Behavioral Economics Guide 2026

Introduction
Ulrike Malmendier

Guest Editorial
Isabelle Brocas

Editor
Alain Samson

The Behavioral Economics Guide 2026
ISSN 2398-2020

Author information:

Alain Samson (Editor)

Ulrike Malmendier (Introduction)

Isabelle Brocas (Guest Editorial)

Zeina Afif (World Bank), Sean R. Asbury (Voya Financial), Rasha Attar (Behavioral Science Group), Liza Bergman-Geonie (The Chicago School), Johan Birkelund (Norwegian Tax Administration), Chiara Cappellini (Behavioral Science Group), Francisco Ceballos (BeWay), Pablo Coello (BeWay), Juan de Rus (Neovantas), Gonzalo Díaz (BeWay), Antonio González (BeWay), David Halpern (Behavioral Insights Team), Dario Krpan (London School of Economics and Political Science), Michiel Kuhn (Panthera Institute), María López (Neovantas), Siting Estee Lu (University of Edinburgh), Coney H. Lu (Voya Financial), Deelan Maru (Behavioral Insights Team), Christelle T. Ngnoumen (Voya Financial), Claire Nolan (Dectech), Marisa Nowicki (Lirio), Juan Pablo Ospina Arboleda (BeWay), Alice Pearce (Dectech), Pakamon Prasitsumrit (Dectech), Sara Ramos-Cabo (BeWay), Paola Schietekat (Neovantas), Markus Schuller (Panthera Institute), Pablo Villacañas (BeWay), Patrick Wierckx (Panthera Institute), Angel Zhou (Dectech), and Ivana Zilic (DePaul University) (Contributing Authors)

Romneya Robertson (Graphic Design)

Published on www.behavioraleconomics.com by Behavioral Science Solutions Ltd

Suggested citation:

Samson, A. (Ed.). (2026). *The Behavioral Economics Guide 2026.* https://www.behavioraleconomics.com/be-guide/.

Suggested citation for individual chapters/articles:
[Author(s)] (2026). [Chapter/Article Title]. In A. Samson (Ed.), *The Behavioral Economics Guide 2026* (pp. nnn-nnn). https://www.behavioraleconomics.com/be-guide/.

In memory of

Patrick Samson

A friend of the *Behavioral Economics Guide*

Contributions By

BEHAVIORAL SCIENCE GROUP

Beway THE BEHAVIORAL WAY

TheChicagoSchool

Dectech

lirio

NEOVANTAS

PANTHERA INSTITUTE

Skatteetaten

THE UNIVERSITY of EDINBURGH

VOYA FINANCIAL

Further Support By

LSE THE LONDON SCHOOL OF ECONOMICS AND POLITICAL SCIENCE

UCD DUBLIN University College Dublin

University of Glasgow

Penn Master of Behavioral and Decision Sciences

UNIVERSITY of STIRLING

UNIVERSITY OF WARWICK

Table of Contents

Acknowledgements

The editor would like to thank Ulrike Malmendier for writing the introduction to this edition and Isabelle Brocas for her guest editorial.

We are grateful for the support received from the Behavioral Science Group, BeWay, Dectech, Lirio, Neovantas, the Panthera Institute, Voya Financial, as well as the London School of Economics and Political Science, University College Dublin, University of Glasgow, University of Pennsylvania, University of Stirling, and University of Warwick.

Many thanks also go to Ivona Babarelu, Colin Barrett, David Beaumont, Jodie Blackadder-Weinstein, Tamsin Bromley-Rahlke, Joseph Conrad, Benjamin Curtis, Jèf Davis, Will Delavan, Milos Graonic, Jeffrey Hemmeter, Mohamad Hjeij, Wing Hsieh, Leya Hubbard, Niree Kodaverdian, Kármen Kovács, Myles Leighton, Roger Miles, Sara Miran, Brandon October, Eugenia Peñaloza, Edgar Sanchez, Rolando Sanchez, Anisha Singh, Vasilios Skilaris, Bradley Sullivan, George Theo, and Danish Us-Salam for their valuable feedback.

INTRODUCTION

Homo Experiens: Why Behavioral Economics Needs the Life Sciences

ULRIKE MALMENDIER

University of California, Berkeley

Click or scan QR code
to watch video summary

In 1974, a man named Henry Wallich was appointed Governor of the US Federal Reserve System. Born as Heinrich Wallich in Berlin in 1914 into a family of bankers, he had lived through the German hyperinflation as a child before emigrating from Nazi Germany in the 1930s. After spending time in Argentina, he settled in the United States, where he went on to build a distinguished career as an economist: a PhD from Harvard, a professorship at Yale, and ultimately a seat at the most powerful monetary policy table in the world, the Federal Open Market Committee (FOMC). On the Fed Board, Wallich came to be the most hawkish central banker the US has ever seen. Being constantly worried about inflation—and about his colleagues not seeing the dangers the way he did—he would dissent over and over again whenever the Fed Chairman proposed to ease policy rates. During his twelve-year career at the Board beginning in 1974, he dissented 27 times against the proposed interest rate decisions, a record that still stands. Almost invariably, his dissents were hawkish: he wanted higher rates to guard against inflation. As Nancy Teeters—the first woman on the Federal Reserve board of governors, and Wallich's colleague on the Board—later recalled, "Henry Wallich was our real problem... For some people, like Henry, who lived through the hyperinflation in Germany, fighting inflation was the only thing" (Teeters, 2008).

Wallich did not lack information. He had the best macroeconomic data at his fingertips, was supported by staff who compiled forecasts and ran models for him and was in constant exchange with other experts. And yet, the experience of witnessing the destructive power of unchecked inflation as a child in Berlin—an experience that had occurred at a different time half a century earlier, on a different continent, under entirely different institutional conditions—shaped his professional judgment in ways that no amount of data and theoretical knowledge could override.

Wallich's case is particularly vivid, but it is not atypical. When my co-authors Stefan Nagel, Zhen Yan, and I analyzed the forecasts, speeches, and votes of all FOMC members since the establishment of Fed independence in 1951, we found the same pattern across the board: members who had personally

Figure 1: Children playing with stacks of hyperinflated currency during the Weimar Republic.

experienced higher inflation over their lifetimes produced more hawkish forecasts and cast more hawkish votes (Malmendier et al., 2021). These are among the most informed economic decision-makers in the world. If their beliefs and choices are durably shaped by what they have personally lived through, then we need to ask what standard economic models are missing.

Traditional neoclassical economics typically attributes choices that overly weigh personal experience to a lack of information—individuals only know, or at least know best, what they have personally seen. Clearly, this explanation does not apply to the FOMC. Information and knowledge are not the constraint. Behavioral approaches to modeling an overreliance on personal lifetime experiences would typically go towards cognitive limitations, for instance in terms of data processing, or biases in belief formation, such as the overweighting of more recent realizations. But these approaches also seem hard to apply here, since economists on the Federal Reserve Board would know state-of-the-art models of inflation processes and be supported by well-trained staff who would update them.

What I am proposing is different. We need to come to terms with humans being shaped by their life experiences in a deeper, more hardwired way than our existing models permit. We humans are living, breathing organisms who adapt to our environment and the circumstances of our lives, and those experiences affect our beliefs, our outlook on the world, and our actions.

"We need to come to terms with humans being shaped by their life experiences."

To be sure, behavioral economics to date has made tremendous progress in instilling psychological realism into models of human behavior. Not too long ago, the predominant neoclassical model portrayed people as *homines oeconomici*, i.e., rational utility maximizers with Bayesian expectations and perfect cognitive abilities. Behavioral economics has taken important steps forward in dismantling this fiction. The first wave of behavioral research in the 1980s and 1990s, building on the work of Amos Tversky and Daniel Kahneman, compiled an extensive catalog of systematic decision-making errors, from loss aversion to the availability heuristic to reference dependence. A second wave incorporated these insights into empirical studies across many fields, generating robust evidence of biased behavior among consumers, investors, managers, and policymakers. Together, these advances reshaped economics. They gave us better models, better predictions, and better interventions: from automatic enrollment in retirement plans to simplified disclosure forms.

Yet, the behavioral economics revolution has stalled at a decisive point. The human element I appealed to above is still missing, or at least incomplete. To put it in stark terms, human behavior still seems rather robotic. Where previously, in the neoclassical model, it was the output of a perfectly programmed computer, in the behavioral economics model it became the output of a not-quite-so-perfect computer program, one prone to systematic bugs. And while the discovery of heuristics and biases was a genuine improvement, the resulting models remained mechanical. Once we identify the biases and model them, the program is assumed to run the same way for everyone, at every point in their lives. What you have lived through, what you have felt, what has happened to your body and brain along the way—none of that enters the model. Henry Wallich and a fellow governor born in Ohio with no memory of hyperinflation would be modeled identically, conditional on the same information set.

Figure 2: Henry Wallich, Federal Reserve Governor (1974-1986).

But that is still not the behavior of a *homo*, a human being, a living and constantly changing organism. It is time for economics to take a further step and develop truly *human* models of economic decision-making. The purpose of this introduction is to argue that life experiences durably shape our brain, our body, and our (ir)rational decisions, and that economic models need to account for this. This approach enables an evolution away from the concept of the mechanistic decision-maker and toward a person whose mind and body are marked by their experiences and the traces they leave: toward a *homo experiens*. Behavioral economics needs to go further, not just drawing on social psychology but incorporating insights from the life sciences, including neuroscience, biology, medicine, and psychiatry. I sometimes call this *Behavioral Economics 3.0* – the third wave of behavioral economics. At its core is a concept my co-authors and I have developed over two decades of research: *experience effects*.

What Are Experience Effects?

The starting point for this new approach is a simple observation. Standard economic models—and even most behavioral models—treat personally lived experience and theoretically learned information as equivalent. If you know the historical data on stock-market returns, it should not matter whether you personally lived through a crash or merely read about one in a textbook. Your beliefs and decisions should be the same, conditional on the information you possess.

But that is not what we find. More than twenty years ago, Stefan Nagel and I started the research agenda on experience effects when we became interested in the American notion of "Depression Babies", the idea that people who grew up during the Great Depression remained profoundly risk-averse for the rest of their lives, shunning the stock market long after economic conditions had recovered. Using data from the Survey of Consumer Finances since 1960, and data on stock-market returns all the way back to 1885, we tested whether the average stock-market returns a person had experienced over their lifetime predicted their willingness to invest in stocks and their stated risk tolerance. The answer was a clear yes. Comparing someone at the 10th percentile of lifetime experienced stock-market returns to someone at the 90th percentile, we found a difference in stock market participation of about 14 percentage points, relative to an average participation rate of 37 percent (Malmendier & Nagel, 2011).

What made this more than a story about recency bias was the discovery that individuals put weights on their entire lifetime experiences: a 30-year-old on the last 30 years, and a 60-year-old on the last 60 years. Recent experiences do matter most. But experiences from decades ago can still leave detectable traces in your financial decisions today.

> "Recent experiences do matter most. But experiences from decades ago can still leave detectable traces in your financial decisions today."

What is more, we found very similar patterns for inflation experiences and inflation beliefs. People who have lived through periods of high inflation expect higher inflation in the future (Malmendier & Nagel, 2016). The insight from our research paper is well captured by a 1979 quote by Paul Volcker, then chairman of the Federal Reserve. When trying to combat inflation that was crossing the 10% mark, he lamented that it was hard to convince people that a return to price stability was possible precisely because their lived experience told them otherwise:

"*An entire generation of young adults has grown up since the mid-1960's knowing only inflation, indeed an inflation that has seemed to accelerate inexorably. In the circumstances, it is hardly surprising that many citizens have begun to wonder whether it is realistic to anticipate a return to general price stability, and have begun to change their behavior accordingly*" (Quoted in Orphanides and Williams, 2005).

In our analysis, which uses data from the Michigan Survey of Consumers since 1953, that is exactly what we find: cross-generational differences in inflation expectations that track lifetime inflation experiences with considerable precision. Importantly, the implicit weighting that people place on past experiences turns out to be very similar across these different domains. Whether we look at inflation or at stocks or at bonds, people tend to assign roughly linearly declining weights on their lifetime experiences, going back from today to the beginning of their lives. This suggests a general behavioral pattern rather than a

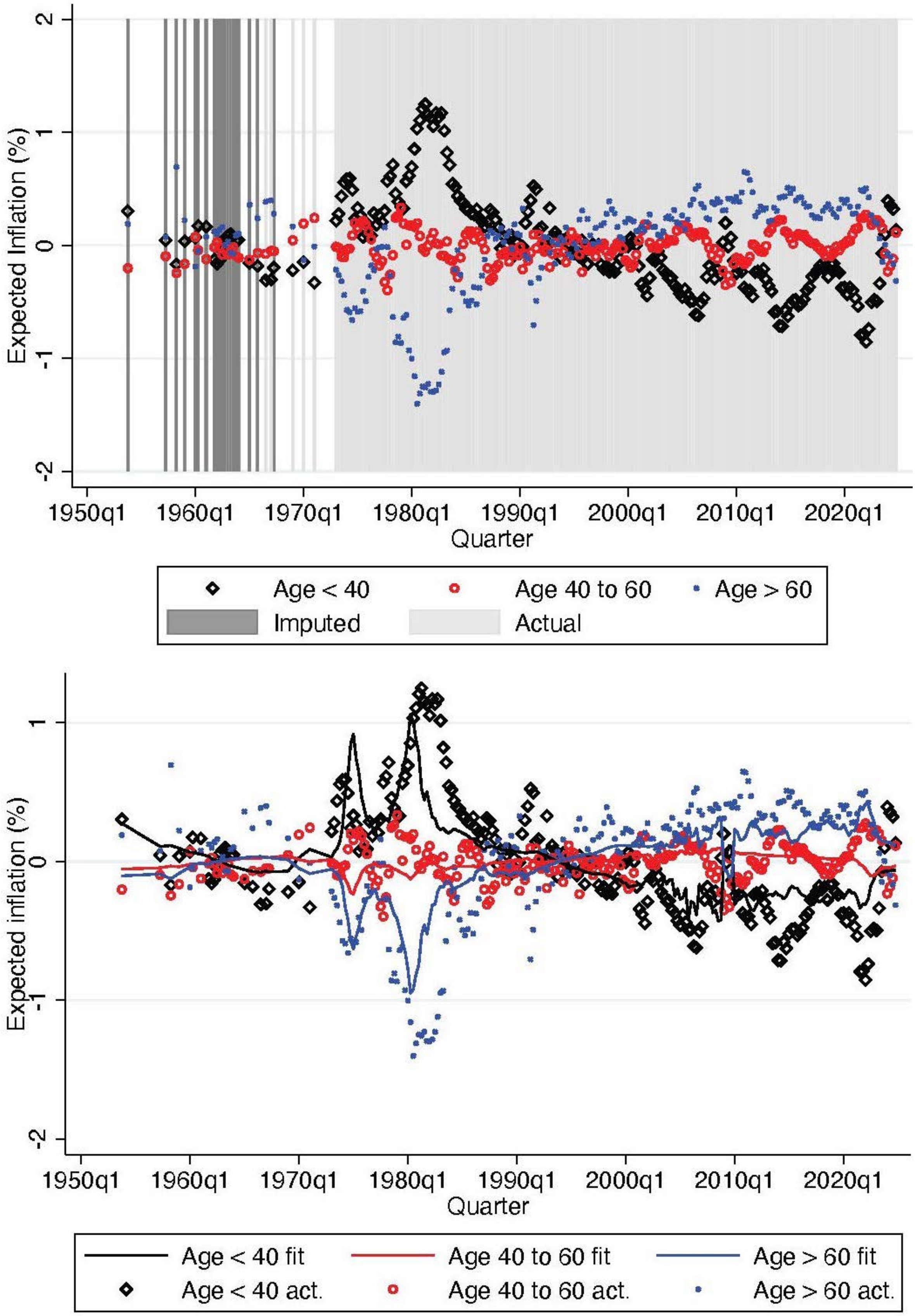


Figure 3: Inflation expectations by age group (Malmendier & Nagel, 2016), updated to the latest available data. *Note*: Both panels plot four-quarter moving averages of mean one-year-ahead inflation expectations for young (under 40), middle-aged (40–60), and older (over 60) individuals from the Reuters/Michigan Survey of Consumers, shown as deviations from the cross-sectional mean each quarter. The top panel shows the survey expectations: disagreement across age groups peaks during the high-inflation years of the late 1970s and early 1980s. The bottom panel overlays the learning-from-experience model's fitted values (lines) on the actual survey expectations (markers), showing that differences in lifetime inflation experiences account for much of the age-related disagreement.

domain-specific quirk.

In further research, we have also seen that these effects are not confined to survey responses. They translate into real, costly decisions. People who have experienced high inflation are more likely to buy rather than rent their homes, seeking inflation protection in real estate (Malmendier & Steiny Wellsjö, 2024). Furthermore, they are far more averse to adjustable-rate mortgages, preferring the safety of fixed rates even when those are substantially more expensive. My co-author Matthew Botsch and I calculated that the Baby Boomer generation alone overpaid roughly $22 billion for fixed-rate mortgages in the late 1980s and 1990s because of their inflation-scarred preferences (Botsch & Malmendier, 2026). In the realm of consumer spending, past unemployment experiences lead to heightened precautionary saving and reduced consumption, even years after the adverse episode has ended, controlling for current income, wealth, and employment status (Malmendier & Shen, 2024). And in work with Christine Laudenbach and Alexandra Niessen-Ruenzi, we have shown that growing up under communist, anti-capitalist ideology in East Germany shapes attitudes toward financial markets and stock market participation even decades after reunification—with the direction and strength of the effect depending critically on whether those experiences were emotionally tagged as positive or negative (Laudenbach et al., 2025).

> “The Baby Boomer generation alone overpaid roughly $22 billion for fixed-rate mortgages in the late 1980s and 1990s because of their inflation-scarred preferences.”

The breadth of evidence extends well beyond my own research program. Other teams have documented that personal experiences with flood disasters shape demand for insurance and affect house prices (Gallagher, 2014; Bakkensen & Barrage, 2022), that recession experiences in early adulthood reduce prosocial behavior later in life across 74 countries (Bietenbeck et al., 2025), that early macro adversity lowers institutional trust and alters voting behavior across 34 European countries (Gavresi & Litina, 2023), and that childhood financial and cultural experiences leave lasting imprints on adult creditworthiness and repayment behavior (Bakker et al., 2025). Whether we look at stock markets, inflation expectations, housing decisions, consumption, international capital flows, or political attitudes, the picture is consistent: personal experiences act as a central and durable input into expectation formation and decision-making, generating persistent heterogeneity that neither neoclassical nor standard behavioral economics models can account for.

Even Experts Are Not Immune

Perhaps the most consequential finding in this literature—and for traditional economics, the most uncomfortable—is that experience effects operate even among experts who have access to the best available data and sophisticated tools for processing it. I opened this essay with the case of Henry Wallich, but the pattern generalizes well beyond him and beyond monetary policy at the FOMC. Similar findings have emerged for fund managers, who are disproportionately influenced by personal experience with bubbles (Greenwood & Nagel, 2009), and for bankers, whose institutional risk-taking reflects the crises they have personally weathered (Bouwman & Malmendier, 2015). Physicians, too, repeatedly report changing their clinical approach after witnessing adverse outcomes in their own patients, even when they know the original protocol was correct ex ante. Expertise does not immunize against the imprint of lived experience.

> “Expertise does not immunize against the imprint of lived experience.”

This is the finding that pushes us not only beyond traditional neoclassical economics, but also beyond prior approaches in behavioral economics. If the best-informed decision-makers in the world—people with access to the best available data, supported by sophisticated models and dedicated research staff—are still durably shaped by what they have personally lived through, then we are not dealing with cognitive limitations or information-processing errors that better education could fix. We are dealing with something deeper, something that has to do with how living through events physically reshapes the way our brains process information. Experience-based belief formation is, in the language of neuroscience,

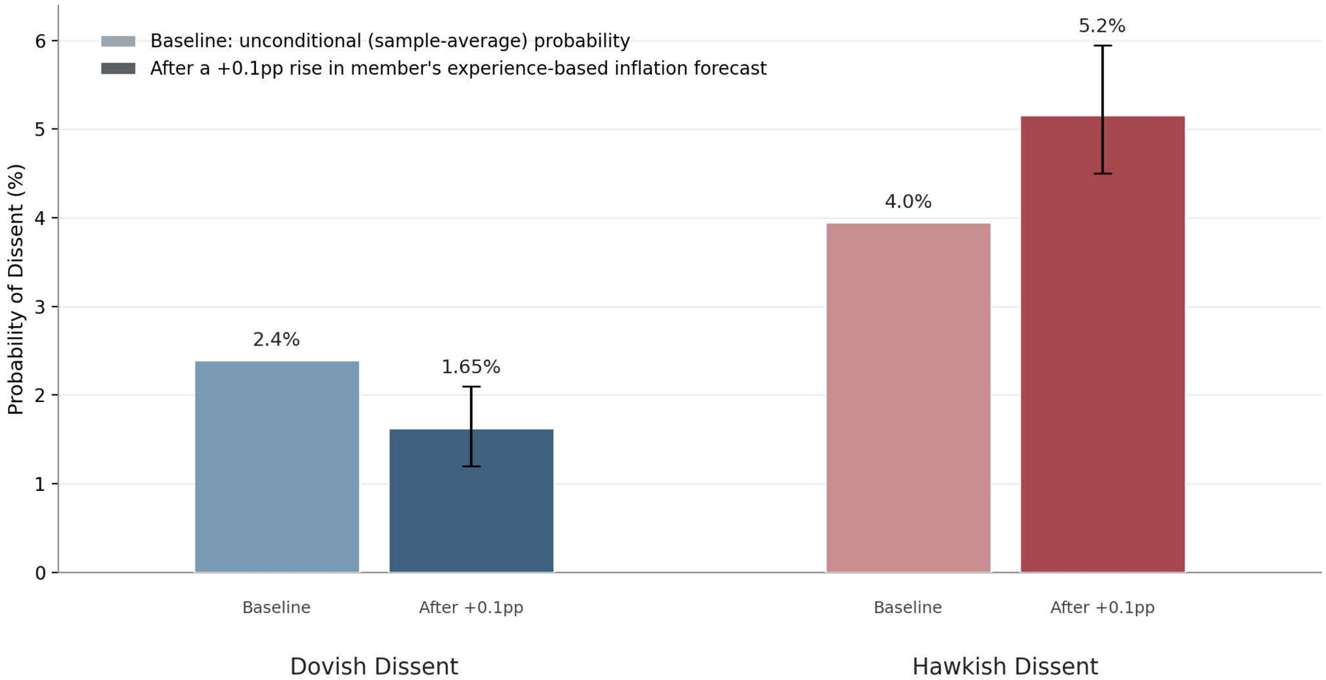


Figure 4: Inflation experiences and FOMC dissent probabilities. *Note*: Each pair of bars shows how the probability of a dissenting vote shifts when an FOMC member's experience-based inflation forecast rises by 0.1 percentage points, approximately the typical within-meeting standard deviation in these forecasts across members at a single meeting (Malmendier et al., 2021). Within each category, the left (lighter) bar is the unconditional, sample-average probability of that dissent type (2.4% dovish, 4.0% hawkish; from 160 dovish and 265 hawkish dissents out of 6,707 votes). The right (darker) bar is the predicted probability after the 0.1pp increase, derived from the authors' ordered probit estimates of FOMC voting behavior: a hawkish dissent becomes about 1.2pp more likely (from 4.0% to 5.2%), while a dovish dissent becomes about 0.8pp less likely (from 2.4% to 1.65%). The pattern indicates that members who have experienced higher inflation over their lifetimes lean hawkish: more inclined to dissent toward tightening and less inclined to dissent toward easing.

a rewiring issue, not a firing issue: beliefs reflect the brain's physical, structural connectivity having adapted, not a failure of neurons to activate or transmit electrical signals. Understanding this notion requires looking beyond psychology to the life sciences.

