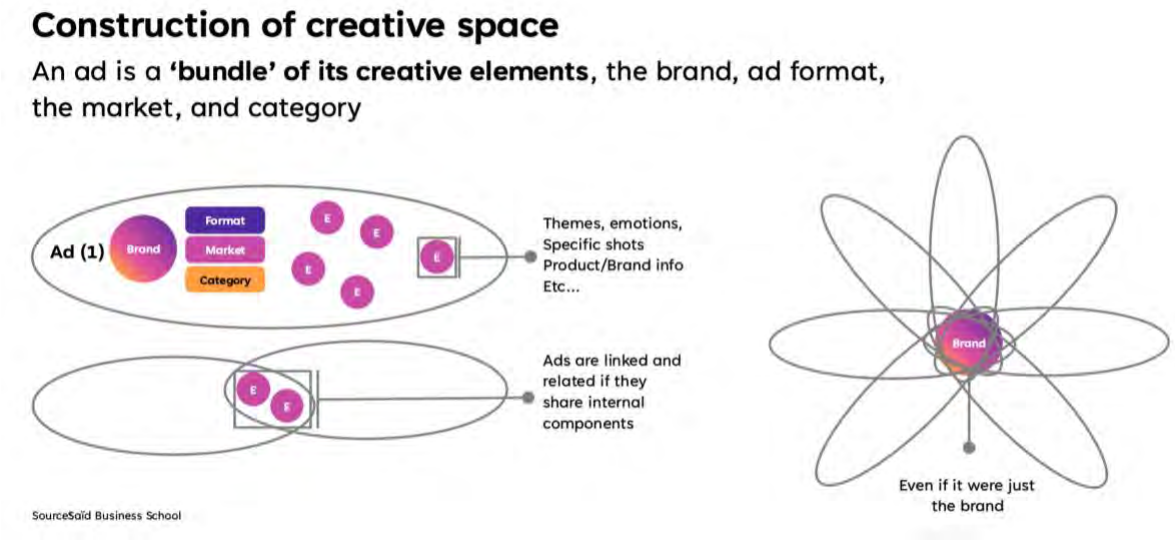



*Figure 5*

Creative space-structures were built using the creative elements identified via LINK AI. One dynamic structure per year was built resulting in millions of interconnecting relationships.