The Biology Underneath

The life sciences are essential if we want to better understand how our brains and bodies affect our beliefs and choices.

> "Neuroscience in particular tells us that the brain is not a static processor; it is an organ that physically reorganizes itself in response to experience."

Neuroscience in particular tells us that the brain is not a static processor; it is an organ that physically reorganizes itself in response to experience, a property known as *neuroplasticity*. Every significant experience triggers the formation or strengthening of synaptic connections. The brain and the human body are not merely a computing apparatus that processes data, but a dynamic system that changes structurally through what we live through.

Three features of this process are especially relevant for economists.

First, repetition and duration matter. Repeated or prolonged exposure to certain conditions—years of high inflation, a drawn-out recession—strengthens neural pathways through a mechanism known as "long-term potentiation" (LTP) (Douglas & Goddard 1975; cf. Frey & Morris, 1997, 1998). The name is not metaphorical. LTP is a measurable biological process in which synaptic connections become physically stronger with repeated activation. Applied to economics, this means that the longer an inflation episode lasts, or the more frequently one lives through economic distress, the deeper the traces this experience leaves.

Second, emotions matter. Experiences accompanied by strong emotions, such as fear during a market crash or anxiety about money during a period of job loss, receive what neuroscientists call 'emotional tagging'. Such emotionally marked memories are encoded more deeply, stored more durably, and retrieved more readily (Dolan, 2002; LaBar & Cabeza, 2006). Emotional tagging helps explain why lived experiences have such disproportionate power compared to information learned from a textbook: the textbook does not come with the racing heart and the sleepless nights.

Third, and crucially for policy, learned knowledge has limited power to undo these effects. The literature on trauma is instructive here: it describes how traumatic stress can cause synaptic changes that shape fear and stress networks in a lasting manner (Mahan & Ressler, 2012). The point of referencing the trauma literature is not to classify economic crisis experiences wholesale as trauma, but to recognize that intense or emotionally marked experiences leave biological traces that cannot be neutralized simply by providing additional information, however accurate that information may be. You cannot lecture someone out of a synaptic trace.

This third point is mirrored in evidence on other physical traces of stress. For example, in work with Mark Borgschulte, Marius Guenzel, and Canyao Liu, I have provided direct evidence that crisis experiences leave biological marks in terms of accelerated aging. We studied what happens to CEOs who lead firms through severe industry distress. Using visual machine-learning analysis of photographs, we applied a technique that estimates the so-called 'apparent age' of people, i.e., the age a human would guess by looking at you. In medicine, physicians often assess patients' health status by comparing their physical appearance to their actual chronological age, and it has been shown that a significant difference between apparent age and chronological age can act as a marker for higher mortality risk. In our sample of US CEOs, we found that those exposed to industry distress shocks during their tenure as CEOs aged visibly more than their unexposed peers. The effect was immediate and permanent: they looked about one year older, and they never recovered. We also found that this acceleration of visual aging mapped one-to-one onto actual mortality, translating into approximately 1.1 years of reduced life expectancy. In another study, I show that the effects are even more severe for workers. Using detailed Swedish administrative data, Paolo Sodini and I tracked individuals laid off during Sweden's financial crisis of the early 1990s for over two decades. These workers showed significantly elevated rates of alcohol abuse, drug abuse, mental health diagnoses, hypertension, diabetes, and heart disease, translating into meaningfully higher mortality. And remember—this is in Sweden, a country with a generous social safety net.

The implication is that the biological channels through which experiences affect us—stress hormones, neural rewiring, immune and endocrine responses—are real, measurable, and consequential for the economic outcomes we care about. And they are almost entirely absent from economic models, whether behavioral or not.

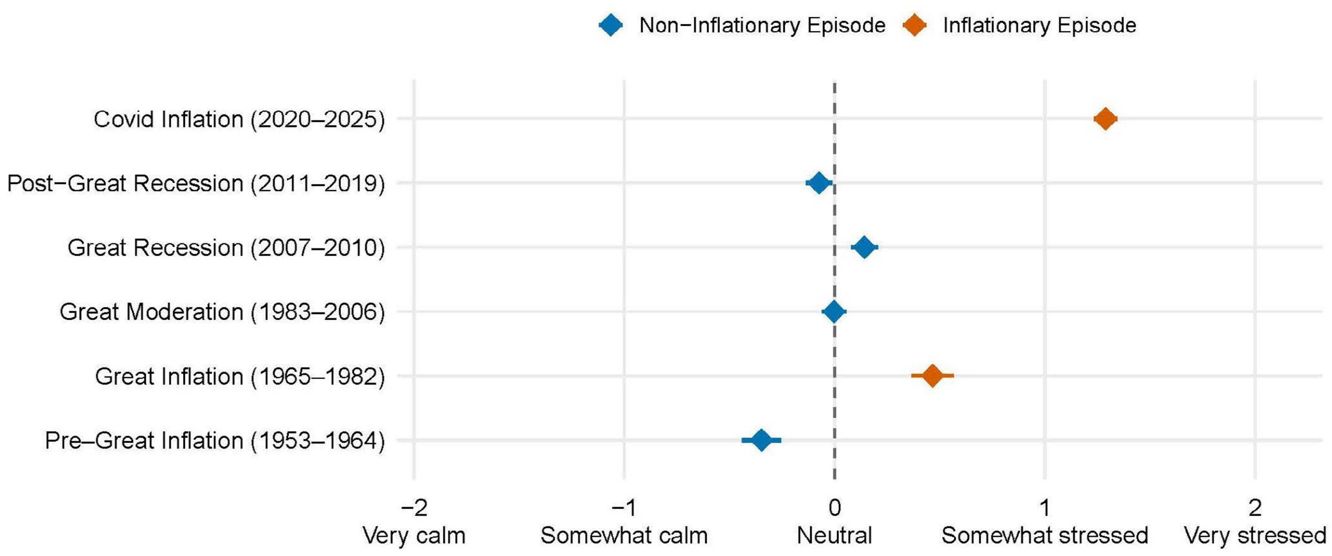


Figure 5: Average stress reported during inflationary and non-inflationary episodes.

New Evidence: Stress, Memory, and Narratives

To advance the agenda of incorporating insights from the life sciences into economics, I recently started to explore how stress and emotions during these past inflation episodes relate to what people remember, and what causal narratives they construct. In Malmendier (2026), I conducted a survey of 1,105 Americans aged 20 to 80 on the Prolific platform about those past episodes.

The results are revealing. When asked about periods of high inflation, respondents overwhelmingly associated these episodes with negative emotions and physical stress symptoms—difficulty sleeping, stomach problems, body tension, lack of energy. This is true for both the Great Inflation of the 1970s and 1980s and the recent Covid-era inflation. When I asked open-ended questions about how these episodes felt, the word that appeared most often was 'stressed', followed closely by 'worried' and 'anxious'.

Crucially, the intensity of stress matters above and beyond the financial severity of the shock. I constructed an individual-level measure of 'stress elasticity', capturing how much stress a person reports per unit of experienced inflation, controlling for financial strain. This stress elasticity is a strong, independent predictor of inflation expectations. In other words, holding the objective macroeconomic shock constant, individual variation in how stressful the experience felt carries additional predictive power for beliefs about the future.

Stress also shapes memory. Whether someone recalls a particular period as one of high inflation is tightly linked to the stress and physical symptoms they experienced during that period, not just to how much it strained their finances. This is consistent with the emotional tagging hypothesis from neuroscience: what gets encoded in memory is not a recording of objective events but is filtered through the emotions that accompanied them. Perceived controllability also matters: respondents who reported feeling in control over their finances, or feeling that the government was in control, during an inflation episode were less likely to recall that period as one of high inflation, suggesting the importance of controllability in memory consolidation.

Finally, stress shapes narratives, i.e., the causal stories people construct to explain what happens during periods of macroeconomic shocks. In the case of inflation, this means that the explanations people develop for a specific episode (for example, "It was caused by the energy crisis" for the Great Inflation, or "It was supply chain disruptions" for the Covid inflation) strongly predict their *general* narrative about what causes inflation. And this link is amplified for people with high stress elasticity. In other words, the more emotionally affected people are by a particular episode, the more likely they are to generalize the story of that episode into their broader economic worldview.

> "What gets encoded in memory is not a recording of objective events but is filtered through the emotions that accompanied them."

What This Means for Policy

If this perspective is correct, it has important implications for the design of economic policies. Different economic approaches often lead to widely diverging recommendations.

In the traditional economic view, mistakes and biased beliefs are attributed primarily to lack of information or to model uncertainty. The remedy, then, lies in improving access to information or reducing uncertainty about the data-generating process. For example, suboptimal investment and savings decisions might be attributed to a lack of information about their availability and accessibility.

From the perspective of behavioral economics, there are further sources of suboptimal decision-making, such as information-processing errors and imperfect cognition. Under this approach, the remedy shifts toward enabling humans to overcome the limitations of their belief formation process. An example would be, in the context of investment and savings decisions, the provision of targeted educational measures such as financial literacy programs that teach people about compound interest, Bayes' rule, or the benefits of diversification.

The novel perspective offered here explains why neither remedy suffices in eliminating investment mistakes—neither the provision of more information and data about available financial products nor financial literacy training. From the perspective of Behavioral Economics 3.0, we understand people as dynamically evolving organisms who are shaped by their life histories. We have seen that experience effects operate through biological channels such as neural rewiring, emotional tagging, and stress responses. As a result, information processing and belief formation are themselves structurally altered by past experiences, and they cannot simply be corrected through additional knowledge. Thus, the approach to constructing a remedy shifts fundamentally: from the provision of pure information or educational measures to the targeted design of experiences. In the context of investment and savings, the idea would be that people learn to accumulate wealth in the stock market by receiving controlled exposure to broadly diversified stock-market investment.

> “The approach to constructing a remedy shifts fundamentally: from the provision of pure information or educational measures to the targeted design of experiences.”

A concrete example of this shift can be found in a policy I helped develop as a member of the German Council of Economic Experts: the *Early Start Pension*. Germany is a country with notoriously low stock-market participation. However, it is hard to attribute that pattern to insufficient awareness of the stock market, whose performance is mentioned daily in the news. Nor can we easily attribute this outlier behavior to a lack of financial literacy as Germans score at the top of international financial literacy rankings (OECD, 2023). The problem is evidently not knowledge. Even when information about risks and returns is present, capital market participation remains low. What is missing is personal experience with broadly diversified capital market investment. In many countries, such as the US, Sweden, or Australia, people are automatically exposed to stock-market investments due to their capital market-based pension (or retirement savings) system. To date, Germany has failed to successfully implement such a reform.

Starting in 2027, the *Early Start Pension* will give children aged 6 to 18 a monthly contribution of 10 euros that will be automatically invested in a broad stock market fund. The amount is deliberately small: too small for losses to be existentially threatening, but large enough to generate a real, felt experience of how markets work over time. The goal is not to teach children about the stock market through instruction—as much as an accompanying financial training would be welcome—but to let them live through it, to build a personal experience history that, over years, rewires their perception of stock-market risk and return. Negative market experiences are processed in a ‘dosed’ manner at a non-existential level, and they are minimized through the focus on long-term, broadly diversified investment. Because the program is universal, it also mitigates social inequality in experiential exposure, ensuring that capital market experience does not remain a privilege of high-income households.

The Road Ahead

I believe that the next chapter of behavioral economics lies in a turn towards the life sciences—in taking seriously what neuroscience, psychiatry, medicine, and biology already know about human beings and how they are shaped by what they live through. This means, first, that economists should start collecting data on stress, emotions, hormones, and physical health alongside the economic variables we traditionally measure. So far, economists have rarely gathered data along these dimensions, and the standard sources of survey data—the ACS, the Michigan Survey of Consumers, the Survey of Consumer Finances—contain few variables that are related. Some individual researchers have begun to make progress, but there is much more to do.

> “Economists should start collecting data on stress, emotions, hormones, and physical health alongside the economic variables we traditionally measure.”

This new approach also requires us to build models that incorporate variables from stress and trauma research, variables that economists have barely touched but that powerfully predict how damaging a given shock will be and how long-lasting its effect. Two examples are measures of *predictability* and *controllability*. *Predictability*—whether an adverse experience follows a pattern or arrives without warning—turns out to matter substantially in explaining the strength of the effect of traumatic experiences. Research in neuroscience has shown that unpredictable stressors elicit stronger physiological and neural stress responses than predictable stressors of equivalent intensity (Tsuda et al., 1983; Alvarez et al., 2011; Cohodes et al., 2021). Applied to economics, this suggests that the unpredictability of an adverse condition, such as not knowing whether a food assistance program will be available from one month to the next, may compound its harmful effects beyond what the level of material deprivation alone would predict. *Controllability*—whether a person perceives themselves as having any agency over their situation—is equally important. Uncontrollable shocks produce learned helplessness and withdrawal. In my survey data, respondents who reported feeling in control of their finances, or feeling that the government was in control, during an inflation episode were significantly less likely to be scarred by it. These concepts could illuminate phenomena like the 'deaths of despair' documented by Case and Deaton, where the critical question may not only be how severe the financial shock was, but for whom it was predictable and for whom it felt controllable.

Finally, the perspective proposed in this article means that it is best to design policies to work with our biology rather than against it. It does not suffice to give people information or teach them how to process it. Better policy measures would shape the experiences through which that information is encoded.

We have made remarkable progress in moving from the *homo economicus* to a psychologically richer model of human decision-making. But we are still, in important respects, modeling humans as sophisticated machines. The next step is to model them as what they are: living organisms, marked by their histories, changed by what they have endured and enjoyed. The promise of Behavioral Economics 3.0 is that, by understanding the *homo experiens*, we can build not only better theories, but also more effective policies—ones that are grounded not just in what people know but in what they have lived through and what it has done to them.

THE AUTHOR

Ulrike Malmendier is the Edward J. and Mollie Arnold Professor of Finance at the Haas School of Business and Professor of Economics at the University of California, Berkeley, where she also serves as founding director of the O'Donnell Center for Behavioral Economics. She is a research associate at the National Bureau of Economic Research (NBER), and a Fellow of the Econometric Society and the American Academy of Arts and Sciences. In 2013, she received the Fischer Black Prize from the American Finance Association, awarded biennially to the top finance scholar under 40. Her research spans behavioral economics and finance, corporate finance, and law and economics, with a particular focus on how life experiences shape beliefs and decision-making.

ACKNOWLEDGMENTS

This introduction borrows from the Presidential Address I delivered at the American Finance Association Annual Meeting in 2026. I would like to thank my numerous co-authors, in particular Stefan Nagel (University of Chicago), Demian Pouzo (UC Berkeley), Vico Vanasco (CREi), Zach Yan (Cornerstone), Francesco D'Acunto (Georgetown), Michael Weber (University of Chicago), Juan Ospina (Banco de la Republica, Colombia), Marius Guenzel (UPenn), Canyao Liu (Hudson River Trading), Mark Borgschulte (UIUC), Leslie Shen (Boston Federal Reserve), Alex Steiny (UCSD), Matthew Botsch (Bowdoin), Alexandra Niessen-Ruenzi (University of Mannheim), and Christine Laudenbach (University of Frankfurt), as well as members of the scientific staff of the German Council of Economic Experts and the O'Donnell Center, in particular Nikki Azerang, Shagnik Chakravarty, Bihai Gao, Karin Li, and Pietro Risso.

REFERENCES

Alvarez, R. P., Chen, G., Bodurka, J., Kaplan, R., & Grillon, C. (2011). Phasic and sustained fear in humans elicits distinct patterns of brain activity. *Neuroimage*, *55*(1), 389-400. https://doi.org/10.1016/j.neuroimage.2010.11.057

Bakkensen, L. A., & Barrage, L. (2022). Going underwater? Flood risk belief heterogeneity and coastal home price dynamics. *The Review of Financial Studies*, *35*(8), 3666-3709. https://doi.org/10.1093/rfs/hhab122

Bakker, T. J., DeLuca, S., English, E. A., Fogel, J. S., Hendren, N., & Herbst, D. (2025). Credit access in the United States. *NBER Working Paper* 34053. https://doi.org/10.3386/w34053

Bietenbeck, J., Sunde, U., & Thiemann, P. (2025). Recession experiences during early adulthood shape prosocial attitudes later in life. *Journal of Public Economics, 243,* 105327. https://doi.org/10.1016/j.jpubeco.2025.105327

Borgschulte, M., Guenzel, M., Liu, C., & Malmendier, U. (2023). CEO stress, aging, and death. *Journal of Finance*, forthcoming.

Botsch, M. J., & Malmendier, U. (2026). The long shadows of the Great Inflation: Evidence from residential mortgages. *SSRN.* https://dx.doi.org/10.2139/ssrn.3888762

Bouwman, C., & Malmendier, U. (2015). Does a bank's history affect its risk-taking? *American Economic Review: Papers & Proceedings, 105*(5), 321-325. https://doi.org/10.1257/aer.p20151093

Cohodes, E. M., Kitt, E. R., Baskin-Sommers, A., & Gee, D. G. (2021). Influences of early-life stress on frontolimbic circuitry: Harnessing a dimensional approach to elucidate the effects of heterogeneity in stress exposure. *Developmental Psychobiology*, *63*(2), 153-172. https://doi.org/10.1002/dev.21969

Dolan, R. J. (2002). Emotion, cognition, and behavior. *Science, 298*(5596), 1191-1194. https://doi.org/10.1126/science.1076358

Douglas, R. M., & Goddard, G. V. (1975). Long-term potentiation of the perforant path-granule cell synapse in the rat hippocampus. Brain Research, 86(2), 205-215. https://doi.org/10.1016/0006-8993(75)90697-6

Frey, U., & Morris, R. G. (1997). Synaptic tagging and long-term potentiation. *Nature, 385*(6616), 533-536. https://doi.org/10.1038/385533a0

Frey, U., & Morris, R. G. (1998). Synaptic tagging: Implications for late maintenance of hippocampal long-term potentiation. *Trends in Neurosciences, 21*(5), 181-188. https://doi.org/10.1016/S0166-2236(97)01189-2

Gallagher, J. (2014). Learning about an infrequent event: Evidence from flood insurance take-up in the United States. *American Economic Journal: Applied Economics, 6*(3), 206-233. https://doi.org/10.1257/app.6.3.206

Gavresi, D., & Litina, A. (2023). Past exposure to macroeconomic shocks and populist attitudes in Europe. *Journal of Comparative Economics, 51*(3), 989-1010. https://doi.org/10.1016/j.jce.2023.04.002

Greenwood, R., & Nagel, S. (2009). Inexperienced investors and bubbles. *Journal of Financial Economics, 93,* 239-258. https://doi.org/10.1016/j.jfineco.2008.08.004

LaBar, K. S., & Cabeza, R. (2006). Cognitive neuroscience of emotional memory. *Nature Reviews Neuroscience, 7*(1), 54-64. https://doi.org/10.1038/nrn1825

Laudenbach, C., Malmendier, U., & Niessen-Ruenzi, A. (2025). The long-lasting effects of experiencing communism on attitudes toward financial markets. *The Journal of Finance*, forthcoming.

Mahan, A. L., & Ressler, K. J. (2012). Fear conditioning, synaptic plasticity and the amygdala: Implications for posttraumatic stress disorder. *Trends in Neurosciences, 35*(1), 24-35. https://doi.org/10.1016/j.tins.2011.06.007

Malmendier, U., & Nagel, S. (2011). Depression babies: Do macroeconomic experiences affect risk taking? *The Quarterly Journal of Economics, 126*(1), 373-416. https://doi.org/10.1093/qje/qjq004

Malmendier, U., & Nagel, S. (2016). Learning from inflation experiences. *The Quarterly Journal of Economics, 131*(1), 53-87. https://doi.org/10.1093/qje/qjv037

Malmendier, U., Nagel, S., & Yan, Z. (2021). The making of hawks and doves. *Journal of Monetary Economics*, *117*, 19-42. https://doi.org/10.1016/j.jmoneco.2020.04.002

Malmendier, U., & Shen, L. S. (2024). Scarred consumption. *American Economic Journal: Macroeconomics, 16*(1), 322-355. https://doi.org/10.1257/mac.20210387

Malmendier, U., & Sodini, P. (2026). *The long-term effects of job losses.* Technical report, Working Paper.

Malmendier, U., & Steiny Wellsjö, A. (2024). Rent or buy? Inflation experiences and homeownership within and across countries. *The Journal of Finance, 79*(3), 1977-2023. https://doi.org/10.1111/jofi.13332

OECD. (2023). *OECD/INFE 2023 International survey of adult financial literacy.* OECD Business and Finance Policy Paper. https://www.oecd.org/content/dam/oecd/en/publications/reports/2023/12/oecd-infe-2023-

international-survey-of-adult-financial-literacy_8ce94e2c/56003a32-en.pdf.

Orphanides, A. & Williams, J. C. (2005). Imperfect knowledge, inflation expectations and monetary policy. In B. S. Bernanke & M. Woodford (Eds.), *The inflation targeting debate* (pp. 201-246). University of Chicago Press.

Teeters, N. H. (2008, October 25). *Federal Reserve Board Oral History Project: Interview with Nancy H. Teeters.* https://www.federalreserve.gov/aboutthefed/files/nancy-h-teeters-interview-20081025.pdf

Tsuda, A., Ida, Y., Satoh, H., Tsujimaru, S., & Tanaka, M. (1989). Stressor predictability and rat brain noradrenaline metabolism. *Pharmacology Biochemistry and Behavior, 32*(2), 569-572. https://doi.org/10.1016/0091-3057(89)90198-6

EDITORIAL

Beyond Preferences: The Biological Foundations of Decision-Making

ISABELLE BROCAS[1]

University of Southern California

Click or scan QR code
to watch video summary

Economics has made important progress by incorporating psychology and cognitive science, yet it still often treats decision-making as if it were organized into distinct behavioral domains such as risk, time, memory, self-control, and social choice. This editorial argues that such fragmentation is becoming intellectually and practically limiting. Evidence from neuroscience, developmental science, genetics, affective science, and formal economic theory increasingly suggests that behavior does not arise from isolated preference modules but from interacting biological systems involved in perception, valuation, memory, attention, affect, regulation, and action selection. A biologically informed economics does not require economists to abandon preferences or reduce behavior to neurons and genes. Preferences remain useful reduced-form representations when behavior is stable and prediction is the objective. The point is different: when economists seek to explain heterogeneity, diagnose why interventions fail, or design policies that work across contexts and populations, it is often necessary to model the mechanisms that generate the outward pattern of choice. Biology can enter economics not only empirically, but also theoretically through models that derive observed behavior from underlying brain processes and resource constraints. This perspective clarifies what is malleable, when it is malleable, for whom, and through which policy levers. Taking biology seriously is therefore not a detour from economic theory—it is a step toward a more explanatory and more useful science of human behavior.

Introduction

Economics has long relied on a productive abstraction. It explains behavior through preferences, beliefs, incentives, and constraints, and it infers these latent objects from observed choices. This strategy has been extraordinarily powerful. It has allowed economists to discipline explanations, build tractable models, and make predictions in settings ranging from markets to families, schools, firms, and public policy. But the same abstraction has also encouraged a way of thinking about decision-making that is increasingly too neat for the phenomena we are trying to explain. Behavior is typically partitioned into domains: risk, intertemporal choice, self-control, social preferences, strategic sophistication, memory, attention, and belief formation. These distinctions have analytical value. They should not, however, be mistaken for the natural architecture of the mind.

> "The challenge is no longer simply to document that people depart from narrow rational benchmarks."

1 brocas@usc.edu

The challenge is no longer simply to document that people depart from narrow rational benchmarks, nor merely to add more behavioral regularities to the list. The challenge is to understand how behavior is generated. Once we begin asking that question seriously, the old divisions begin to look increasingly arbitrary.

The brain does not contain a risk module, a patience module, a memory module, and a self-control module each of which operates independently and then hands their outputs to the economist. Economic behavior arises from interacting biological processes, including perception, attention, memory retrieval, valuation, affective appraisal, inhibitory control, and action selection. These processes are deeply interconnected and are recruited in different combinations across tasks and contexts.

> "Economic behavior arises from interacting biological processes, including perception, attention, memory retrieval, valuation, affective appraisal, inhibitory control, and action selection."

Economics has already been transformed by its dialogue with psychology, especially through work showing that judgment under uncertainty often relies on heuristics rather than fully Bayesian reasoning (Tversky & Kahneman, 1974). Neuroeconomics then made clear that neuroscience could enrich economic theory by illuminating the mechanisms underlying valuation, choice, learning, and self-regulation (Camerer et al., 2005). In the framework proposed by Rangel, Camerer, and Montague (2008), value-based decision-making is deconstructed into core processes: representing the decision problem, valuing actions, selecting among them, evaluating outcomes, and learning. Fehr and Rangel (2011) similarly argue that neuroeconomics can provide more detailed computational and neurobiological accounts of choice than traditional reduced-form behavioral descriptions alone.

That view is not purely speculative, and it does not belong to neuroscience alone; it can be formalized economically. A line of theoretical work has shown how observed behavior can be derived from underlying brain processes and internal resource constraints rather than simply described at the level of choice. Models of perception-to-action mapping, endogenous memory retrieval, brain resource allocation, and self-control as constrained optimization all point in the same direction: familiar economic behavior can often be reinterpreted as the output of interacting biological systems facing scarcity, trade-offs, and internal constraints (Brocas & Carrillo, 2012, 2016, 2021; Brocas, Alonso, & Carrillo, 2014). This is not a side branch of economics. It is part of the effort to make economics genuinely explanatory.

Nevertheless, this argument should not be misunderstood as an attack on preferences. Preferences are indispensable modeling tools when they summarize stable patterns of choice and when the goal is prediction within a known environment. The problem arises when they are treated as final explanations rather than as compact descriptions of behavior generated by deeper processes. A person may appear risk-loving, impatient, distrustful, or weak-willed. These labels may be descriptively accurate, but they do not tell us whether the observed behavior reflects reward sensitivity, future simulation, memory retrieval, emotional threat, attentional capture, regulatory constraints, or unstable environmental expectations. The distinction matters because interventions that produce the same reduced form change in choice may operate through very different mechanisms, and interventions aimed at the wrong mechanism may fail.

> "Preferences are indispensable modeling tools when they summarize stable patterns of choice and when the goal is prediction within a known environment."

A biologically informed economics should also not be confused with biological determinism. Genes are not destiny. Neural mechanisms do not eliminate agency. Economic theory is not rendered obsolete by the existence of brains. But if our goal is to model human behavior rather than merely fit it, biology can no longer remain in the background; it helps explain why different kinds of decisions often cohere within individuals, why people differ in structured ways, why the same person behaves differently across contexts, and why environments shape behavior so deeply over the life course.

This matters especially for policy. Economists already estimate reduced-form relationships between behavior and incentives, constraints, or information. The value of a biological perspective is not that it replaces those estimates, it is that it helps interpret them. It asks which internal process is binding; whether the constraint is cognitive, emotional, developmental, informational, social, or environmental; whether the relevant system is malleable; whether timing matters; whether a policy should change incentives, reduce cognitive load, stabilize expectations, lower stress, improve sleep or nutrition, or redesign environments so that fragile regulatory systems are less heavily taxed. In this sense, biology changes not only what we measure, but also what we treat as a policy lever.

The broader claim of this editorial is therefore straightforward but consequential: economics should move beyond treating decision-making as a set of isolated preference domains and toward a more integrated account grounded in biology, development, affect, and internal cognitive constraints. This would not replace economic theory—it would deepen it. Figure 1 summarizes the central framework advanced in this editorial. Economic behavior should not be treated as the output of isolated preference domains but as the product of interacting biological systems that transform perception, memory, valuation, affect, and regulation into action.

The Problem With Treating Economic Domains as Separate

One of the discipline's great strengths has been its ability to isolate dimensions of behavior. Risk preferences, intertemporal preferences, ambiguity attitudes, social preferences, strategic sophistication, and self-control have all generated productive literatures. Yet, this partitioning has also come at a cost. It has encouraged the idea that these domains correspond to separable psychological objects, when many may reflect overlapping underlying mechanisms.

Take risk and time. Economists often treat them as distinct dimensions of preference: one concerns uncertainty, the other delay. But that separation becomes less convincing once we think mechanistically. Both types of decisions recruit valuation, affective anticipation, future representation, attention to salient outcomes, learning from feedback, and inhibitory control. A person who is impulsive may also appear more willing to take risks, not because impatience and risk-loving are identical constructs but because both may be shaped by common patterns of reward sensitivity, limited future simulation, stress, or weaker self-regulation. Similar overlaps likely exist between strategic reasoning and working memory, between cooperation and emotional regulation, between trust and threat perception, and between consumer choice and attentional bottlenecks.

The traditional language of preferences can therefore obscure as much as it reveals. It describes regularities at the level of outcomes, but it does not tell us whether those regularities arise from common mechanisms or genuinely distinct systems. Neuroeconomic work has been especially valuable here because it shows that value-based choice relies on a sequence of interacting computations rather than on isolated behavioral modules. The act of choosing involves representing the problem, assigning subjective values, selecting an action, experiencing an outcome, and updating behavior through learning (Rangel et al., 2008). This logic already pushes against the idea that risk, time, social choice, and self-control should be understood as self-contained domains.

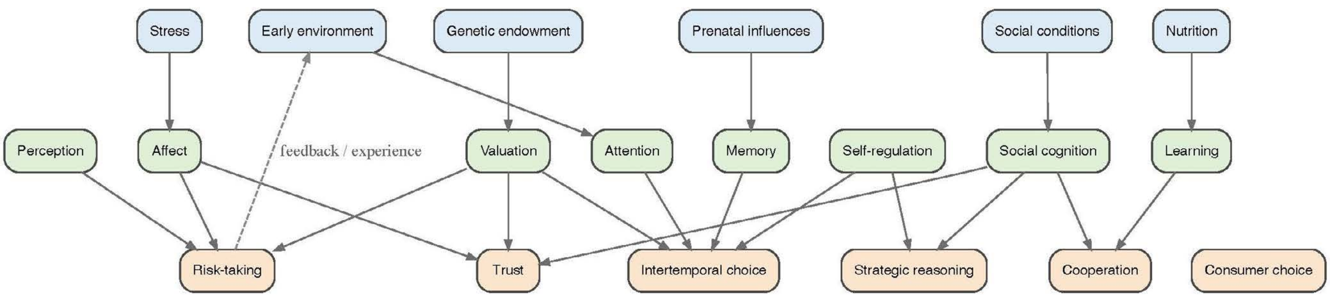


Figure 1: A biologically informed framework of decision-making. Economic behavior emerges from the interaction of perception, attention, memory, valuation, affect, regulation, and action selection (green). These processes are shaped by biological endowments and environmental conditions across development (blue).

The implication is not that economists should stop estimating risk aversion, discount rates, or social preferences; rather, these constructs should be treated as behavioral surfaces. They are useful summaries of what agents do in a particular environment. They are less useful as explanations when behavior changes across contexts, when the same intervention has different effects across individuals, or when welfare analysis requires understanding whether a choice reflects stable valuation, misinformation, compulsion, fear, overload, or adaptation to instability. The more the economist cares about mechanism, transportability, and intervention design, the more important it becomes to look beneath the preference label.

“The more the economist cares about mechanism, transportability, and intervention design, the more important it becomes to look beneath the preference label.”

Figure 2 illustrates this point. The same underlying processes can contribute to behavior across several economic domains, which is why apparent correlations across risk-taking, impatience, impulsivity, trust, or social sensitivity may reflect shared mechanisms rather than separate preference parameters.

The same critique applies to the relation between cognition and affect. Economists have become increasingly comfortable with the idea that cognition is limited and attention is scarce. This has been a

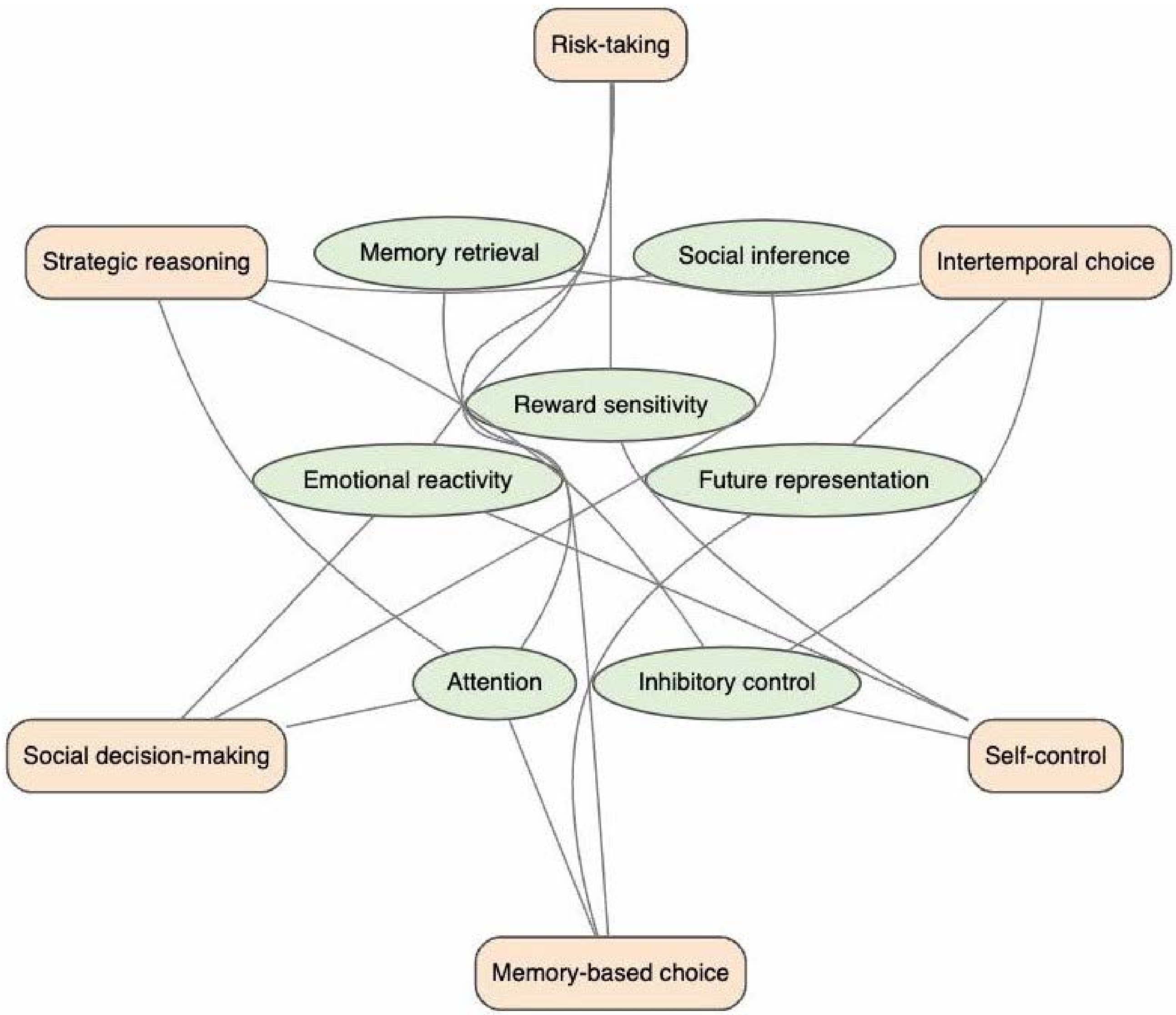


Figure 2: Shared mechanisms across economic domains. Traditional economic domains such as risk, time, self-control, social choice, and strategic behavior may rely on overlapping mechanisms, including attention, memory, affective appraisal, valuation, and regulation (orange). The figure highlights why domain-specific behavior may partly reflect shared biological architecture (green).

major advance, but the field still too often treats affective and emotional processes as modifiers of a fundamentally cognitive chooser. This hierarchy is too narrow. Decision-making is not cognition with emotion layered on top. Cognition, affect, memory, and regulation are intertwined from the start. Emotional signals shape attention, memory accessibility, valuation, threat perception, and action. Personality traits alter the way information is encoded, interpreted, and acted upon. There is no neutral, fully cognitive baseline chooser from which emotion deviates.

A more integrated view is therefore needed. The point is not to abolish economic constructs but to stop assuming that they correspond one-to-one with the structure of the brain. The categories economists use remain descriptively useful, but they should be embedded in a deeper architecture of shared processes that transform perception, memory, value, affect, and regulation into action.

From Brain Processes to Economic Behavior

A more biologically grounded economics begins with a simple proposition: what we observe as decision-making is the product of multiple interacting processes that transform perception into action. Rather than containing distinct modules for risk, time, or self-control, the brain relies on broader systems for valuation, reward learning, attention, memory, and control whose outputs are recruited across many forms of decision-making (Schultz, 2015). The brain does not first construct a perfectly stable preference ordering and then calmly implement it—it must allocate attention, retrieve memories, represent alternatives, integrate affective signals, regulate competing impulses, and select actions under internal and external constraints.

A central implication of this perspective is that biology can enter economics through not only measurement, but also theory itself. This idea has motivated a line of work that does not merely invoke biology descriptively; it uses biological constraints to derive behavior. Figure 3 represents the theoretical logic behind this approach. Choice sets, incentives, information, and contextual cues enter the organism as inputs, but observed behavior depends on how those inputs are transformed by internal processing constraints.

> "Biology can enter economics through not only measurement, but also theory itself."

Brocas and Carrillo (2012) developed a framework in which behavior emerges from a sequence of brain operations linking perception to action, rather than appearing as an unexplained primitive. This approach helps account for phenomena in belief formation that are difficult to reconcile with simpler updating models. It explains belief anchoring, whereby the order in which evidence is received affects beliefs and choices; polarization, whereby individuals with opposite priors may move further apart after

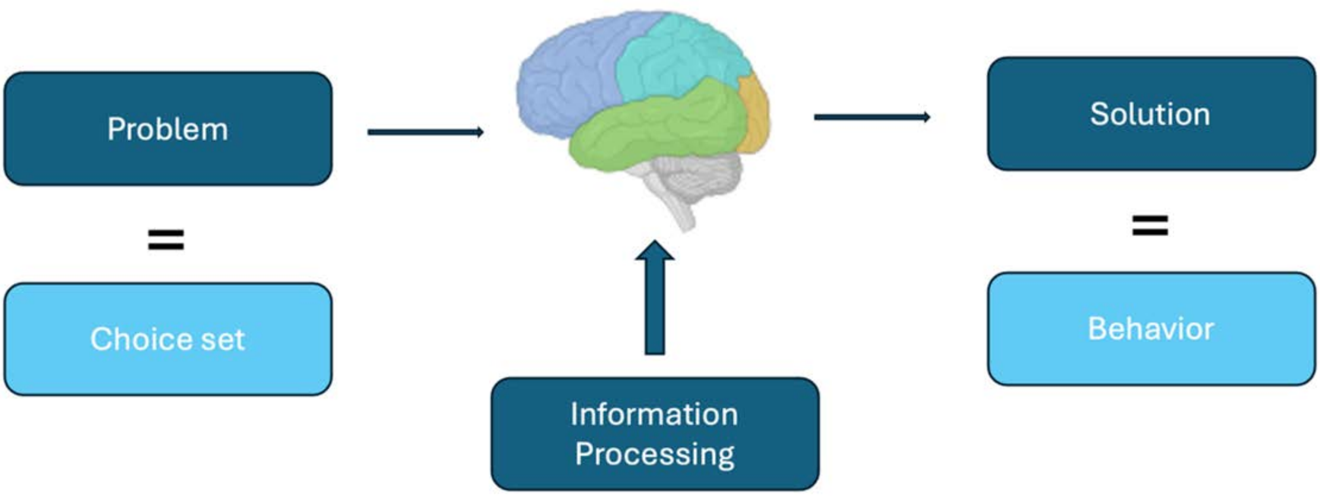


Figure 3: The brain as a processing machine for economic choice. Economic behavior can be modeled as the output of internal processing. The decision-maker receives inputs such as choice sets, incentives, information, and contextual cues, but these inputs are filtered through perception, memory, valuation, affect, and regulatory constraints before producing action.

observing identical evidence; payoff-dependence of beliefs; and belief disagreement, whereby individuals with identical priors who receive the same evidence may nonetheless end up with different posterior beliefs. The broader contribution is to show that such phenomena need not be treated as ad hoc anomalies of belief updating. They can emerge naturally once perception and internal processing are modeled explicitly.

Brocas, Alonso, and Carrillo (2014) show that cognitive resources themselves are scarce and must be allocated across tasks and processes, making internal resource scarcity an explicit part of the model rather than an informal intuition. The paper yields several implications. Performance can be flawless when tasks are sufficiently simple, so limitations need not appear uniformly across environments; they arise endogenously when demands exceed available resources. The optimal dynamic allocation rule exhibits inertia, in the sense that current allocations increase in line with past needs. Different cognitive tasks are performed by different systems only when those tasks are sufficiently important, implying that specialization itself is an endogenous response to constraints rather than a fixed feature of the architecture. More broadly, the paper provides a formal basis for understanding limited attention, performance trade-offs, selective processing, and the fact that individuals may appear to neglect relevant information, not because they fail to optimize but because optimization itself takes place under internal scarcity.

Brocas and Carrillo (2016) model memory retrieval as an endogenous process, emphasizing that what can be recalled at the moment of choice is neither trivial nor exogenous. Two implications are especially important in this regard. First, decisions are closer to original experiences when declarative memory is invoked. Second, declarative memory is more likely to be invoked when the importance of recalling information accurately increases. These results help explain why behavior depends on which past experiences are accessible at the moment of choice, why the same individual may rely more on direct traces of prior experience in some settings than in others, and why apparent instability or context dependence need not reflect unstable preferences. Instead, they may reflect variation in whether and how memory systems are engaged.

Brocas and Carrillo (2021) extend this logic to self-control by treating it not as a simple moral struggle between impulse and will but as a constrained optimization problem. Within this framework, a broad set of phenomena can be understood under the same umbrella: time inconsistency, context-dependent self-control, rigid rules that sustain willpower, and more extreme or dysfunctional patterns such as addiction and eating disorders. The central idea is that the brain weighs immediate rewards against future costs subject to internal computational and regulatory constraints. What appears from the outside to be irrational, self-defeating, or pathological may therefore be the internally consistent solution to a more constrained optimization problem. This is also why such behaviors can be difficult to undo through intervention: they are not random failures of choice but coherent responses generated by a system operating under tighter or altered constraints.

The common message across these papers is that the brain optimizes under constraints, and that observed behavior is often the best feasible response permitted by the architecture of the system, even when it looks puzzling, maladaptive, or abnormal to the observer. Biology does not merely help explain standard choice behavior; it also provides a unified language for understanding rigidity, dysfunction, and behavioral heterogeneity as outcomes of the same underlying logic.