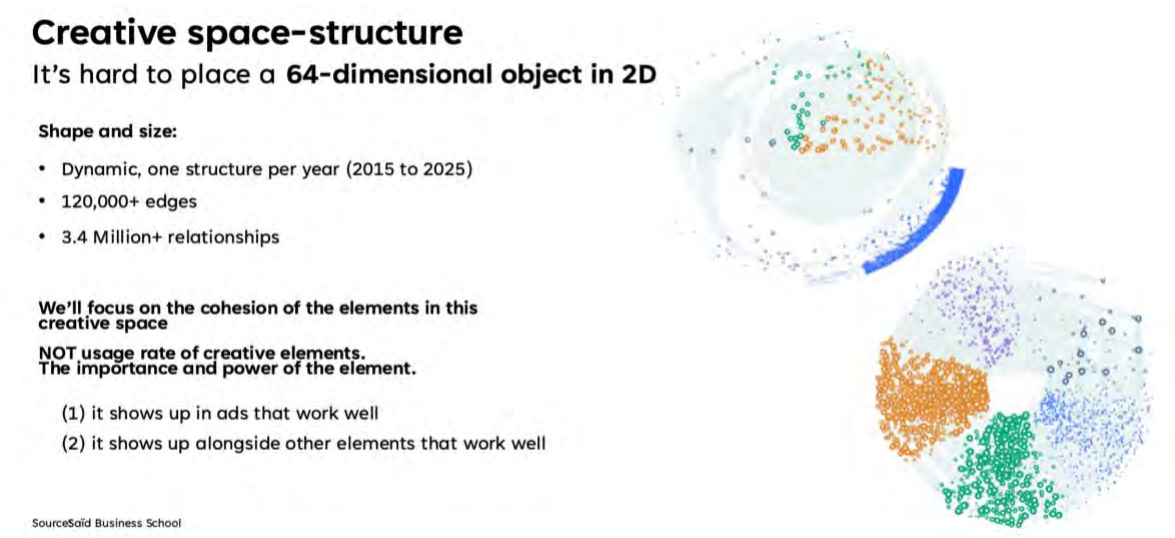



*Figure 6*

Within this space, the focus moves to cohesion; an element-level measure that accounts simultaneously for how crucial an element is and how strongly it relates to other important elements. This allowed for a more nuanced understanding of how creative choices work together, and how specific combinations of creative elements created better outcomes, rather than examining them in isolation. For example, product-focused elements, such as demonstrating, describing or showcasing the product form a highly cohesive cluster associated with both short-term sales potential and long-term brand equity. These combinations appear frequently and consistently among high-performing ads falling into the top 90% of cohesive elements, and were evident across 35% of all ads.

*Figure 7*

Other combinations are less obvious. Elements such as sustainability messaging show relatively low standalone cohesion, but become more effective when combined with complementary elements such as human relationships, innovation or everyday settings. This highlights the importance of context and combination, rather than relying on single cues.

## *The Impact of Cohesion on Brand Outcomes*

To get to a level of cohesion for each ad, a score is calculated by the sum of all the creative element cohesion values in the ad, with further aggregations possible by brand, time, market, etc. Then, to examine the effect of brand-level impact of cohesion across its ads, we regress brand metrics as a function of a change in cohesion score, while accounting for where a brand sat in the previous year, yearly time effect, the market and category effects, as well as the number of ads included for that brand (see Figure 8). Higher cohesion scores, those in the top quartile of the distribution, are associated with stronger brand outcomes across multiple dimensions, including increases in likelihood of choice, meaningfulness, salience and measures related to consumer preference and pricing strength. However, we should also be cautious not to accept that this result suggests that all creative efforts should converge to the same effective and highly cohesive sets, and much less to the exact same execution. We must recall that when similar combinations of elements become widespread, for example during the COVID-19 global health pandemic when many brands used the "We'll get through this together" slogan, their impact diminished. Elements as described here are more general (themes, pacing, shots, logo representation, etc.), and their cohesion is structural, leaving room for executional variation without value-destroying creativity loss and homogeneity.

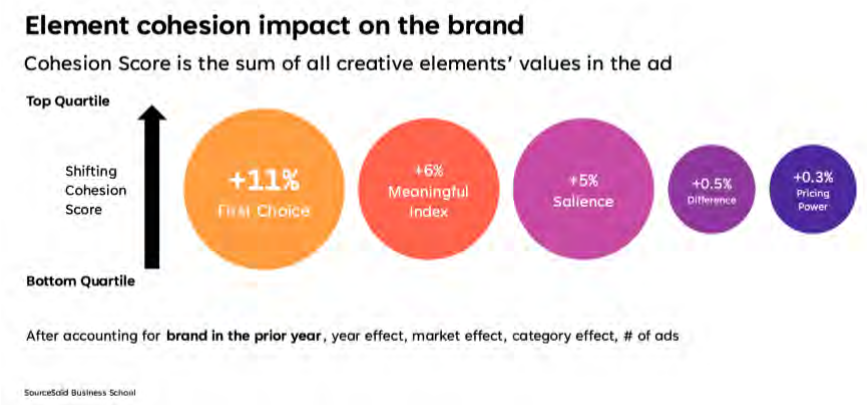



*Figure 8*

Copyright © **Esomar** 2026



### *Tracking Cohesion Over Time*

The 10-year data period allowed us to assess cohesion over time. This shows us whether a brand is using the right creative elements at the right moment, and when those elements begin to lose their effectiveness. Representing creative structure as a year-by-year, time-indexed view allows for genuine longitudinal analysis, rather than relying on a static snapshot, proving that even elements that work well can eventually plateau or decay in value. Cohesion therefore needs to be actively refreshed, not simply banked.

### *Disruption: What Happens After an Agency Switch?*

To examine how creative structure responds to a change in strategy, we use an event-analysis structure with an agency switch as our target event. This creates a natural point of disruption from which to observe what happens to creative cohesion, brand equity, and longer-term performance. Creative space-structures were set up as follows:

- History: How many elements do they use per ad?
- Prior creative structure: How many unique elements to they use across all ads?
- The agency switch event: What is the change in average cohesion score?
- Test of outcome: What happens to brand equity?
- Future consequence: What happens in the long run?