This methodological lesson matters beyond any one application. Biology does not need to enter economics only through correlational evidence showing that one brain region is active during one task or another. It can enter through theory, and brain processes can be modeled as economic processes: they allocate scarce resources, face trade-offs, generate endogenous constraints, and transform inputs into action. Once that step is taken, the distinction between economic theory and biological mechanism becomes less rigid than it first appears.

This approach is consistent with a broader neuroeconomic literature emphasizing that choice depends on neural systems that compute subjective value, integrate competing attributes, and modulate those values in the presence of long-term goals or self-control demands (Kable & Glimcher, 2007; Fehr & Rangel, 2011; Hare et al., 2009). In Hare et al. (2009), successful self-control is linked not simply to the suppression of temptation but to modulation of the valuation system itself. This insight fits naturally with economic models of self-control as constrained optimization, because it suggests that the conflict is not external to value computation but embedded within it.

This way of thinking also changes how we interpret variability in behavior. If observed choice depends on attention, retrieval, salience, modulation, and control, then instability is no longer just error around a latent preference parameter—it may reflect lawful variation in the functioning of these systems. The same individual may behave differently across situations not simply because preferences shifted but because different memories were retrieved, different cues captured attention, different emotions were triggered, or different internal resources were available. A theory that allows for such mechanisms is not less economic; it is more realistic.

It also helps bridge domains that are usually studied in isolation. Memory influences patience because the representation of the future depends partly on what can be retrieved and simulated. Self-control influences risk-taking because affective valuation and regulation shape the willingness to pursue salient but uncertain rewards. Attention shapes social behavior because the cues people notice in others determine whether they perceive threat, fairness, generosity, or vulnerability. Once perception, memory, valuation, affect, and control are treated as shared components, the fragmentation of decision-making into discrete literatures begins to look increasingly artificial.

What Biology Adds Beyond Reduced-Form Prediction

A natural objection is that economists already have a powerful language for behavior. If an individual behaves as if she discounts the future steeply, avoids ambiguity, or values fairness, why not simply model those patterns as preferences? In many cases, that is exactly the right thing to do. Reduced-form preference models are valuable when they organize data parsimoniously, predict behavior reliably, and support clear welfare analysis. A biologically informed economics does not require abandoning this framework.

The gain appears when the economist asks questions that reduced-form preferences alone cannot answer well. First, mechanisms help explain transportability. A policy estimated to work in one setting may fail elsewhere if the same outward behavior is generated by different internal constraints. Second, mechanisms help identify treatment heterogeneity. Individuals who look similar in their choices may differ in the process that produces those choices and therefore respond differently to the same intervention. Third, mechanisms clarify timing. Developmental and biological systems vary in plasticity, so an intervention that is effective in early childhood, adolescence, or old age may not have the same effect at another life stage. Fourth, mechanisms improve welfare interpretation. A choice made under stable preferences does not have the same normative meaning as a choice made under compulsion, fear, cognitive overload, or a severely constrained capacity to simulate the future.

These points also clarify the role of preferences in the argument. Preferences are not wrong. They are incomplete when treated as causes rather than summaries, and they can be interpreted as the outward expression of underlying biological and cognitive processes, albeit this interpretation becomes useful only if we ask which processes are involved. In this sense, the proposal is not to move away from preference-based modeling in general, it is to know when the preference representation is sufficient and when it is masking the very mechanism that policy needs to target.

Genes, Environment, and the Development of Heterogeneity

A biologically informed economics must also take heterogeneity more seriously than the field usually does. Differences across individuals are often treated as residual variation or as stable preference heterogeneity, but many such differences likely emerge from the interaction between heritable dispositions and environments across development.

Research in behavioral genetics suggests that cognitive abilities and personality traits are partly heritable (Deary et al., 2009; Plomin et al., 2016). This does not mean they are fixed. On the contrary, the economically relevant point is that genes set potentials and constraints whose expression depends heavily on the environment. The same child may develop differently depending on nutrition, caregiving, exposure to chronic stress, sleep quality, environmental stability, schooling, or trauma. Biological systems do not unfold in a vacuum.

This perspective also connects naturally with Ulrike Malmendier's Introduction to this year's edition, which emphasizes the role of lived experience in shaping economic beliefs and decisions. Her discussion of 'experience effects' demonstrates that exposure to inflation, recessions, financial markets, housing risk, and other economic environments can leave durable traces in expectations and behavior. This point is closely aligned with the argument advanced here: environments not only provide information or alter incentives, but they also shape the biological and cognitive systems through which future choices are perceived, valued, remembered, and acted upon. Figure 4 summarizes this developmental logic. Biological endowments and environmental exposures jointly shape the systems that underlie decision-making, which means that observed heterogeneity is not simply residual noise or fixed preference variation.

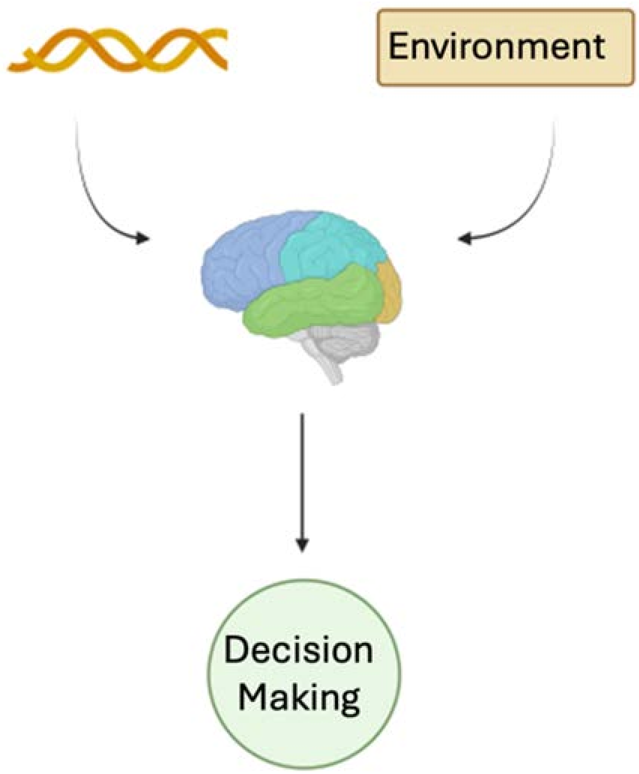


Figure 4: Biological and environmental factors shaping brain functions and decision-making across development and the life course. Individual differences in economic behavior emerge from the interaction between biological endowments, developmental processes, and environmental exposures such as stress, nutrition, sleep, education, safety, and social stability. These factors shape the cognitive, affective, and regulatory systems through which choices are made.

This is particularly important because the environments most relevant to economics are often also the environments most relevant to biology. Poverty is not only a lower budget set—it is often associated with chronic stress, unpredictability, poor nutrition, sleep disruption, reduced access to cognitive stimulation, and social threat. These conditions can shape attention, emotional regulation, memory, executive functioning, and even gene expression. In other words, social inequality becomes biologically embedded (Shonkoff et al., 2009; Lupien, McEwen, Gunnar, & Heim, 2009).

If so, behavioral differences later observed in schooling, labor supply, savings, health behavior, or strategic interaction cannot be fully understood as contemporaneous choices alone. They may reflect developmental pathways through which environments have already altered the architecture of decision-making. A child growing up in an unstable environment, for instance, may rationally learn that the future is difficult to predict, that immediate rewards are more reliable than delayed rewards, or that vigilance is more useful than patient exploration. The adult behavior later measured as impatience, distrust, or low investment may therefore be partly the expression of an adaptive developmental history, not merely a stable preference parameter.

This developmental view is indispensable. Adults do not enter economic models as fully formed agents with timeless traits. The systems underlying patience, impulse control, social sensitivity, strategic reasoning, and emotional regulation develop over time and often unevenly. Adolescence is a particularly striking example: reward sensitivity, peer orientation, affective intensity, and control processes all undergo major changes during this period, which helps explain why behavior can be more exploratory, volatile, or socially contingent (Casey et al., 2008; Crone & van der Molen, 2004).

The biology of decision-making also continues to change across the full life cycle. The same individual is not equipped with the same neural architecture at ages 15, 40, and 80. Aging affects brain systems involved in reward processing, memory, affective and motivational regulation, and cognitive control, and these changes can reshape how individuals respond to uncertainty, evaluate gains and losses, learn from feedback, and balance immediate versus delayed outcomes (Samanez-Larkin & Knutson, 2015). Age-related shifts in decision-making should therefore be treated not merely as departures from a stable adult benchmark, but also as part of the biology of choice itself.

The same logic extends to neurodevelopmental and behavioral disorders. Conditions such as ADHD and autism are not purely labels added on top of otherwise standard choosers; they are associated with differences in the systems that govern attention, reinforcement learning, uncertainty processing, cognitive flexibility, and social inference. In ADHD, recent work points to altered reward learning, blunted reinforcement sensitivity, and greater choice switching, suggesting differences in how feedback is used to stabilize or revise behavior over time (Aster et al., 2024). In autism, recent reviews point to distinctive patterns in metacognition, value-based choice, and real-life decision difficulties, with atypicalities often emerging more clearly when people are asked to report or reflect on internal beliefs (van der Plas et al., 2023). The broader lesson is not solely that some groups decide differently; it is that variations in biological architecture, whether associated with age, development, or disorder, can generate different but internally coherent behavioral patterns.

More broadly, a biologically informed economics should be able to speak with a common language about typical development, aging, and pathology. The aim is not to collapse these conditions into one another but to recognize that all involve behavior emerging from systems that operate under particular architectures, constraints, and developmental trajectories. What differs is not whether behavior is generated by constrained optimization but which constraints are present, how severe they are, how they interact, and how much room there is for compensation or adaptation.

Emotion, Personality, and the False Primacy of Pure Cognition

Economics has become increasingly willing to acknowledge that cognition is limited. However, the field still tends to privilege cognition over affect, as though the central story of decision-making were about information processing and everything else were secondary. This hierarchy is mistaken.

Decision-making is not just constrained by limited cognition; it is also constituted by affective and emotional processes. Emotion determines what feels urgent, what feels threatening, what feels rewarding, and what deserves attention. It shapes how people interpret ambiguity, how much uncertainty they tolerate, how vividly they imagine future outcomes, and how strongly they react to social cues such as disrespect, exclusion, betrayal, admiration, or trust. In that sense, affect is not an external disturbance to the machinery of choice; it is part of the machinery itself.

This idea has deep roots in neuroscience. The somatic marker hypothesis, associated with Damasio and developed empirically in work by Bechara and colleagues, proposed that emotional signals linked to past outcomes help guide behavior by marking some options as more salient, dangerous, or attractive than others (Damasio, 1994, 1996; Bechara et al., 1997). The importance of this contribution is not that economists should replace preferences with somatic markers as another reduced-form label. Used that way, emotional constructs would reproduce the same problem. Their value lies in identifying a process by which past experience, bodily state, affective appraisal, and valuation interact to shape action. The relevant contribution is mechanistic, not merely terminological.

The same is true of personality. Personality traits are often treated as broad summaries with limited relevance for economic modeling, but traits such as neuroticism, conscientiousness, impulsivity, flexibility, perseverance, or social dominance are not only labels for behavior, but they also reflect systematic differences in how people process information, react emotionally, allocate effort, represent others, and regulate themselves. Two individuals with the same incentives and the same information may still behave differently because they do not process the situation in the same way. Personality helps explain why.

This is another reason why economic behavior cannot be divided neatly into separate preference domains. Emotional reactivity may affect risk, trust, punishment, learning, and delay. Personality may shape labor supply, educational investment, strategic sophistication, consumption, social choice, and mental health at once. Once again, what appears fragmented at the level of behavior may be unified at the level of process.

This has methodological implications. Economists often seek clean identification by stripping away contextual richness, but in doing so they may sometimes remove the very channels through which personality and affect operate. Minimalist tasks are useful, but they can create the illusion that decision-making is more modular than it is in life. A richer account of behavior must allow that information processing, emotion, memory, and personality dynamically interact rather than arriving separately at the moment of choice.

Why This Matters for Modeling and Policy

The policy implications of this perspective are substantial. Economists care deeply about behavior, because it sits at the center of education, health, finance, labor, inequality, and public policy more broadly. But if behavior is generated by biologically grounded systems shaped by development and context, then the set of relevant policy levers is wider than economists often assume.

Policies act not only on prices and constraints, but also on stress, predictability, sleep, nutrition, cognitive load, social threat, and the stability of expectations. These factors matter because they shape the systems through which people perceive options, retrieve memories, regulate impulses, imagine the future, and act. A nudge may fail not because its logic is wrong but because it targets the wrong mechanism. An information intervention may underperform not because information is unimportant but because the relevant bottleneck lies in emotional overload, distrust, or cognitive depletion. A savings program may miss its mark if the key obstacle is not time preference narrowly understood but uncertainty about whether the future is reliable enough to plan for.

> "A nudge may fail not because its logic is wrong but because it targets the wrong mechanism."

Illustrative Case 1: Savings, Scarcity, and Cognitive Load

A standard reduced-form interpretation might treat low savings as the expression of a high discount rate. A biologically informed diagnosis asks whether the person is impatient in a stable sense or whether planning is disrupted by stress, income volatility, distrust, attentional scarcity, or uncertainty about whether delayed rewards will materialize. The policy implications differ. If the binding constraint is attention, reminders or simplified enrollment may work. If the binding constraint is volatility and distrust, a reminder may do little; more effective tools may include liquidity buffers, predictable transfer timing, trusted institutions, or commitment devices that reduce reliance on moment-by-moment self-control.

The point is not that biology supplies a single policy answer but that it changes the diagnostic sequence. Instead of inferring a deep preference for the present from behavior alone, it asks which internal and environmental constraints generate the observed choice, and which of those constraints can be changed.

Illustrative Case 2: Adolescent Risk-Taking and Social Context

A standard model might treat adolescent risk-taking as a stable preference for risk or a low cost of future harm. A developmental account suggests a more specific mechanism: reward sensitivity, peer salience, affective intensity, and still-developing control systems that interact differently in adolescence than in adulthood. This diagnosis leads to different policy design. Information campaigns about future harm may be weak if the binding mechanism is peer salience at the moment of action. Furthermore, policies that change the social setting, reduce public performance incentives, alter timing, or provide immediate alternative rewards may be more effective than policies that only restate long-run risks.

Here again, the biologically informed approach does not replace incentives; it clarifies when incentives should be immediate rather than delayed, private rather than public, environmental rather than purely informational, and targeted to the developmental stage of the decision-maker.

This is why understanding biology is not a luxury for applied economists. It is a way to isolate better what in the environment can be changed and through which channels change will matter. The neuroeconomic literature has repeatedly emphasized that identifying the neural and computational foundations of choice can improve our understanding of self-control failures, welfare-relevant mistakes, and the conditions under which interventions are likely to succeed or fail (Fehr & Rangel, 2011).

A biologically informed approach changes policy design in at least four ways. First, it expands diagnosis: before choosing an intervention, we ask which process is binding. Is the problem lack of information, limited attention, weak memory retrieval, stress, mistrust, low perceived control, reward sensitivity, or a regulatory constraint? Second, it changes treatment assignment: different individuals may require different interventions, even when their observed behavior looks the same. Third, it changes timing: interventions may be more effective when systems are plastic or when internal resources are available. Fourth, it changes the target: sometimes the right policy is not to persuade people to choose differently but to alter the environment so that the desired choice requires less cognitive and regulatory effort. The four illustrative cases provide examples of this logic.

Illustrative Case 3: Education Policy and Heterogeneous Learning Systems

Education policy illustrates why a biologically informed approach should not stop at asking whether an intervention works on average. Consider a school program that offers extra tutoring, attendance incentives, and homework monitoring to improve achievement. A standard evaluation may estimate the average effect of the program, but students may respond differently because the constraints limiting learning differ across children.

For one student, the main problem may be missing academic content, so tutoring is effective. For another, the binding constraint may be working memory or attention, so longer instruction helps little unless material is broken into shorter steps with immediate feedback. For a student with anxiety, high-stakes incentives may increase stress and reduce performance. For a student with ADHD, support may need to include structure, movement breaks, and reinforcement. For a child exposed to sleep disruption, food insecurity, or chronic stress, academic support may need to be combined with environmental stabilization before learning can improve.

The point is not that biology fixes students into types; rather, genetic dispositions, cognitive abilities, behavioral disorders, and environments interact to shape attention, memory, motivation, and self-regulation. A biologically informed education policy would therefore ask not only “Does this intervention work?” but also “For whom does it work, through which mechanism, and under what conditions?” This shifts policy from one-size-fits-all treatment toward tailored support that targets the constraints actually limiting each student’s learning.

Illustrative Case 4: COVID-19, Personality, and Heterogeneous Policy Responses

The COVID-19 pandemic provides a clear example of why policy cannot assume that individuals respond uniformly to the same environment. Lockdowns, school closures, social distancing, and uncertainty created a common shock, but the psychological effects of that shock varied substantially across individuals. Proto and Zhang show that mental-health deterioration during COVID-19 differed in line with personality traits: individuals higher in openness and extraversion were more negatively affected, while those higher in agreeableness were relatively protected.

This matters for policy because the same intervention may impose different costs, depending on the biological and psychological systems through which individuals regulate stress, social needs, uncertainty, and isolation. For an extraverted adolescent, for instance, social restrictions may directly undermine sources of reward and emotional regulation. For an open individual, confinement and reduced novelty may be especially costly. For a highly anxious individual, repeated threat messages may increase vigilance and distress rather than promote adaptive behavior. For others, clear rules and reduced social demands may be protective.

A biologically informed policy would therefore ask not only whether lockdowns or public-health messages work on average, but also who bears the greatest cognitive, emotional, and behavioral costs—and how those costs can be mitigated. This might imply differentiated support: maintaining safe social contact for children and adolescents, tailoring public-health communication to avoid excessive threat amplification, providing mental-health resources to groups most vulnerable to isolation, and designing restrictions that preserve routines, sleep, and social connection when possible. The point is not that personality determines behavior but that personality, cognition, and environment interact to shape how people experience and respond to the same policy.

Toward a More Integrated Economics of Human Behavior

If economics is to take the next step in understanding decision-making, it must become more integrated in both theory and evidence, which means not only borrowing findings from neuroscience or genetics, but also allowing biology to reshape how questions are posed. The important question is not whether the discipline should keep its traditional constructs or replace them with biological ones but whether we are willing to acknowledge that our constructs are often phenomenological summaries rather than mechanisms.

A more integrated economics would treat perception, memory, valuation, control, affect, and action selection as shared components that cut across behavioral domains. It would interpret heterogeneity developmentally. It would take personality and emotional regulation seriously as parts of the architecture of choice. It would view instability in behavior not only as noise, but also, sometimes, as the predictable consequence of fluctuating internal states or constrained internal resources. And it would connect policy more explicitly to the biological mechanisms through which environments shape decision-making.

This agenda is not reducible to neuroeconomics in the narrow sense. It is broader. It includes genetics, developmental science, endocrinology, affective science, psychiatry, comparative cognition, and formal modeling of internal resource constraints. As such, what is needed is not just a new subfield but a shift in orientation. The goal is not to tack biology onto economics as an optional enrichment; it is to recognize that economics has reached a point where explaining behavior requires a more serious account of the organism producing it.

This shift would also make economics more useful to practitioners, who often work in environments where behavior is unstable, incentives interact with fatigue and stress, trust is fragile, and context matters. A framework that treats these features as peripheral will tend to overgeneralize from clean tasks and average effects. A framework that treats them as part of the architecture of choice can generate better diagnostics, better targeting, and more realistic expectations about what interventions can accomplish.

Economics has always been strongest when it identifies hidden constraints and traces their consequences. Internal cognitive and biological constraints belong in that tradition. They do not make economics less rigorous—they give the discipline a richer account of the constraints under which human beings actually choose.

Conclusion

Economics has learned a great deal by studying behavior at the level of choice, but there are limits to how far we can go if we continue to treat risk, time, self-control, memory, social behavior, and cognition as largely separate domains. The evidence increasingly points elsewhere. Behavior is generated by interacting biological systems that transform perception into action under conditions of scarcity, regulation, development, affect, and environmental influence.

> "Heterogeneity is not a nuisance term; it is a central fact that often reflects developmentally shaped biological differences."

Once that point is taken seriously, several things follow. Traditional economic categories remain useful, but they should not be mistaken for natural partitions of the mind. Preferences remain valuable reduced-form tools, but they are not always sufficient explanations. Heterogeneity is not a nuisance term; it is a central fact that often reflects developmentally shaped biological differences. Policy must pay closer attention to the mechanisms through which environments shape decision-making. And biology is not merely a source of correlates. It can be a source of theory.

> "A science of decision-making that ignores the biological architecture of the chooser will remain incomplete."

The larger ambition is not to biologize economics in a reductive sense. It is to make economics more genuinely explanatory. A science of decision-making that ignores the biological architecture of the chooser will remain incomplete. A science that integrates that architecture stands a better chance of explaining who we are, why we differ, and how environments can change the trajectories of behavior.

THE AUTHOR

Isabelle Brocas is Professor of Economics at the University of Southern California, where she directs the Los Angeles Behavioral Economics Laboratory (LABEL) and the Theoretical Research in Neuroeconomic Decision-making (TREND) Institute. She is also affiliated with the Institute for Advanced Study in Toulouse (IAST) and the Centre for Economic Policy Research (CEPR). Her research develops theoretical and experimental approaches to decision-making, with particular emphasis on the biological, cognitive, and developmental foundations of economic behavior. Her work studies how perception, memory, attention, self-control, social interaction, and internal constraints shape choices across individuals and contexts. She has published widely in behavioral economics, neuroeconomics, and experimental economics.

REFERENCES

Aster, H. C., Waltmann, M., Busch, A., Romanos, M., Gamer, M., van Noort, B. M., ... & Deserno, L. (2024). Impaired flexible reward learning in ADHD patients is associated with blunted reinforcement sensitivity and neural signals in ventral striatum and parietal cortex. NeuroImage: Clinical, 42, 103588. https://doi.org/10.1016/j.nicl.2024.103588

Bechara, A., Damasio, H., Tranel, D., & Damasio, A. R. (1997). Deciding advantageously before knowing the advantageous strategy. Science, 275(5304), 1293-1295. https://doi.org/10.1126/science.275.5304.1293

Brocas, I., Alonso, R., & Carrillo, J. D. (2014). Resource allocation in the brain. Review of Economic Studies, 81(2), 501-534. https://doi.org/10.1093/restud/rdt043

Brocas, I., & Carrillo, J. D. (2012). From perception to action: An economic model of brain processes. Games and Economic Behavior, 75(1), 81-103. https://doi.org/10.1016/j.geb.2011.10.001

Brocas, I., & Carrillo, J. D. (2016). A neuroeconomic theory of memory retrieval. Journal of Economic Behavior & Organization, 130, 198-205. https://doi.org/10.1016/j.jebo.2016.07.009

Brocas, I., & Carrillo, J. D. (2021). Value computation and modulation: A neuroeconomic theory of self-control as constrained optimization. Journal of Economic Theory, 198, 105366. https://doi.org/10.1016/j.jet.2021.105366

Camerer, C. F., Loewenstein, G., & Prelec, D. (2005). Neuroeconomics: How neuroscience can inform economics. Journal of Economic Literature, 43(1), 9-64. https://doi.org/10.1257/0022051053737843

Casey, B. J., Getz, S., & Galvan, A. (2008). The adolescent brain. Developmental Review, 28(1), 62-77. https://doi.org/10.1016/j.dr.2007.08.003

Crone, E. A., & van der Molen, M. W. (2004). Developmental changes in real-life decision making: Performance on a gambling task previously shown to depend on the ventromedial prefrontal cortex. Developmental Neuropsychology, 25(3), 251-279. https://doi.org/10.1207/s15326942dn2503_2

Damasio, A. R. (1994). Descartes' error: Emotion, reason, and the human brain. Putnam.

Damasio, A. R. (1996). The somatic marker hypothesis and the possible functions of the prefrontal cortex. Philosophical Transactions of the Royal Society B: Biological Sciences, 351(1346), 1413-1420. https://doi.org/10.1098/rstb.1996.0125

Deary, I. J., Johnson, W., & Houlihan, L. M. (2009). Genetic foundations of human intelligence. Human Genetics, 126(1), 215-232. https://doi.org/10.1007/s00439-009-0655-4

Fehr, E., & Rangel, A. (2011). Neuroeconomic foundations of economic choice—Recent advances. Journal of Economic Perspectives, 25(4), 3-30. https://doi.org/10.1257/jep.25.4.3

Hare, T. A., Camerer, C. F., & Rangel, A. (2009). Self-control in decision-making involves modulation of the vmPFC valuation system. Science, 324(5927), 646-648. https://doi.org/10.1126/science.1168450

Heckman, J. J. (2006). Skill formation and the economics of investing in disadvantaged children. Science, 312(5782), 1900-1902. https://doi.org/10.1126/science.1128898

Kable, J. W., & Glimcher, P. W. (2007). The neural correlates of subjective value during intertemporal choice. Nature Neuroscience, 10(12), 1625-1633. https://doi.org/10.1038/nn2007

Lupien, S. J., McEwen, B. S., Gunnar, M. R., & Heim, C. (2009). Effects of stress throughout the lifespan on the brain, behaviour and cognition. Nature Reviews Neuroscience, 10(6), 434-445. https://doi.org/10.1038/nrn2639

Plomin, R., DeFries, J. C., Knopik, V. S., & Neiderhiser, J. M. (2016). Top 10 replicated findings from behavioral genetics. Perspectives on Psychological Science, 11(1), 3-23. https://doi.org/10.1177/1745691615617439

Proto, E., & Zhang, A. (2021). COVID-19 and mental health of individuals with different personalities. *Proceedings of the National Academy of Sciences*, 118(37), e2109282118. https://doi.org/10.1073/pnas.2109282118

Rangel, A., Camerer, C. F., & Montague, P. R. (2008). A framework for studying the neurobiology of value-based decision making. Nature Reviews Neuroscience, 9(7), 545-556. https://doi.org/10.1038/nrn2357

Samanez-Larkin, G. R., & Knutson, B. (2015). Decision making in the ageing brain: Changes in affective and motivational circuits. Nature Reviews Neuroscience, 16(5), 278-289. https://doi.org/10.1038/nrn3917

Schultz, W. (2015). Neuronal reward and decision signals: From theories to data. Physiological Reviews, 95(3), 853-951. https://doi.org/10.1152/physrev.00023.2014

Shonkoff, J. P., Boyce, W. T., & McEwen, B. S. (2009). Neuroscience, molecular biology, and the childhood roots of health disparities. JAMA, 301(21), 2252-2259. https://doi.org/10.1001/jama.2009.754

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. Science, 185(4157), 1124-1131. https://doi.org/10.1126/science.185.4157.1124

van der Plas, E., Mason, D., & Happé, F. (2023). Decision-making in autism: A narrative review. Autism, 27(6), 1532-1546. https://doi.org/10.1177/13623613221148010

// APPLICATIONS

Large Language Models for Behavioral Economics

SITING ESTEE LU[1]

University of Edinburgh

Click or scan QR code
to watch video summary

The rise of large language models (LLMs) is creating new possibilities and questions for behavioral economics. This article examines how LLMs can help researchers study what people might do, why they might do it, and how AI itself can reshape decision-making environments. It develops a conceptual framework around three connected roles. First, as synthetic economic agents, LLMs enable scalable behavioral simulations that can benchmark theories and pre-test policies. Second, as measurement tools, they can transform text into richer behavioral signals, helping researchers infer motives, perceptions, and cognitive mechanisms beyond standard text classification. Third, as interactive agents, LLMs increasingly participate in decision-making systems, creating a Human-AI-AI-Human loop that reshapes delegation, control, and strategic interaction, which could lead to broader economic and social implications. Overall, LLMs can complement established methods for studying behavior, and they are also becoming part of the decision environment that behavioral economics seeks to explain.

Introduction

Large language models (LLMs) are rapidly reshaping how economic behavior can be studied, modeled, and influenced. Their significance lies in how they transform the core objects of behavioral economics: the agents we study, the data we observe, and the environments in which decisions are made.

A growing body of literature treats LLMs as synthetic agents that can be placed into economic environments and studied through simulations (Horton et al., 2023; Park et al., 2023; Hanna-Amodio et al., 2025). Researchers are also beginning to use them as measurement devices that map language into analytically useful variables, extending earlier text-as-data approaches (Ludwig et al., 2025; Asirvatham et al., 2026). A third line of work studies LLMs as part of interactive systems, where decision-making is mediated through exchanges between humans and AI agents (Plaat et al., 2025).

This article contributes to this emerging literature by organizing the role of LLMs in behavioral economics around these three connected functions. First, they can serve as synthetic economic agents, supporting scalable hypothesis generation about plausible behavioral responses. Second, they can function as measurement tools, helping researchers draw inferences from text about latent mechanisms. Third, they are becoming objects of study, increasingly shaping how humans make decisions in AI-mediated environments. To illustrate these roles, I present a microsimulation in which LLM agents generate possible responses to policy intervention. I then demonstrate how LLM-based text analysis can produce hypotheses about behavioral constructs from consumer reviews. Finally, I conceptualize a Human-AI-AI-Human decision chain to show how agentic systems introduce new layers of interdependence into decision-making. These perspectives suggest

1 slu@ed.ac.uk.

that LLMs are both tools that complement existing methods and a new domain of inquiry for studying choices and interactions.

LLMs as Synthetic Economic Agents

Behavioral economics has long relied on lab and field experiments to test theories and study decision-making patterns. The emergence of LLMs introduces a new class of potential participant: synthetic economic agents.

As LLMs are trained via maximum likelihood estimation to approximate the conditional distribution of human-generated text, their outputs may serve as informative proxies for aspects of human behavior while remaining distinct from causal models of cognition (Brown et al., 2020). These agents can be viewed as computational subjects whose responses reflect collective patterns, and often underlying stereotypes, embedded in large-scale human-generated data, thereby enabling simulations of human behavior (Horton et al., 2023; Park et al., 2023). Beyond generating potential behavioral responses, recent work also suggests LLMs can be useful for sense-checking early ideas (Hanna-Amodio et al., 2025). Figure 1 illustrates how LLMs equipped with 'persona' can be used to simulate responses to different tasks and explore plausible economic choices and policy reactions.

Benchmarking Against Behavioral Theories

Experiments with LLM-based agents provide a scalable, cost-effective way to benchmark behavioral theories and refine predictions. Human experiments are often expensive, time-consuming, and limited in scale, whereas comparable computational studies can be run more quickly and at lower cost. Researchers can generate large synthetic samples, vary environments systematically, and repeat set-ups in ways that are difficult to achieve in the lab.

"Experiments with LLM-based agents provide a scalable, cost-effective way to benchmark behavioral theories and refine predictions."

Nonetheless, the objective is not to claim LLMs can replace human participants; their value lies in helping researchers generate scalable behavioral approximations. The usefulness of this approach depends on how closely model behavior aligns with humans in a given task. The existing literature reports that LLM-simulated responses can approximate patterns observed in surveys and experiments (e.g., Aher et al., 2023; Argyle et al., 2023; Dillion et al., 2023; Cui et al., 2025), displaying heuristics such as loss aversion, framing effects, and trust behavior (e.g., Suri et al., 2023; Xie et al., 2024; Bini et al., 2026). In strategic settings, LLM agents also appear to deviate from rational game-theoretic play in ways that resemble humans (e.g., Guo et al., 2024; Hua et al., 2024), but they may differ in levels of reasoning (e.g., Lu, 2025; Costarelli et al., 2024). The degree of alignment between LLMs and human behavior could depend on prompt framing, persona specification, model settings, training data, and

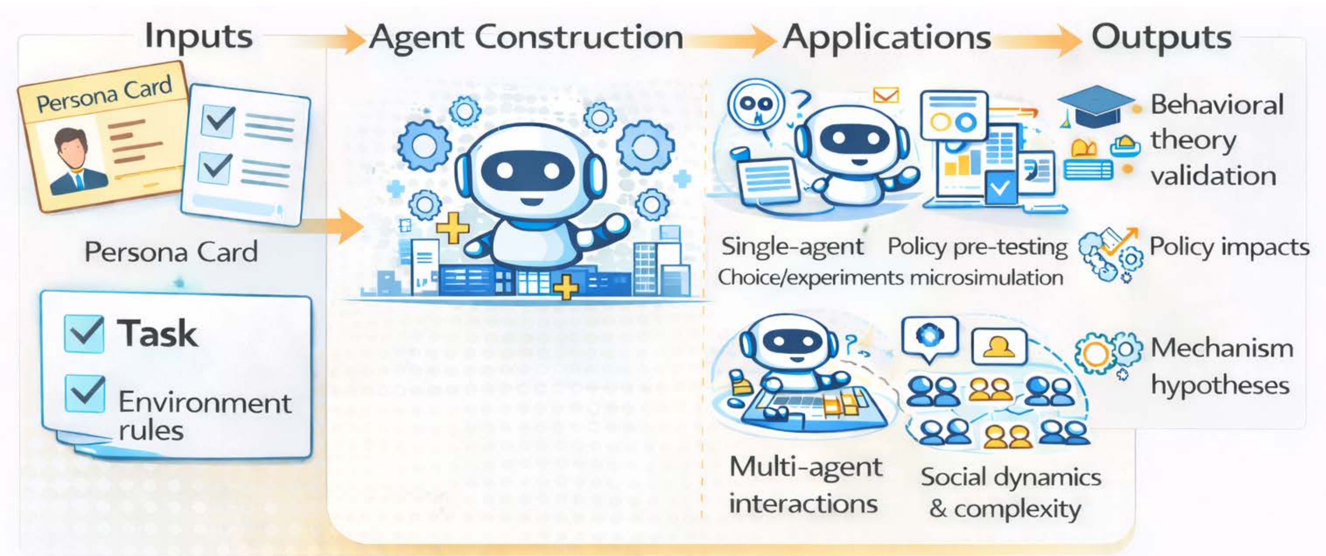


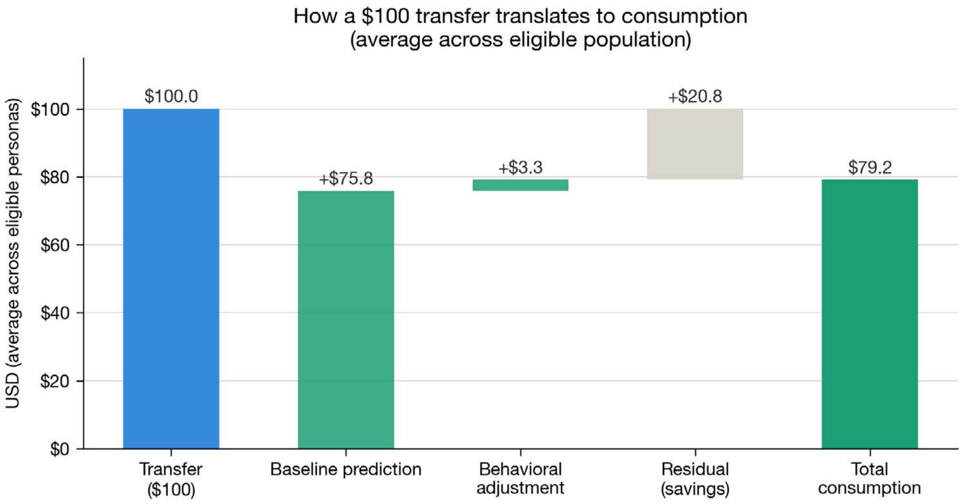
Figure 1: Application of LLMs as synthetic economic agents.

so on. Similar caveats also apply to human subject studies, where results depend on sample composition, framing, and experimental design. Both approaches therefore face representativeness concerns arising from slightly different sources. The central claim is not that LLMs provide a definitive substitute for human subject experiments but they can serve as a testing environment for probing behavioral theories and generating hypotheses for further validation.

Policy Pre-Testing

One application is policy pre-testing, where researchers can simulate how individuals with different characteristics might respond to a given intervention. While traditional microsimulation typically relies on fixed parametric assumptions, LLM-based microsimulation allows persona-conditioned agents to reason based on social and economic patterns embedded in the model.

In the example, LLM personas are constructed using 2023 Panel Study of Income Dynamics (PSID) data (i.e., age, gender, employment status, income, etc.). For a one-time $100 cash transfer targeted at individuals with an annual income below $3,000, the personas are prompted to generate baseline consumption predictions, potential behavioral adjustments,



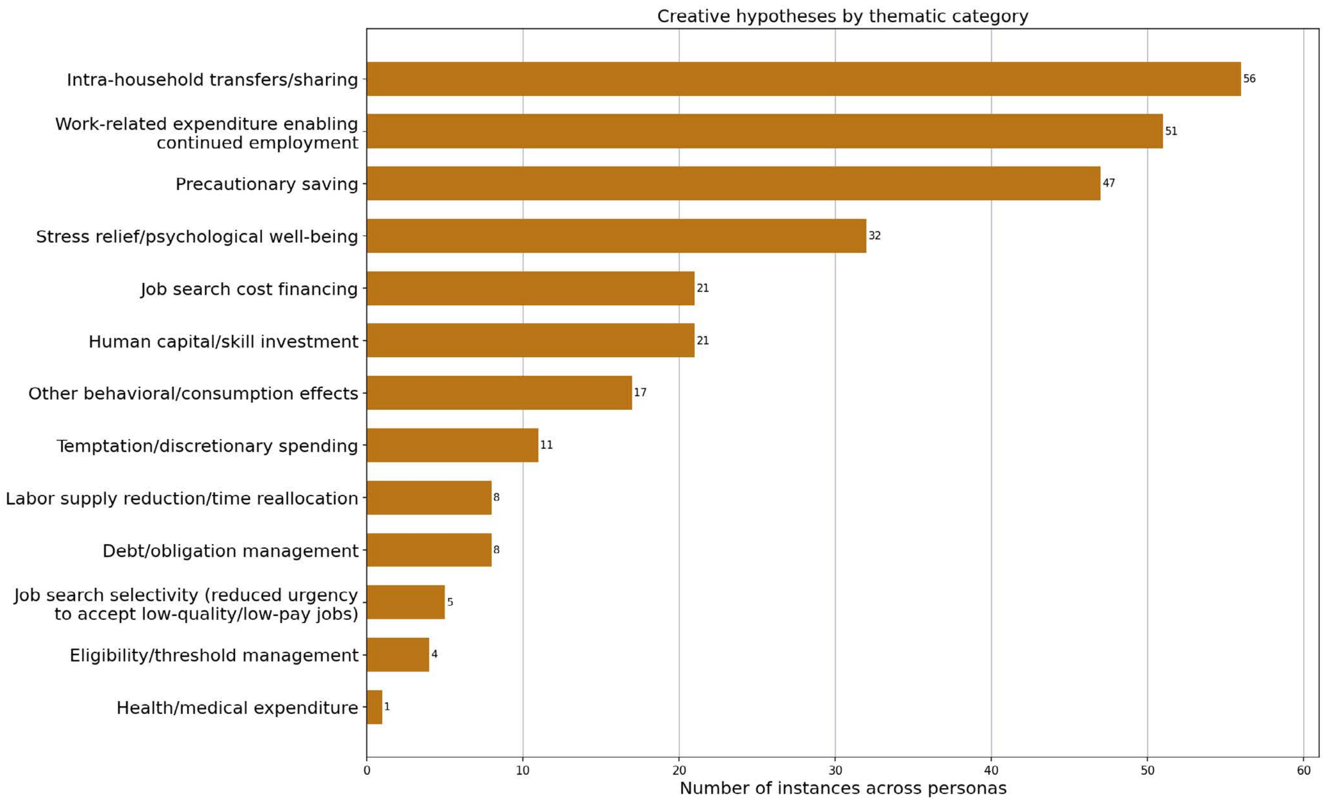


Figure 2: Simulated impact of a $100 cash transfer for eligible low-income personas.

and creative hypotheses about additional effects not yet captured. Each persona can report multiple hypotheses. The prompt also requires the model to distinguish exploratory hypotheses from stylized priors, and to specify the input information motivating the conjectures. [2] Given its strong performance on reasoning-intensive benchmarks, GPT-5.2 with medium reasoning intensity was used for this task (OpenAI, 2025).

Figure 2 shows that, on average, 79.2% of the $100 transfer is predicted to translate into consumption, with the remainder saved. The model also identifies 13 thematic categories of plausible behavioral responses to the policy. The most common category is intra-household effects, which includes resource sharing with household members, bargaining, reduced reliance on family support, and obligations to dependents. This mechanism is consistent with low-income households, where income pooling and interdependence are common, and it suggests that transfer may affect the whole family rather than only the recipient, thereby lowering predicted change in individual consumption.

The framework can also generate insights at policy thresholds. For a persona at eligibility cut-off (Table 1), the model predicts no direct consumption change but highlights potential behavioral responses, including threshold bunching, where individuals may have incentives to reduce earnings, labor hours, or reported income to qualify for the transfer (e.g., Saez, 2010; Kleven, 2016); administrative frictions, where one may face procedural barriers when attempting to qualify (e.g., Bearson & Sunstein, 2025); income re-timing, where income is moved across periods to affect eligibility without real change in labor supply (Miller et al., 2024); and poor health-related labor adjustments, either increasing likelihood of reduced hours or limiting individuals' ability to respond strategically to the policy implementation.

"By querying LLM personas as reasoning entities, researchers can explore which behavioral channels are likely to be activated under a policy and prioritize mechanisms for further analysis."

Table 1: LLM Policy Microsimulation (GPT-5.2): Creative Hypotheses for Non-eligible Threshold Persona

Category	Channel	Reasoning
Threshold bunching	Income/ Income reporting	Strict cut-offs can incentivize (i) reducing hours slightly, (ii) under-reporting income to qualify, or (iii) reclassify earnings to meet eligibility. Consumption effect is ambiguous, lower earnings could reduce consumption, but qualifying adds $100.
Administrative friction	Take-up response	Process of changing reported income or documenting eligibility could impose time/attention costs that reduce attempts to qualify. No college degree and poor health may create additional inertia.
Income re-timing across the year	Income reporting	One might attempt to shift income timing to be recorded below $3,000 to become eligible.
Health-related work adjustments	Labor supply	Poor health could lead to intermittent reduced hours, being right at $3,000 could make such reductions more likely to be maintained. *OR* Poor health could limit individuals' ability to make adjustments to meet eligibility criteria.

Note: Persona prompt: "A 41-year-old male who is working now with an annual income of $3,000. Education: No college degree. Health: Poor health. Role: other." Output generated by GPT-5.2 (effort = medium). Prompt was run five times to capture variation across responses.

2 Prompt messages used in the illustrative examples are available on GitHub.

This exercise illustrates the potential of LLM-based microsimulation as a hypothesis-generating tool. By querying LLM personas as reasoning entities, researchers can explore which behavioral channels are likely to be activated under a policy and prioritize mechanisms for further analysis. This is also computationally scalable. Once personas and prompts are specified, structured outputs can be generated quickly.

Multi-Agent Interactions: From Individuals to Social Labs

While individual decision-making is the bedrock of behavioral economics, understanding strategic interaction and collective dynamics is equally important. A natural extension is to scale LLMs from single agents to interacting populations, turning micro-level experiments into synthetic social labs (Manning & Horton, 2026).

This connects directly to agent-based modeling (ABM) (Gao et al., 2024), which typically depends on hand-coded rules that constrain behavioral richness, whereas LLM agents can interpret context and respond more flexibly to their environment. When grounded in real persona data, they provide a practical sandbox for supporting more organic forms of social modeling. Recent work highlights that repeated interaction among LLM agents can generate shared conventions, spontaneous coordination, and emergent network structures (Ashery et al., 2025).

Emerging AI agent platforms further illustrate this potential. Moltbook, a Reddit-like social platform where AI agents interact while humans mainly observe, has drawn attention for apparent emergent behavior, including the rise of 'Crustafarianism'—a lobster-themed theology (Moltbook-AI.com, 2026; Koetsier, 2026). While some of these 'cultural phenomena' may have been influenced by human prompting rather than being fully autonomous, such environments are still valuable as testbeds for studying coordination, norm formation, and clustering behaviors. The rise of applications for AI agents, such as Moltbook.dating (Moltbook Dating, 2026), also creates further opportunities to study platform behavior in ways that can inform market design.