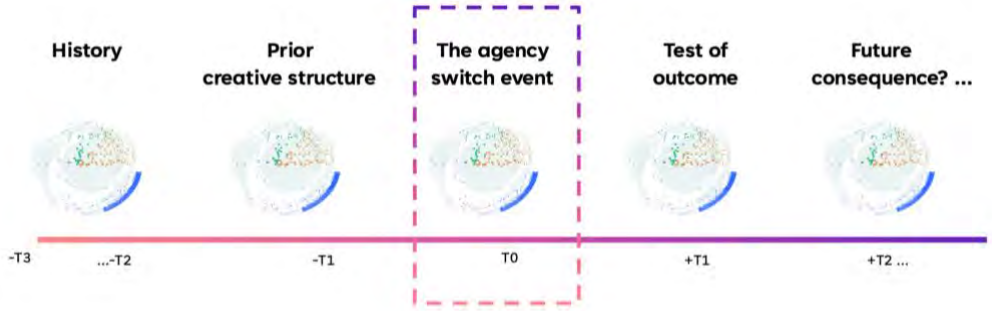



*Figure 9*

Following an agency switch, cohesion falls sharply, with an overall decline of -35.2%. Interestingly, post-switch ads show no change in the number of elements used per ad, but there is a slight -2.5 decline in the number of unique elements. In practice, this suggests a narrowing and consolidation of the creative toolkit, while new agencies additionally explore peripheral components in search of new creative innovation (lower cohesion tends to indicate underused, disconnected, novel unproven ideas and similarly disconnected to elements that themselves have high cohesion).

### *The Immediate Trade-Off: Differentiation Rises, Trust Falls*

Although there is a fall in cohesion points to a loss of structural consistency, some brand measures improve in the short term. After an agency switch, brands show a more than 3.28% increase in difference, a more than 2.1% increase in willingness to pay more, as well as a more than 3.4% increase in consideration. However, trust falls by -2.8 points, suggesting that a “new idea shock” can create both opportunity and risk, particularly if the brand is being repositioned. For marketers, this is a critical trade-off: agency change can inject distinctiveness and momentum, but it may also unsettle established trust.

### *The Longer-Term Consequence*

Over the longer-term, agency change looks less like a permanent loss and more like a time-limited, cyclical disruption. After two to three years, 80% of brands regain some breadth in the creative elements they use and 50% return to the same level of breadth as before the switch. These may not be the same elements, but the overall creative range recovers after an initial fall in use of unique elements that was seen immediately post agency switch.

*Figure 10*

Importantly, 30% of brands exceed their pre-switch cohesion levels after two to three years, with an average gain of 19.5% versus their previous position. These are the brands that appear to have searched for innovation and found new combinations that worked. However, the fact that 70% do not reach this level shows that agency change is not a guarantee of creative renewal. Understanding why some brands convert disruption into stronger cohesion, while most do not, is an important area for further analysis.

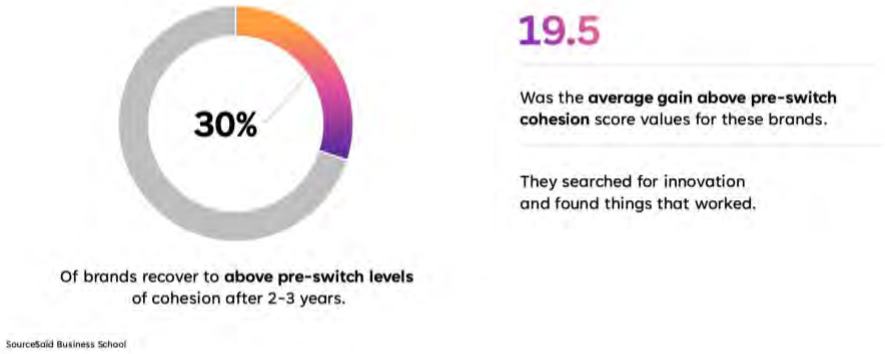



*Figure 11*

Kantar's vast dataset has enabled us to find a novel way to map creative elements, their interactions and the overall consequence of how they affect brand equity. It has allowed us to define the structural cohesion of "creative space" to explain their overall value, particularly on brands being consumers' first choice, which aligns with market share and likelihood to purchase. The shock from agency switch on the creative space-structure is not so straight forward. We see an immediate loss of cohesion, and an overall narrowing of creative focus. At the same time some gains are seen in differentiation, consideration and a willingness to pay more for a brand. However, these come at the risk of a loss of trust. Over time, around two to three years, it is possible to regain some breadth in the use of different creative elements and potentially exceed earlier cohesion. However, the cyclical disruption of an agency switch may mean it is then time to start the process again.

## Summary and Recommendations

Copyright © **Esomar** 2026


This paper shows that creative effectiveness is not an isolated communications outcome, but a measurable contributor to brand equity and long-term brand growth. By bringing together more than a decade of Kantar LINK+ and BrandZ data, the analysis provides robust evidence that stronger predicted creative quality is systematically associated with stronger future brand outcomes. Demand Power emerges as the clearest predictive measure, with meaningfulness and salience (components of it) also showing strong alignment, confirming that advertising which builds affinity, impact and relevance can act as an early signal of future brand strength. Crucially, the relationship is not instantaneous or uniform. On average, changes in creative quality precede changes in brand equity by around 10 to 11 months, but the analysis reveals how creative works on two clocks as two distinct patterns emerged:

1. Brands where creative and brand equity move together with little or no lag;
2. Brands where effects accumulate more gradually, sometimes over as long as two years.

This distinction matters for planning. Some brands can expect relatively rapid feedback from improvements in creative performance, while others need a longer-term view, greater consistency and more patience before the full equity effect becomes visible. Across both patterns, the starting quality of the creative and its direction of travel are critical: strong creative gives brands a better chance of growth, while weak creative that fails to improve risks stagnation or decline.

The second contribution of the paper uses Kantar data to move beyond measuring creative performance to mapping the structure of creative itself. By treating ads as combinations of interacting elements, rather than as a collection of individual devices, the analysis shows that cohesion is a powerful way to understand why some creative systems work harder than others. Product-focused combinations, for example, appear highly cohesive and strongly linked to both short-term and long-term outcomes, while other themes, such as sustainability, become more effective when supported by complementary elements. The implication is clear: effectiveness is rarely the result of a single creative cue. It comes from the way creative choices work together, in context, at a particular moment in time. The longitudinal view also shows that even effective creative structures cannot simply be repeated indefinitely. Cohesion helps brands build meaning, salience, preference and pricing strength, but when the same combinations become widely adopted or over-used, their impact can decay. Distinctiveness therefore remains an essential counterweight to consistency. The strongest creative systems are not static templates; they are managed systems that retain enough continuity to build memory structures, while evolving sufficiently to remain fresh, relevant and differentiated.

The agency-switch analysis reinforces this point by showing what happens when creative systems are disrupted. A change of agency produces an immediate fall in cohesion and a narrowing of the creative toolkit, yet it can also generate short-term gains in difference, consideration and willingness to pay more. At the same time, trust declines, underlining the strategic risk of creative disruption. Over two to three years many brands recover creative breadth, and a third exceed their previous cohesion levels, but most do not. Agency change can therefore create the conditions for creative renewal, but it does not guarantee it. The brands that benefit are those that convert disruption into better, more effective combinations of creative elements. Taken together, the findings argue for a more dynamic view of creative effectiveness. Creative should not be judged only as an executional output, nor managed on a campaign-by-campaign basis. It should be understood as a living brand asset: one that has structure, momentum, memory and consequence. For marketers, the practical challenge is to calibrate that system deliberately, using evidence to know when to stay consistent, when to refresh and when to take the risk of change. The world's strongest brands create some of the world's best advertising—not because creativity and brand strength are separate forces, but because each reinforces the other. When creative quality is sustained, structured and intelligently evolved, it becomes one of the most powerful ways to build and protect long-term brand equity.

## About the Authors

Vera Sidlova, Global Creative Thought Leadership Director, Kantar, Prague, Czechia.

Polly Wyn-Jones, Global Knowledge Director, Creative & Media, Kantar, Leamington Spa, UK.

Horacio Gonzalez, Head of Unstructured Data Analytics & AI, Kantar, London, UK.

Copyright © **Esomar** 2026

Congress 2026

Felipe Thomaz, Associate Professor of Marketing, Saïd Business School, University of Oxford, UK.

Neta Livneh, Postdoctoral Research Fellow, Saïd Business School, University of Oxford, UK.