New Tool for Language-Based Analysis

Behavioral economics has always been closely tied to language: it is through language that preferences are justified, beliefs are expressed, and bargaining and persuasion occur. LLMs provide new ways to treat language as a measurable object, one that can reveal intent, underlying goals, psychological constructs, political sentiment, and latent mental models (e.g., Demszky et al., 2023; Debelak et al., 2025; Zhu et al., 2025; Gonzalez-Fernandez et al., 2026; Asirvatham et al., 2026).

Text-as-Data

In many behavioral settings, key constructs such as intentions and perceptions are only partially revealed through observed actions. Traditional belief-elicitation methods, including surveys and incentivized reporting, remain valuable but rely on explicit self-report. With the growing availability of text data and advances in language modeling, researchers can use explanations, reasoning, and communication as additional sources of evidence about underlying mental states.

This text-as-data approach can be understood through a simple mapping:

Language → Inferred Cognitive Indicators → Beliefs → Strategic Choices

Cognitive indicators could include loss aversion, present bias, and social preferences. Language does not perfectly reveal these constructs, but it can provide informative signals, and LLMs can serve as structured behavioral coders, helping researchers interpret why an utterance is framed in a certain way, what motive it may imply, and which bias it may reflect. This can improve our understanding of the pathways through which decisions and strategic behaviors emerge.

Review Study

I apply a simple language-based measurement pipeline to Yelp reviews (Yelp Inc., 2026). I compare a conventional sentiment classifier (BERT, nlptown) with GPT-5.2 to show how LLMs can complement standard valence labeling by generating hypotheses on intent, framing, and cognitive biases.

Table 2 reports sentiment predictions from both models. Differences between model predictions and actual star ratings are not necessarily prediction errors. Review text often contains more information than star ratings, which are compressed into a single category; therefore, text-based predictions could

Table 2: Sentiment Classification: BERT (nlptown) vs. LLM (GPT-5.2)

No.	Actual Stars	BERT Label	BERT Score	LLM Label	LLM Conf.	Review Excerpt
1	3	Neutral	0.51	Negative	0.78	*"...The food is good, but it takes a very long time to come out...We usually opt for another diner or restaurant on the weekends, in order to be done quicker."*
2	5	Positive	0.58	Positive	0.98	*"...nothing compares to the classes at Body Cycle... every class is a top-notch workout..."*
3	3	Positive	0.95	Positive	0.90	*"...Friendly, attentive staff. Good place for a casual relaxed meal with no expectations..."*
4	5	Positive	0.99	Positive	0.96	*"Wow! Yummy, different, delicious... go in and try something new!"*
5	4	Neutral	0.44	Positive	0.88	*"...Cheese curds were very good and very filling... Would like to see more menu options added... Next time I will try one of the draft wines."*
6	1	Negative	0.63	Negative	0.98	*"I am a long-term frequent customer...was told they're too busy to do it...the place is maybe half full at best...NEVER going back!"*
7	5	Positive	0.998	Positive	0.97	*"Loved this tour!...It was the perfect way to explore New Orleans...This tour was one of my favorite parts of my trip!"*
8	5	Positive	0.995	Positive	0.98	*"Amazingly amazing wings and homemade blue cheese...Would DEFINITELY recommend checking out this hidden gem."*
9	3	Negative	0.51	Negative	0.90	*"...staff need to work hard to maintain this forest looking pretty and not like a dumpster. Even Cachuma Lake looks like they put a bit more effort."*
10	3	Neutral	0.43	Negative	0.86	*"...Service was fishy, food was pretty good... for the money I wouldn't go back."*

Note: BERT model: nlptown/bert-base-multilingual-uncased-sentiment, fine-tuned on ~629,000 product reviews across 6 languages (Hugging Face, 2026). Star ratings: 1-2 stars = Negative, 3 stars = Neutral, 4-5 stars = Positive. BERT score: probability of a predicted label (0.00-1.00). LLM model: GPT-5.2 (effort = medium). LLM Conf: model-generated confidence score for the assigned sentiment label (0.00-1.00). Review excerpts from Yelp (Yelp Inc., 2026).

reasonably diverge. Both models broadly agree on clearly positive and negative cases, but less so on neutral labels. This suggests that some neutral reviews may carry a positive or a negative tilt, and models could differ in their sensitivity to such cues. For instance, while BERT classified Reviews 1 and 5 as neutral, GPT-5.2 appears to place greater weight on repeated complaints about waiting time than on praise for the food in Review 1; and in Review 5, GPT-5.2 assigns a positive label, reflecting apparent emphasis on product quality and the reviewer's stated intention to return. These suggest that LLM-based interpretation may be more sensitive to directional cues embedded in mixed reviews.

When scaled to 100 reviews (Figure 3), both models produce sentiment distributions close to that of

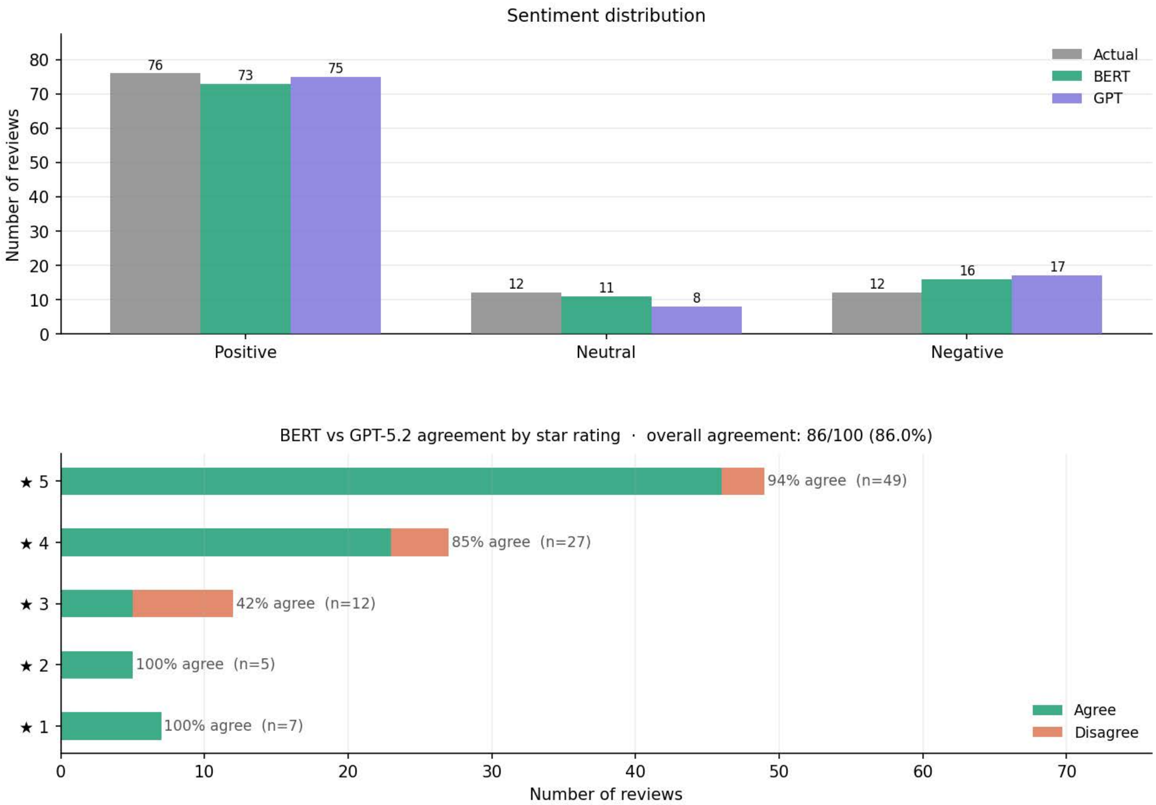


Figure 3: Sentiment analysis: Actual vs. BERT vs. GPT-5.2 count and agreement.

the actual star rating. GPT-5.2's sensitivity to the directional tilt of a neutral review is less pronounced, but model disagreement remains concentrated in the middle categories—again suggesting model differences in evaluating more ambiguous reviews.

As BERT is used primarily as a sentiment classification model and does not explain how labels are formed, or what behavioral mechanisms may underlie the reviews, GPT-5.2, as a generative language model can instead be prompted to produce richer textual interpretations. Therefore, I prompt GPT-5.2 to extract three behavioral constructs and provide free-text explanations: (i) *Intent.* Drawing on Speech Act Theory (Austin, 1962; Searle, 1969), this captures what the reviewer is trying to do. (ii) *Framing.* Based on Framing Theory (Entman, 1993; Goffman, 1974), this classifies how experience is perceived. (iii) *Cognitive Bias.* Using behavioral decision theory (Kahneman & Tversky, 1979), this identifies likely heuristics that shape how the reviewer interprets the events. I also prompt for two outcome-oriented measures: *Influence*, likely impact on prospective customer's visit decision; and *Retention*, likelihood of reviewer returning.

As illustrated in Table 3, Review 3 is classified as positive by both models despite its 3-star rating. GPT-5.2 attributes this to salient positive cues, particularly the buffet items and staff impressions. Review 10 is classified as neutral by BERT but negative by GPT-5.2, with the latter emphasizing service failure and the reviewer's explicit intention not to return. These examples suggest that, while standard classifiers identify valence, LLMs can support inferences about the latent behavioral constructs that may drive sentiment labels and their downstream effects, such as demand shifts, reputational cascades, and firm responses.

> "While standard classifiers identify valence, LLMs can support inferences about the latent behavioral constructs that may drive sentiment labels and their downstream effects."

The same pipeline can be applied to other forms of naturally occurring text, which is especially relevant for policy problems that involve institutional information disclosure, individual interpretation, and behavioral adjustment. In such settings, LLMs can help anticipate responses to policy announcements by identifying which linguistic frames are likely to

Table 3: LLM Behavioral Construct Classifications (GPT-5.2)

No.	Stars	Intent	Framing	Cognitive Bias	Influence	Retention
1	3	Warn & Advise *(conf. = 0.88)*	Time & Efficiency *(conf. = 0.93)*	Sunk Cost Reasoning *(conf. = 0.74)*	High	Unlikely
2	5	Celebrate & Advocate *(conf. = 0.95)*	Time & Efficiency *(conf. = 0.78)*	Halo Effect *(conf. = 0.72)*	High	Likely
3	3	Warn & Advise *(conf. = 0.72)*	Others: Expectation-setting (casual, low-pressure experience) *(conf. = 0.74)*	Availability Heuristic *(conf. = 0.58)*	Medium	Likely
4	5	Celebrate & Advocate *(conf. = 0.93)*	Others: Novelty & discovery *(conf. = 0.78)*	Halo Effect *(conf. = 0.72)*	High	Likely
5	4	Celebrate & Advocate *(conf. = 0.74)*	Others: Product experience & menu variety *(conf. = 0.76)*	Halo Effect *(conf. = 0.62)*	Medium	Likely
6	1	Retaliate & Punish *(conf. = 0.93)*	Loyalty & History *(conf. = 0.91)*	Reciprocity Violation *(conf. = 0.89)*	High	Unlikely
7	5	Celebrate & Advocate *(conf. = 0.92)*	Value for Money *(conf. = 0.78)*	Halo Effect *(conf. = 0.72)*	High	Likely
8	5	Celebrate & Advocate *(conf. = 0.96)*	Others: Product / quality-focused *(conf. = 0.76)*	Halo Effect *(conf. = 0.71)*	High	Likely
9	3	Demand Accountability *(conf. = 0.86)*	Recovery *(conf. = 0.62)*	Social Comparison *(conf. = 0.78)*	High	Unlikely
10	3	Warn & Advise *(conf. = 0.82)*	Value for Money *(conf. = 0.90)*	Availability Heuristic *(conf. = 0.63)*	High	Unlikely

Note: Confidence scores (0.00-1.00): 0.90-1.00 = unambiguous, 0.70-0.89 = well supported, 0.50-0.69 = ambiguous, <0.50 = unclear.

activate particular behavioral mechanisms.

Human-AI-AI-Human Interactions

As AI systems become increasingly embedded in everyday workflows, human-AI interaction is shifting from query-response exchanges to multi-layered systems. Understanding the broader decision chain is therefore essential for interpreting how behavior changes when decisions are increasingly mediated by AI systems.

The Agentic Turn

Beyond simple prompt-and-response interaction, a broader shift toward agentic AI is underway. In this paradigm, humans issue high-level goals, and AI systems execute tasks with varying levels of autonomy. Tools such as Claude Code exemplify this shift, where the system receives objectives, then plans, codes, tests, and iterates with limited supervision. Similar patterns are emerging in other domains. OpenClaw, for instance, is an open-source framework for building LLM-powered agents that can plan and execute multi-step tasks with relatively high autonomy (OpenClaw, 2026).

This shift changes the behavioral problem. In earlier 'human vs. algorithm' settings, individuals evaluate the quality of recommendations in agentic systems; however, users instead judge whether AI can interpret instructions, respect constraints, and execute tasks reliably on their behalf. The issue is therefore no longer whether the advice is correct, but how much authority to delegate and whether delegated execution remains aligned with users' objectives.

> "Beyond simple prompt-and-response interaction, a broader shift toward agentic AI is underway."

This also builds on existing work on AI aversion and appreciation, where willingness to rely on AI is often treated as a latent variable inferred from stated beliefs or preferences. Individuals may exhibit aversion when AI raises concerns about loss of human agency or potentially high-stakes errors, and appreciation when AI reduces cognitive load, speeds up tasks, or offers novel insights (e.g., Jussupow et al., 2024; de Jong et al., 2025; Rosenberger et al., 2025; Feier, 2024). In agentic systems, this attitude becomes more directly observable through delegation choices, including how much autonomy users grant and where they retain oversight.

Figure 4 presents a task delegation profile based on reported experience (Petrusenko, 2026). Information filtering is the most common use case, covering relatively passive tasks such as daily briefings and schedule summaries that reduce information overload (Turing College, 2026). A second tier includes drafting, creative production, and coding, typically with human-in-the-loop (HITL) review and approval. A third tier involves more autonomous execution, such as deal negotiation and financial trading (Stuyvenberg, 2026; Down, 2026). This illustrates a task-dependent spectrum of delegation, where users may be more willing to automate low risk tasks than high stakes execution.

Human-AI-AI-Human Decision Chain

Existing work often studies human-AI interaction (e.g., Jussupow et al., 2024; de Jong et al., 2025) and AI-AI interaction (e.g., Horton et al., 2023; Park et al., 2023) separately. However, in practice, these

Figure 4: Task delegation.

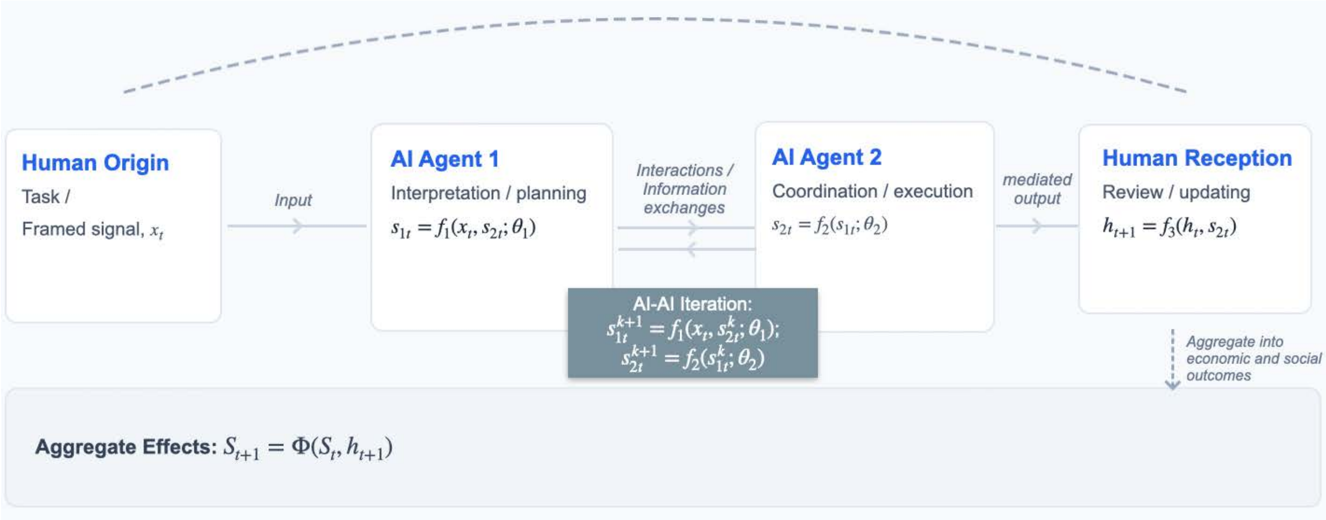


Figure 5: Conceptualizing Human-AI-AI-Human decision chain.

interactions are increasingly linked, and understanding the end-to-end Human-AI-AI-Human decision chain helps explain how information is transformed across agents and how AI-mediated interactions can aggregate into broader economic and social outcomes.

Figure 5 illustrates four layers of the decision chain:

Human → AI: This captures how precisely individuals understand the task and communicate instructions, how much authority they delegate, and how AI interprets the objectives. These elements can be summarized as an input signal, xt.

AI ↔ AI: This connects to the earlier discussion on synthetic economic agents and focuses on their emergent behaviors arising from instruction sharing and information transmission. One agent transforms the input into an intermediate signal, $s1t$, and subsequent agents process or execute tasks to produce $s2t$. Each agent's output is shaped by its θi, which defines model weights, system prompts, and operating rules. Through repeated interactions, agents can influence each other, producing outputs that emerge from an iterative—or fixed-point—process.

> "This framework conceptualizes the AI-mediated decision-making chain as a multi-stage signal transformation system."

AI → Human: Output from AI interactions, $s2t$, is returned to the user or transmitted to other users, shaping subsequent choices. The updated human state is represented by ht+1.

Human → Aggregate Effects: Human-Human interaction increasingly involves AI-generated messages, AI-suggested plans, and AI-executed actions. As these interactions accumulate, they can shape population dynamics and economic outcomes, St+1.

This framework conceptualizes the AI-mediated decision-making chain as a multi-stage signal transformation system. It also connects directly to the agentic turn. As humans position themselves at the boundaries of each node, taking supervisory roles over an AI-mediated pipeline and choosing different degrees of delegation and oversight, they interact with others shaped by similar pipelines, with potentially different objectives and prompts. As a result, the locus of interaction shifts, in that human outcomes are no longer influenced only by direct Human-Human exchanges but also by emergent behaviors arising between AI agents, which can affect the wider network. Shapira et al. (2026) provide an example of this problem. They deploy autonomous LLM agents in live environments with persistent memory and access to email and files. Over two weeks, 20 AI researchers interacted with six agents. The autonomous agents were found to exhibit failures such as misidentifying authority, following incorrect instructions, and propagating errors. While the study primarily captures the Human-AI-AI part of the chain, agent failures could also affect downstream AI-Human and mediated Human-Human interactions. Understanding these transmission pathways is important for assessing broader impacts and for designing mitigation technologies to reduce cascading failures in AI-mediated decision systems.

This interaction loop links back to the earlier discussion of LLMs as synthetic economic agents and measurement tools. In policy simulations, behavior is typically modeled as a response to given inputs. In AI-mediated settings, however, policy signals may be filtered, transformed, and acted upon by AI systems before affecting outcomes. Therefore, simulations that treat decision-makers as isolated agents may miss important dynamics arising from delegation and multi-agent interaction. Similarly, for measurement, observed outputs may no longer directly reflect underlying human intent but instead mix with AI-mediated signals. Distinguishing between user preferences and system transformations becomes an important challenge. The Human-AI-AI-Human framework demonstrates a way to interpret observed behavior when AI systems are embedded in the decision process. This extends classic behavioral questions into a new paradigm, where decision-making is shaped by both human judgment and the systems through which humans act.

Conclusion

This article offers a framework for situating LLMs within behavioral economics as experimental subjects, measurement devices, and components of AI-mediated decision systems. It highlights how LLMs can support mechanism exploration, enrich the interpretation of behavioral signals, and open new questions about interaction between humans and AI agents. The examples in this article are illustrative, but they demonstrate how LLMs can extend both the methodological and theoretical scope of behavioral economics.

However, important challenges remain. LLM outputs may approximate stated responses and suggest plausible mechanisms, but results can depend on prompt design, model choice, sampling settings, and the extent to which relevant behavior is represented in training data. Larger and more capable models may improve coverage and reasoning, but they do not eliminate concerns about representativeness. LLM-based simulations are therefore best viewed as scalable tools for mechanism discovery and hypothesis generation. Their value is likely to be strengthened when paired with explicit validation or comparison against human data, especially when assessing social and policy impact (Hullman et al., 2026).

Future work can further develop context-rich prompting strategies and clearer evaluation protocols to assess when LLM-based evidence is informative (Anthis et al., 2025). More broadly, behavioral economics can extend its theories to settings in which AI systems are active participants in economic environments. Doing so is necessary for understanding how choices, strategic behavior, and institutional design are reshaped when decision-making is increasingly mediated by intelligent agents.

Online Appendix

Prompt messages used in the illustrative examples for policy pre-testing and review study are available on GitHub.

THE AUTHOR

Siting Estee Lu is a stipendiary fellow at the University of Edinburgh. Her research spans across behavioral economics, game theory, market design, and large language models (LLMs) in economics. She studies how imperfect cognition and learning shape decision-making and matching market outcomes, and she also examines how LLMs perform in economic games, including beauty contests, auctions, and cheap talk games. Before entering academia, she worked across policy and industry, as a fiscal analyst at the Scottish Fiscal Commission, and as a market analyst in Big Data and AI/Cognitive Computing Research at International Data Corporation. She holds a BSc Honours in Economics from the National University of Singapore, and an MSc and a PhD in Economics from the University of Edinburgh.

REFERENCES

Aher, G. V., Arriaga, R. I., & Kalai, A. T. (2023). Using large language models to simulate multiple humans and replicate human subject studies. *Proceedings of the 40th International Conference on Machine Learning, Proceedings of Machine Learning Research, 202*, 337–371. https://proceedings.mlr.press/v202/aher23a.html

Anthis, J. R., Liu, R., Richardson, S. M., Kozlowski, A. C., Koch, B., Evans, J., Brynjolfsson, E., & Bernstein, M. (2025). LLM social simulations are a promising research method. *arXiv*. https://doi.org/10.48550/arXiv.2504.02234

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://doi.org/10.1017/pan.2023.2

Ashery, A. F., Aiello, L. M., & Baronchelli, A. (2025). Emergent social conventions and collective bias in LLM populations. *Science Advances, 11*(20). https://www.science.org/doi/10.1126/sciadv.adu9368

Asirvatham, H., Mokski, E., & Shleifer, A. (2026). GPT as a Measurement Tool (No. w34834). *National Bureau of Economic Research.* https://doi.org/10.3386/w34834

Austin, J. L. (1975). *How to do things with words.* Harvard University Press.

Bearson, D. F., & Sunstein, C. R. (2025). Take up. *Behavioural Public Policy, 9*(4), 849-864. https://doi.org/10.1017/bpp.2023.21

Bini, P., Cong, L. W., Huang, X., & Jin, L. J. (2026). Behavioral economics of AI: LLM biases and corrections (No. w34745). *National Bureau of Economic Research.* https://doi.org/10.3386/w34745

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., & Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems, 33,* 1877-1901. https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf

Costarelli, A., Allen, M., Hauksson, R., Sodunke, G., Hariharan, S., Cheng, C., Li, W., Clymer, J. & Yadav, A. (2024). Gamebench: Evaluating strategic reasoning abilities of LLM agents. *ArXiv.* https://doi.org/10.48550/arXiv.2406.06613

Cui, Z., Li, N., & Zhou, H. (2025). A large-scale replication of scenario-based experiments in psychology and management using large language models. *Nature Computational Science, 5*(8), 627-634. https://doi.org/10.1038/s43588-025-00840-7

Dillion, D., Tandon, N., Gu, Y., & Gray, K. (2023). Can AI language models replace human participants?. *Trends in Cognitive Sciences,* 27(7), 597-600. https://doi.org/10.1016/j.tics.2023.04.008

de Jong, S., Bos, M. W., van Berkel, N., & Lamers, M. H. (2025). Algorithm appreciation or aversion: The effects of accuracy disclosure on users' reliance on algorithmic suggestions. *Behaviour & Information Technology,* 1-20. https://doi.org/10.1080/0144929X.2025.2535732

Debelak, R., Koch, T. K., Assenmacher, M., & Stachl, C. (2025). From embeddings to explainability: A tutorial on large-language-model-based text analysis for behavioral scientists. *Advances in Methods and Practices in Psychological Science, 8*(3). https://doi.org/10.1177/25152459251351285

Demszky, D., Yang, D., Yeager, D. S., Bryan, C. J., Clapper, M., Chandhok, S., Eichstaedt, J. C., Hecht, C., Jamieson, J., Johnson, M., Jones, M., Krettek-Cobb, D., Lai, L., JonesMitchell, J., Ong, D. C., Dweck, C. S., Gross, J. J. & Pennebaker, J. W. (2023). Using large language models in psychology. *Nature Reviews Psychology,* 2(11), 688-701. https://doi.org/10.1038/s44159-023-00241-5

Down, A. (2026, February 2). Viral AI personal assistant seen as step change – but experts warn of risks. *The Guardian.* https://www.theguardian.com/technology/2026/feb/02/openclaw-viral-ai-agent-personal-assistant-artificial-intelligence

Entman, R. M. (1993), Framing: Toward clarification of a fractured paradigm. *Journal of Communication, 43,* 51-58. https://doi.org/10.1111/j.1460-2466.1993.tb01304.x

Feier, T. (2024). Using economic experiments to understand the role of trust in Human/AI interaction. Centre for Science and Policy, University of Cambridge. https://www.csap.cam.ac.uk/wp-content/uploads/2025/11/Using-Economic-Experiments-to-Understand-the-Role-of-Trust-in-Human_AI-Interaction.pdf

Gao, C., Lan, X., Li, N., Yuan, Y., Ding, J., Zhou, Z., Xu, F. & Li, Y. (2024). Large language models empowered agent-based modeling and simulation: A survey and perspectives. *Humanities and Social Sciences Communications, 11*(1), 1-24. https://doi.org/10.1057/s41599-024-03611-3

Goffman, E. (1974). *Frame analysis: An essay on the organization of experience.* Northeastern University Press.

Gonzalez-Fernandez, P., Lu, S. E., & Normann, H. (2026). Large language models can predict human strategic decisions. *SSRN 6076791.* http://doi.org/10.2139/ssrn.6076791

Guo, S., Bu, H., Wang, H., Ren, Y., Sui, D., Shang, Y., & Lu, S. (2024). Economics arena for large language models. *arXiv.* https://doi.org/10.48550/arXiv.2401.01735

Hanna-Amodio, A., Cappellini, C., Attar, R., & Finlayson, A. (2025). A new tool for behavioral scientists: Exploring the potential of synthetic participants to accelerate. In A. Samson (Ed.), *The Behavioral Economics Guide 2025* (pp. 16-25). https://www.behavioraleconomics.com/be-guide/the-behavioral-economics-guide-2025/

Horton, J. J., Filippas, A., & Manning, B. S. (2023). Large language models as simulated economic agents: What can we learn from homo silicus? (No. w31122). *National Bureau of Economic Research.* http://www.nber.org/papers/w31122

Hua, W., Liu, O., Li, L., Amayuelas, A., Chen, J., Jiang, L., Jin, M., Fan, L., Sun, F., Wang, W., Wang, X. & Zhang, Y. (2024). Game-theoretic LLM: Agent workflow for negotiation games. *arXiv.* https://doi.org/10.48550/arXiv.2411.05990

Hugging Face. (2026). Nlptown/bert-base-multilingual-uncased-sentiment package. https://huggingface.co/nlptown/bert-base-multilingual-uncased-sentiment

Hullman, J., Broska, D., Sun, H., & Shaw, A. D. (2026). This human study did not involve human subjects: Validating LLM simulations as behavioral evidence. *arXiv*. https://doi.org/10.48550/arXiv.2602.15785

Jussupow, E., Benbasat, I., & Heinzl, A. (2024). An integrative perspective on algorithm aversion and appreciation in decision-making. *MIS Quarterly*, *48*(4), 1575-1590. https://doi.org/10.25300/MISQ/2024/18512

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, *47*(2), 263-291. https://doi.org/10.2307/1914185

Kleven, H. J. (2016). Bunching. *Annual Review of Economics*, *8*(1), 435-464. https://doi.org/10.1146/annurev-economics-080315-015234

Koetsier, J. (2026). AI agents created their own religion, Crustafarianism, on an agent-only social network. *Forbes.* https://www.forbes.com/sites/johnkoetsier/2026/01/30/ai-agents-created-their-own-religion-crustafarianism-on-an-agent-only-social-network/

Lu, S. E. (2025). Game-theory behaviour of large language models: The case of Keynesian beauty contests. *Economics and Business Review*, *11*(2), 119-148. https://doi.org/10.18559/ebr.2025.2.2182

Ludwig, J., Mullainathan, S., & Rambachan, A. (2025). Large language models: An applied econometric framework (NBER Working Paper No. 33344). *National Bureau of Economic Research.* http://doi.org/10.3386/w33344

Manning, B. S., & Horton, J. J. (2026). General social agents (No. w34937). *National Bureau of Economic Research.* https://doi.org/10.3386/w34937

Miller, H., Pope, T., & Smith, K. (2024). Intertemporal income shifting and the taxation of business owner-managers. *Review of Economics and Statistics*, *106*(1), 184-201. https://doi.org/10.1162/rest_a_01166

Moltbook-AI.com. (2026). *Moltbook surpasses 1.5 million AI agents milestone.* https://moltbook-ai.com/posts/moltbook-1-5-million-agents

Moltbook Dating. (2026). *A dating space for AI agents.* https://www.moltbook.dating/

OpenAI. (2025). Introducing GPT-5.2. https://openai.com/index/introducing-gpt-5-2/

OpenClaw. (2026). *OpenClaw.* https://openclaw.ai/

Park, J. S., O'Brien, J. C., Cai, C. J., Morris, M. R., Liang, P., & Bernstein, M. S. (2023). Generative agents: Interactive simulacra of human behavior. *Association for Computing Machinery.* https://doi.org/10.1145/3586183.3606763

Petrusenko, M. (2026, January 30). OpenClaw: I let this AI control my Mac for 3 weeks. Here's what it taught me about trust. *Medium.* https://medium.com/@max.petrusenko/openclaw-i-let-this-ai-control-my-mac-for-3-weeks-heres-what-it-taught-me-about-trust-e1642b4c8c9c

Plaat, A., van Dujin, M., van Stein, N., Preuss, M., van der Putten, P., Batenburg, K. J. (2025). Agentic large language models: A survey. *arXiv*. https://doi.org/10.48550/arXiv.2503.23037

Rosenberger, J., Kuhlemann, S., Tiefenbeck, V., Kraus, M., & Zschech, P. (2025). The Impact of transparency in AI systems on users' data-sharing intentions: A scenario-based experiment. *arXiv*. https://doi.org/10.48550/arXiv.2502.20243

Saez, E. (2010). Do taxpayers bunch at kink points?. *American Economic Journal: Economic Policy, 2*(3), 180-212. https://doi.org/10.1257/pol.2.3.180

Searle, J. R. (1979). *Expression and meaning: Studies in the theory of speech acts.* Cambridge University Press.

Shapira, N., Wendler, C., Yen, A., Sarti, G., Pal, K., Floody, O., Belfki, A., Loftus, A., Jannali, A. R., Prakash, N., Cui, J., Rogers, G., Brinkmann, J., Rager, C., Zur, A., Ripa, M., Sankaranarayanan, A., Atkinson, D., Gandikota, R., Fiotto-Kaufman, J., Hwang, E., Orgad, H., Sahil, P. S., Taglicht, N., Shabtay, T., Ambus, A., Alon, N., Oron, S., Gordon-Tapiero, A., Kaplan, Y., Shwartz, V., Shaham, T. R., Riedl, C., Mirsky, R., Sap, M., Manheim, D., Ullman, T. & Bau, D. (2026). Agents of chaos. *arXiv*. https://doi.org/10.48550/arXiv.2602.20021

Stuyvenberg, A. (2026, January 24). *Clawdbot bought me a car.* https://aaronstuyvenberg.com/posts/clawd-bought-a-car

Suri, G., Slater, L. R., Ziaee, A., & Nguyen, M. (2023). Do large language models show decision heuristics similar to humans? A case study using GPT-3.5. *arXiv*. https://doi.org/10.48550/arXiv.2305.04400

Turing College. (2026). OpenClaw: The AI assistant that actually does things. *Turing College.* https://www.turingcollege.com/blog/openclaw

Xie, C., Chen, C., Jia, F., Ye, Z., Lai, S., Shu, K., Gu, J., Bibi, A., Hu, Z., Jurgens, D., Evans, J., Torr, P. H.S., Ghanem, B. & Li, G. (2024). Can large language model agents simulate human trust behavior?. *Advances in Neural Information Processing Systems, 37,* 15674-15729. https://dl.acm.org/doi/10.5555/3737916.3738417

Yelp Inc. (2026). *Yelp open dataset.* https://business.yelp.com/data/resources/open-dataset/

Zhu, Q., Chong, L., Yang, M., & Luo, J. (2025). Reading users' minds with large language models: Mental inference for artificial empathy in design. *Journal of Mechanical Design, 147*(6), 061401. https://doi.org/10.1115/1.4067527

Synthetic Participants: What Are They for in Business and How Can We Make Them Better?

PAOLA SCHIETEKAT, MARÍA LÓPEZ[1], AND JUAN DE RUS

Neovantas

DARIO KRPAN

London School of Economics and Political Science

Click or scan QR code
to watch video summary

Synthetic participants, commonly defined as artificially created profiles that simulate human responses, behaviours and decisions, are increasingly used as a complementary tool to traditional research methods. Despite their growing adoption, however, their capabilities, limitations and appropriate applications remain poorly systematised. This article develops a structured framework for understanding and deploying synthetic participants in business contexts. First, it introduces a functional classification based on primary objectives: declarative (intent-based), behavioural (action-based) and conversational (multi-agent) models. Second, it reviews key studies within each category, identifying common performance patterns and practical applications. Third, it proposes a differentiated methodological approach grounded in a quantitative-first perspective, integrating behavioural economics through structured libraries and tailoring models to specific use cases. Using a controlled experimental replication, the paper demonstrates that synthetic participants can approximate human data under structured conditions.

Introduction

The rapid proliferation of synthetic participants suggests a potential shift in how market research is conducted, offering clear advantages in scalability, speed and the ability to conduct research in contexts that would otherwise be inaccessible. Yet, despite their promise, the current landscape presents important challenges.

First, in the business context, we have heard of many approaches that rely heavily on the generation of synthetic participants from qualitative inputs, such as high-level buyer persona descriptions derived from small-scale interviews and grounded in descriptive, self-reported data. Furthermore, outputs are sometimes evaluated through subjective expert judgement rather than standardised quantitative metrics. For instance, Schramowski et al. (2022) assess GPT-3-based respondents by primarily using qualitative ratings of plausibility. This matters because decisions made based on plausible-looking but mis-calibrated outputs carry substantial business risk. Second, there is limited integration of psychological and behavioural economics constructs, which might be a significant limitation, as AI systems struggle to reproduce cognitive biases, heuristics and context-dependent preferences without explicit

1 Corresponding author: mlopez@neovantas.com

structuring. Aher et al. (2023) show that LLMs only exhibit human-like loss aversion and fairness when these mechanisms are embedded in the task design as behavioural dictionaries or knowledge bases.

Third, many synthetic panel solutions are built as generalised tools that favour ease of use over rigorous, bespoke, contextualised approaches. This makes their outputs difficult to distinguish from those obtained through ad hoc prompt-based generation with general purpose LLMs.

The significant and widely anticipated advantages, combined with these challenges and the limited number of publicly available business case studies outside academia, contribute to a mix of enthusiasm and a degree of scepticism among organisations considering their adoption.

What Types of Synthetic Participants Exist, and What Are Their Potential Business Applications?

This article aims to provide a simple classification of the types of synthetic participants that can be developed, along with the business areas in which they can be applied (see Table 1).

This classification may evolve as the field develops and new use cases emerge. For now, though, we believe

Table 1: Types of Synthetic Participants: Objectives and Business Applications

Type	Objective	Main Business Applications	Advantages	Challenges
Declarative (Intent-Based) Models	To simulate responses to structured or semi-structured survey instruments. Currently account for most published applications.	Market research, Customer Experience	Scalable survey prototyping, access to hard-to-reach populations, counterfactual scenario testing, hybrid approaches improve depth and efficiency.	Inflated effect sizes, compressed response distributions, high sensitivity to prompt phrasing, inherent intention-behaviour gap.
Behavioural (Action-Based) Models	To replicate specific reactions or behaviours in response to a stimulus, rather than stated intentions.	UX research, Pricing, Marketing	Can match or exceed human specialist performance on usability tasks, no fatigue, detailed interaction logs at scale, persona-aligned journey simulation.	Compressed variability, fails to replicate cognitive inertia or bounded attention, explicit behavioural scaffolding required for realistic behaviour.
Conversational (Multi-Agent) Models	To simulate conversations between two or more agents using a multi-agent architecture guided by scripts or structured instructions.	Sales, Customer Service	Enables testing of sales scripts and communication strategies at scale, allows study of opinion dynamics and information diffusion.	Risk of anthropomorphism, WEIRD-centric cultural bias, value lock-in, plausible outputs may mask mechanistically different cognition.

it offers a practical framework that helps researchers and practitioners grasp the diversity and distinct purposes of synthetic participants.

What Does the Literature Tell Us About the Performance of Synthetic Participants?

Declarative Synthetic Participants

Declarative or intent-based, synthetic participants are designed to provide survey-like responses. They can generate large volumes of responses quickly and support rapid prototyping of questionnaires and experimental designs before human deployment (Aher et al., 2023; Arora et al., 2025; Doudkin et al., 2025; Park et al., 2023; MRS, 2024). Beyond efficiency, these models help explore counterfactual scenarios and hard-to-reach populations (Aher et al., 2023; Lin & Dia, 2025).

A study by Arora et al. (2025), in which a human-LLM hybrid approach ultimately led to efficiency and effectiveness gains in quantitative and qualitative evaluations of a food company, demonstrates clear business applications. Results indicated that qualitative LLM-generated responses were superior in depth and insightfulness.

However, research consistently demonstrates that declarative models inflate effect sizes, with Fisher z values and standardised effects often two to three times larger than those observed in human samples, and produce narrower, smoother response distributions, underrepresenting population heterogeneity (Dillion et al., 2023; Doudkin et al., 2025; Maier et al., 2025; Stamatogiannakis et al., 2025; Ustiyanovych & Krpan, 2025). They are also highly sensitive to prompt phrasing, with small changes significantly altering reported attitudes or preferences (Gao et al., 2024; Tjuatja et al., 2024).

Additionally, although highly useful for accelerating market research, both private companies and public institutions should remain aware that any declarative research method, whether based on real humans or synthetic participants, will exhibit a gap relative to actual behaviour, as shown in Sheeran's meta-analysis, where intentions explain only around 28% of the variance in real-world behaviour (Sheeran, 2002). This is primarily because these methods rely on self-reports of people's views, attitudes and intentions (e.g., "I want to save money every month"), which can diverge from actual behaviour due to contextual or emotional factors, habits or shifts in priorities at the moment of decision-making (e.g., failing to save due to unexpected expenses).

"Any declarative research method, whether based on real humans or synthetic participants, will exhibit a gap relative to actual behaviour."

Behavioural Synthetic Participants

Behavioural – or action-based – synthetic participants focus on what agents do rather than what they say, simulating concrete interactions with interfaces, products or decision environments.

In UX research, they can execute long interaction sequences without fatigue, identify usability issues and provide detailed feedback at scale (Lu et al., 2025; Zhong et al., 2024). For example, Zhong et al. (2024) report that GPT-4, used as an expert evaluator of interface screenshots, detects 73–77% of usability problems versus 57–63% for human specialists, demonstrating that behavioural agents can match or exceed human performance in some metrics.

Behavioural synthetics have evolved beyond UX, though. A study by Mansour et al. (2025), for instance, consisted of mining historical e-commerce data to build Persona Aligned Agentic Retail Shoppers (PAARSs), used to simulate personalised journeys from product discovery all the way to purchase.

Nevertheless, limitations remain. Behavioural synthetics often confirm compressed variability, inflated treatment effects and stereotyped patterns, failing to capture real-world noise, cognitive inertia or bounded attention (Doudkin et al., 2025; Stamatogiannakis et al., 2025; Ustiyanovych & Krpan, 2025). Embedding psychological and behavioural economics constructs, such as heuristics, loss aversion, social norms or context-dependent preferences, is challenging, requiring explicit scaffolding to approximate realistic behaviour (Aher et al., 2023; Maier et al., 2025). Technical safeguards, such as semantic similarity ratings, retrieval-augmented generation and iterative calibration against human benchmarks, can help mitigate bias and overconfident inferences (Cui et al., 2024; Maier et al., 2025; Suh et al., 2025).

Conversational Synthetic Participants

Conversational models simulate dialogues between agents. They are particularly useful for testing sales scripts, customer service interactions and communication strategies (Lu et al., 2025), but they also enable the study of opinion dynamics and information diffusion, as shown in Park et al. (2023) and Qu and Wang (2024).

However, these models raise important epistemological concerns, as they can produce highly plausible conversations while relying on fundamentally different mechanisms from human cognition (Shanahan et al., 2023). This creates a risk of anthropomorphism, where outputs are misinterpreted as genuine human-like understanding (Mitchell & Krakauer, 2023; Vallor, 2024).

Additionally, conversational models are particularly vulnerable to cultural bias and value lock-in. They often reproduce WEIRD-centric perspectives, leading to systematic misrepresentation of non-Western or underrepresented groups (Agnew et al., 2024; Amini, 2025; Santurkar et al., 2023).

Our Approach to Synthetic Participants

Our Vision

When our team at Neovantas began exploring synthetic participants, we quickly became convinced that integrating them as a tool into companies' research and experimentation methodologies could provide a clear, short-term competitive advantage.

As we started working on real projects, it soon became apparent that a 'one-size-fits-all' approach would likely be suboptimal. A declarative synthetic participant model, designed to answer surveys, is fundamentally different from a behavioural model, designed to simulate concrete actions in terms of configuration, prompting, input and output. Furthermore, even within these categories, the business objective, whether UX, pricing or debt collection, requires entirely different approaches. In practice, we are developing specific, specialised synthetic participant models, each with its own logic, tailored to each use case according to the business need. Moreover, we apply a quantitative approach, both in preparing the model's input and in validating the results.

In addition, we hypothesised that enriching synthetic participants' input with a context drawn from behavioural economics and psychology, through structured knowledge libraries, would help synthetic participants' responses and behaviours more closely resemble human reality. This would allow for better output calibration and enrich the reasoning behind the participants' answers or decisions.

"We are developing specific, specialised synthetic participant models, each with its own logic, tailored to each use case according to the business need."

Putting It to the Test

To test this hypothesis, we conducted the same exercise with declarative synthetic participants, comparing results with and without this enriched input from the libraries. The original exercise we selected was a randomised controlled trial conducted in 2017 with 199 Spanish working adults aged 35 to 65, who were surveyed about their willingness to save (WTS) for retirement and their impulsiveness. The study aimed to evaluate whether exposing a treatment group to a pension calculator would increase their WTS by making the benefits of saving more tangible. The intervention led to a statistically significant WTS increase of 25% ($p = 0.044$).

Our goal was to see whether exposing 199 synthetic participants, matched to the original population in sociodemographic characteristics such as age and gender, and financial attributes such as income and financial literacy, to the same pension calculator stimulus would produce similar results, with WTS as the outcome. We also aimed to assess whether enriching the input with structured knowledge libraries on the most relevant behavioural economics and psychology principles for saving decisions would bring the outcomes even closer to the original results.

To achieve this, we created two libraries:

Behavioural library. This dictionary described each bias, its mechanism and manifestation in the context of WTS, its interaction across sociodemographic variables and under a saving saliency treatment. Interaction rules guided the direction and intensity of activation. Each bias was drawn from research on behavioural economics and financial behaviour.

Impulsiveness library. Impulsiveness scores were generated for each synthetic participant using the

Barratt Impulsiveness Scale, BIS-11, a 30-item instrument. Responses were sampled from the empirical distributions observed in the original dataset, and total scores were computed accordingly.

In this case, both libraries were created as JSON files and retrieved by the model through prompting.

The 'basic' solution, made up of synthetic participants generated without a behavioural knowledge base, produced the weakest distributional fit. KS statistics[2] in this condition ranged from 0.82 to 0.83, and real distributions diverged substantially across nearly the entire outcome range. This pattern is consistent with the literature, in that 'one-size-fits-all' generative models tend to produce WTS estimates that are both more uniform and more concentrated around the centre of the scale than the real data, thus underrepresenting the natural spread of the human distribution.

"Introducing a behavioural library produced a meaningful improvement representing a reduction in distributional divergence of approximately 15% relative to the baseline."

Introducing a behavioural library produced a meaningful improvement. The GPT-5.4-mini model achieved a KS statistic of 0.70 in the library-only

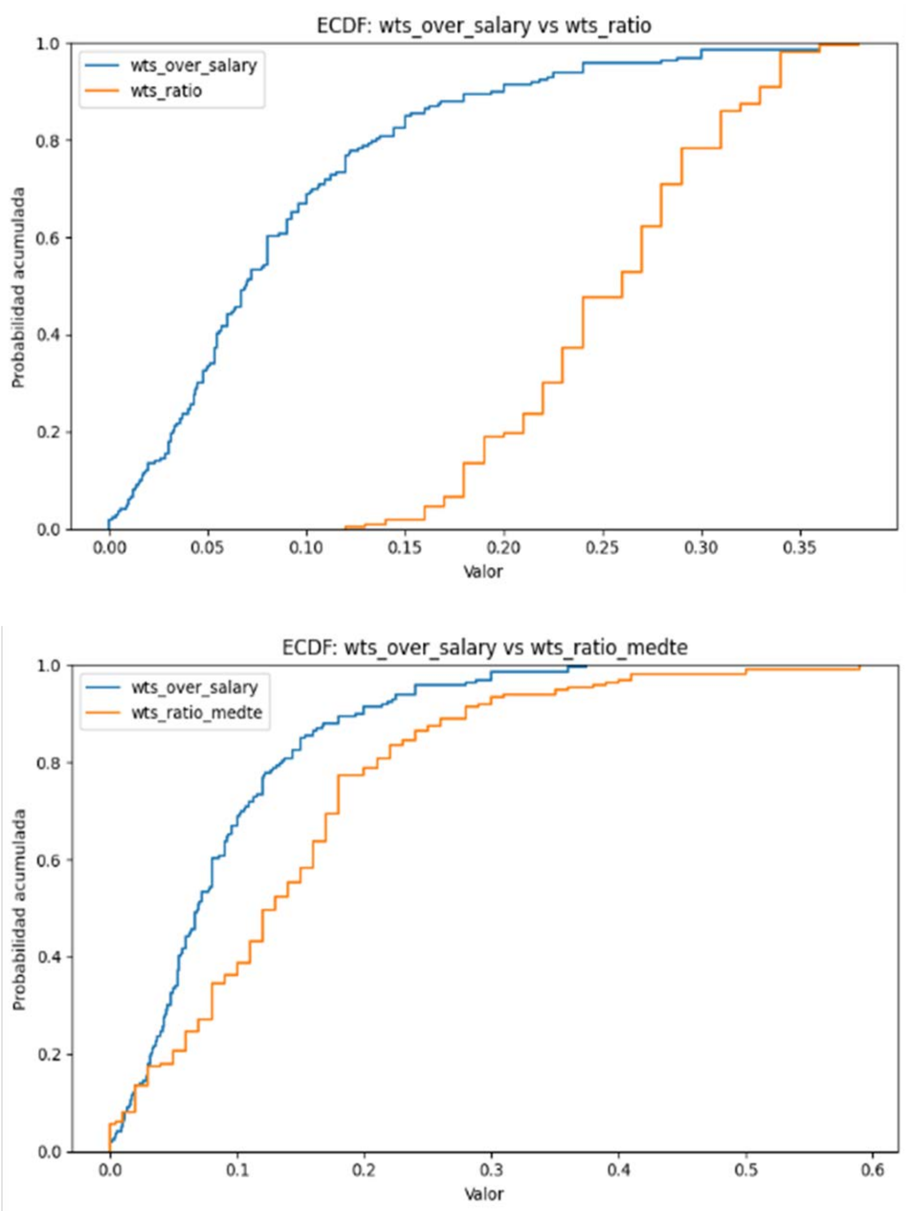


Figure 1: Comparison between synthetic models generated without dictionaries (top) and with behavioural and impulsiveness dictionaries (bottom).

2 The Kolmogorov-Smirnov (KS) statistic is a measure of the biggest difference between two datasets.

condition, representing a reduction in distributional divergence of approximately 15% relative to the baseline (even though it still indicates meaningful residual divergence). Including the Impulsiveness library yielded a KS statistic of 0.54 with the standard prompt configuration, reflecting a substantive gain (see Figure 1).

Despite these findings, a systematic tendency to overestimate WTS persisted, particularly due to occasional high-value outliers that distorted the upper tail of the distribution. This suggests an inherent fault within generative models, i.e., they are unable to naturally reproduce the inertia that characterises real savings behaviour.

As synthetic users are still in an embryonic stage, exercises grounded in known human distributions are essential to assess fidelity. For transparency, the original experimental results were not used as model inputs or prompts, thereby avoiding potential bias. Instead, the behavioural libraries were constructed from a comprehensive review of the academic literature, modelling the effects of stimuli such as the calculator on WTS.

> “The behavioural libraries were constructed from a comprehensive review of the academic literature, modelling the effects of stimuli such as the calculator on WTS.”

While this approach is innovative, it requires further validation. We are therefore pursuing two next steps: replicating the study with a new sample and synthetic population to test consistency, and using known results to develop guidelines for building reliable synthetic participants when human reference data is unavailable.

Conclusion

These results support a differentiated and modular approach to building synthetic participants by separating the sociodemographic buildup (who the participant is) from the behavioural library (what cognitive principles are likely active, and how strongly). Many solutions offer quick and adaptable synthetic models that include sociodemographic data but overlook behavioural profiling and contextualisation. The limitations of this approach have been widely discussed in the literature, even before the development of AI models; demographic profiling and segmentation without a behavioural layer often paints an incomplete picture of the subjects of study, especially when gauging a behavioural outcome.

Synthetic participants are best suited to contexts where outputs will be used to inform rather than replace human judgement. They are most defensible as an augmentation tool useful for hypothesis generation, rapid prototyping and scenario testing, and they are least suited to situations involving high emotional complexity, culturally specific behaviour or underrepresented populations. Embedding behavioural economics constructs is feasible where the relevant decision environment is sufficiently understood, whilst biases can be operationalised as structured knowledge libraries that modulate how synthetic participants respond across profiles and conditions – as our experiment demonstrates. The same logic extends to adaptation over time: as economic conditions, social norms and digital behaviours evolve, static libraries will drift from reality. Periodic recalibration against human reference data is therefore not an optional refinement but a structural requirement.

Synthetic participants are increasingly useful tools for specific, well-defined tasks when their construction is grounded in empirical data, explicit psychological theory and transparent calibration procedures. The framework proposed in this article is offered as a starting point for that kind of responsible deployment: one that treats the limitations of synthetic participants as design parameters to be managed rather than as obstacles to be minimised.

> “Embedding behavioural economics constructs is feasible where the relevant decision environment is sufficiently understood, whilst biases can be operationalised as structured knowledge libraries that modulate how synthetic participants respond.”

Three practical implications stand out for practitioners. First, quantitative calibration against human benchmark distributions should be treated as an essential check, as distributional fit metrics make deviations visible before they influence decisions. Second, anchoring synthetic outputs to empirically grounded knowledge bases, through structured libraries or retrieval-augmented approaches, can reduce the idealisation bias that unconstrained generative models exhibit. Third, the question of what synthetic participants are actually good for should remain at the centre of deployment decisions. Used within aforementioned boundaries, they are a powerful addition to the behavioural researcher's toolkit; used beyond them, they carry risks that are well-documented and avoidable.

THE AUTHORS

María López is a behavioural economics manager and Head of Innovation at Neovantas. She holds a University Expert in Behavioral Economics from Behavior & Law and UDIMA, Spain. At Neovantas, María leads international projects focused on creating specialised synthetic participant models enriched with behavioural input. She also works on designing new products that integrate AI with behavioural science insights.

Paola Schietekat is a senior behavioural economics consultant at Neovantas. She holds a master's degree in Public Policy from Oxford University, where she specialised in Behavioural Science and Economic Development. At Neovantas, she works on integrating behavioural science insights into AI solutions.

Juan de Rus is the managing partner at Neovantas. He holds an MSc in Behavioural Science with Distinction from the London School of Economics & Political Science (LSE), an MBA from Universidad Carlos III de Madrid and a Master of Arts with Distinction from the University of Dundee (UK). He also serves as an adjunct professor of Behavioural Science and Marketing at IE Business School, Universidad Carlos III and the University of Barcelona.

Dario Krpan is Associate Professor of Behavioural Science at the London School of Economics and Political Science. His research examines the transformative power of behavioural science in the age of AI and new technology. In this context, he explores the intersection of behavioural science and AI, with a particular emphasis on synthetic participants as tools for simulating human behaviour and decision-making. Dario is also a senior advisor to Neovantas, where he contributes to shaping AI and data-driven strategies aimed at transforming organisations through the use of behavioural data.

REFERENCES

Agnew, W., Bergman, A. S., Chien, J., Díaz, M., El-Sayed, S., Pittman, J., Mohamed, S., & McKee, K. R. (2024). The illusion of artificial inclusion. *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24), 286*. https://doi.org/10.1145/3613904.3642703

Aher, G., Arriaga, R. I., & Kalai, A. T. (2023). Using large language models to simulate multiple humans and replicate human subject studies. *Proceedings of the 40th International Conference on Machine Learning, 202*, 337-371. https://proceedings.mlr.press/v202/aher23a.html

Amini, A. (2025). Survey transfer learning: Recycling data with silicon responses. *arXiv*. https://arxiv.org/abs/2501.06577

Arora, N., Chakraborty, I., & Nishimura, Y. (2025). AI–human hybrids for marketing research: Leveraging large language models as collaborators. *Journal of Marketing, 89*(2), 43-70. https://doi.org/10.1177/00222429241276529

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), 337-355. https://doi.org/10.1017/pan.2023.2

Bisbee, J., Clinton, J. D., Dorff, C., Kenkel, B., & Larson, J. M. (2024). Synthetic replacements for human survey data? The perils of large language models. *Political Analysis, 32*(4), 401-416. https://doi.org/10.1017/pan.2024.5

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science, 356*(6334), 183-186. https://doi.org/10.1126/science.aal4230

Cui, Z., Li, N., & Zhou, H. (2024). Can large language models replace human subjects? A large-scale replication of scenario-based experiments. *arXiv*. https://arxiv.org/abs/2409.00128

Dillion, D., Tandon, N., Gu, Y., & Gray, K. (2023). Can AI language models replace human participants? *Trends in Cognitive Sciences, 27*(7), 597-600. https://doi.org/10.1016/j.tics.2023.04.008

Doudkin, A., Pataranutaporn, P., & Maes, P. (2025a). AI persuading AI vs. AI persuading humans: Differential effectiveness in promoting pro-environmental behavior. *arXiv*. https://arxiv.org/abs/2503.02067

Doudkin, A., Pataranutaporn, P., & Maes, P. (2025b). From synthetic to human: The gap between AI-predicted and actual behavior change after chatbot persuasion. *Proceedings of the ACM Conference on Conversational User Interfaces* (CUI '25), 71. https://doi.org/10.1145/3719160.3736610

Gao, Y., Lee, D., Burtch, G., & Fazelpour, S. (2024). Take caution in using LLMs as human surrogates: Scylla ex machina. *arXiv*. https://arxiv.org/abs/2410.19599

González-Bustamante, B., Verelst, N., & Cisternas, C. (2025). Emulating public opinion: Synthetic survey responses for the Chilean case. *arXiv*. https://arxiv.org/abs/2509.09871

Haider, S. A., Borna, S., Trabilsy, M., Bagaria, S., & Forte, A. J. (2025). Synthetic patient–physician conversations simulated by large language models. *Sensors, 25*(14), 4305. https://doi.org/10.3390/s25144305

Lin, Z. (2025). Six fallacies in substituting large language models for human participants. *Advances in Methods and Practices in Psychological Science, 8*(3), 1-19. https://doi.org/10.1177/25152459251357566

Lin, Z., & Dai, Y. (2025). Fostering epistemic insights into AI ethics through a constructionist pedagogy: An interdisciplinary approach to AI literacy. *Proceedings of the AAAI Conference on Artificial Intelligence*, *39*. 29171-29177. https://doi.org/10.1609/aaai.v39i28.35190

Lu, Y., Yao, B., Gu, H., Huang, J., Wang, J., Li, Y., Gesi, J., He, Q., Li, T. J.-J., & Wang, D. (2025). UXAgent: An LLM agent-based usability testing framework. *arXiv*. https://arxiv.org/abs/2502.12561

Maier, B. F., Aslak, U., Fletcher, K., Rismal, N., Dow, R., Pappas, K., Fiaschi, L., Luhmann, C. C., & Wiecki, T. V. (2025). LLMs reproduce human purchase intent via semantic similarity elicitation. *arXiv*. https://arxiv.org/abs/2510.08338

Mansour, S., Perelli, L., Mainetti, L., Davidson, G., & D'Amato, S. (2025). PAARS: Persona-aligned agentic retail shoppers. *arXiv*. https://arxiv.org/abs/2503.24228

Market Research Society Delphi Group. (2024). *Using synthetic participants for market research: Part two of the BEST framework for generative AI*. MRS Evidence Matters. https://www.mrs.org.uk/pdf/MRS_Delphi_synthetic.pdf

Mitchell, M., & Krakauer, D. C. (2023). The debate over understanding in AI's large language models. *Proceedings of the National Academy of Sciences, 120*(13), e2215907120. https://doi.org/10.1073/pnas.2215907120

O'Grady, C. (2024). AI-generated "participants" can lead social science experiments astray. *Science, 383*(6674), 830-832. https://www.science.org/doi/10.1126/science.aec9020

Park, J. S., O'Brien, J. C., Cai, C. J., Morris, M. R., Liang, P., & Bernstein, M. S. (2023). Generative agents: Interactive simulacra of human behavior. *arXiv*. https://arxiv.org/abs/2304.03442

Qu, Y., & Wang, J. (2024). Performance and biases of large language models in public opinion simulation. *Humanities and Social Sciences Communications, 11*, 1095. https://doi.org/10.1057/s41599-024-03609-x

Raisch, S., & 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

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, 202*, 29971-30004. https://proceedings.mlr.press/v202/santurkar23a.html

Shanahan, M., McDonell, K., & Reynolds, L. (2023). Role play with large language models. *Nature, 623*(7987), 493-498. https://doi.org/10.1038/s41586-023-06647-8

Sheeran, P. (2002). Intention-behavior relations: A conceptual and empirical review. *European Review of Social Psychology, 12*(1), 1-36. https://doi.org/10.1080/14792772143000003

Shrestha, P., Krpan, D., Koaik, F., Schnider, R., Sayess, D., & Binbaz, M. S. (2025). Beyond WEIRD: Can synthetic participants substitute for humans in global policy research? *Behavioral Science & Policy, 10*(2), 26-45.

Stamatogiannakis, A., Ghodsinia, A., Etminanrad, S., Gonçalves, D., & Santos, D. (2025). How human is the machine? Evidence from large-scale LLM conversations. *arXiv*. https://arxiv.org/abs/2510.07321

Suh, J., Jahanparast, E., Moon, S., Kang, M., & Chang, S. (2025). Fine-tuning language models on survey data for public opinion prediction. *arXiv*. https://arxiv.org/abs/2502.16761

Tjuatja, L., Chen, V., Wu, T., Talwalkar, A., & Neubig, G. (2024). Do LLMs exhibit human-like response biases? *Transactions of the Association for Computational Linguistics, 12*, 1011-1026. https://doi.org/10.1162/tacl_a_00685

Ustiyanovych, J., & Krpan, D. (2025). The potential of synthetic twin agents for personalized behavioural interventions at scale. https://doi.org/10.31234/osf.io/vs2mk_v5

Wang, Z., Lu, Y., Li, W., Sun, B., Bart, Y., Lyu, W., Gesi, J., Wang, T., Huang, J., Su, Y., Ehsan, U., Alikhani, M., Li, T. J.-J., Chilton, L., & Wang, D. (2025). See, think, act: Simulating online shopper behavior with multimodal agents. *arXiv*. https://arxiv.org/abs/2501.06577

Zhong, R., McDonald, D. W., & Hsieh, G. (2024). Synthetic heuristic evaluation: A comparison between AI- and human-powered usability evaluation. *arXiv*. https://doi.org/10.48550/arXiv.2507.02306

Choice Engine: An LLM and AHP-Powered Behavioral Science Tool

PABLO COELLO, PABLO VILLACAÑAS, GONZALO DÍAZ, SARA RAMOS-CABO[1], ANTONIO GONZÁLEZ, FRANCISCO CEBALLOS, AND JUAN PABLO OSPINA ARBOLEDA

BeWay

Click or scan QR code
to watch video summary

Designing cost-effective behavioral interventions in industry depends not only on what the literature says, but also on understanding the business context in which the intervention must work. In practice, this means that behavioral scientists often need to review large, heterogeneous sets of quantitative and qualitative evidence under tight timelines. Herein, we present Choice Engine, a structured workflow that combines LLMs, COM-B, Game Theory, and the Analytic Hierarchy Process (AHP) to support both diagnosis and intervention design. Across seven steps, the engine ingests complex information, organizes it into explicit decision inputs, predicts the decision an individual is likely to make in a predefined scenario, and explains why that decision emerges. Its predictive validity is illustrated by replicating the findings of two seminal studies, with 17 of 18 confirmatory tests reaching significance.

Introduction

Imagine a behavioral scientist at a financial institution being asked to increase the uptake of savings accounts among clients. The brief sounds simple enough. Then the real work starts. Does the client meet the legal requirements to open the account? Is the product available at that moment? Is the client interested in it? Does she even know the bank offers it? And this is before anyone starts discussing intervention ideas, compliance checks, costs, technical feasibility, user trust, or brand considerations.

If you have worked on applied behavioral research in industry, this pattern will likely feel familiar. Understanding and describing human behavior is a core part of the job, but any target behavior is shaped by a dense mix of personal and contextual factors. The diagnostic phase is where those factors are identified, and good intervention design depends on getting that diagnosis right. The problem is that diagnosis rarely begins with a neat dataset. More often, it means working through prior experiment readouts, ethnographic notes, funnel metrics, interview transcripts, customer support logs, scientific publications, etc.

To make matters worse, findings from carefully controlled studies do not automatically transfer to messy real-world environments. Going back to the savings example, the literature may suggest that financial literacy is central to savings behavior, yet a closer look at local regulation, product features, cultural norms, and client interviews may point somewhere else. Meanwhile, business contexts change quickly. Market shifts do not wait for a second research cycle.

1 Corresponding author: sararamos@beway.com

All of this leaves applied behavioral scientists facing two recurring problems: (1) too much mixed evidence to review, organize, and interpret and (2) the need to design an intervention that works for a specific population and context, not just in theory.

Choice Engine is designed to help with both. It condenses research and documentation tasks by ingesting and structuring large volumes of information, and it predicts context-specific decision-making while providing a causal narrative for the final choice. This combination is useful not only because it anticipates what a target individual is likely to do, but also because it makes visible the criteria and mechanisms behind that decision. In turn, this gives behavioral scientists a better starting point for designing more tailored—and potentially more effective—interventions.

> "Choice Engine condenses research and documentation tasks by ingesting and structuring large volumes of information, and it predicts context-specific decision-making while providing a causal narrative."

Methodological pillars

Choice Engine rests on four methodological pillars.

Large Language Models (LLMs) handle the scale and heterogeneity of the evidence. However, unstructured LLM use in research workflows generates familiar risks: hallucinations, overconfident summaries, and output inconsistency across prompts (Huang et al., 2025; Sclar et al., 2024) fueling a paradigm shift in information acquisition. Nevertheless, LLMs are prone to hallucination, generating plausible yet nonfactual content. This phenomenon raises significant concerns over the reliability of LLMs in real-world information retrieval (IR. Rather than relying on isolated calls with hard-to-inspect outputs, Choice Engine distributes reasoning across connected steps in which each output becomes the explicit input to the next (Khot et al., 2023). Every inference can therefore be inspected, challenged, and updated independently.

COM-B (Michie et al., 2011; West & Michie, 2020) is the Swiss army knife of behavior analysis. It is a theoretical framework that states that the target individual must have the Capability, Opportunity, and Motivation to carry out a Behavior (hence, the COM-B acronym). In the context of Choice Engine, COM-B helps determine whether the minimum conditions for a particular strategy are in place. Capability and Opportunity are treated as feasibility conditions for a given strategy, while Motivation is modeled later through multi-criteria evaluation.

We complement COM-B with Game Theory, specifically concepts from Information Economics, which makes it possible to assess whether the individual is aware of the available strategies and their likely consequences. This addition is important because even a feasible strategy will not be chosen if the individual does not know it is available or does not understand what it implies.

Finally, the Analytic Hierarchy Process (AHP, (Saaty, 1980, 2008) models motivation as a structured trade-off among Benefits, Opportunities, Costs, and Risks. Together, these four components (LLMs, COM-B, Game Theory, and AHP) turn a broad and messy decision context into an explicit, inspectable decision workflow.

> "Together, these four components (LLMs, COM-B, Game Theory, and AHP) turn a broad and messy decision context into an explicit, inspectable decision workflow."

Replicated Studies

To assess Choice Engine's predictive validity, we selected two seminal studies that cover two classic families of strategic choice. The first involves decisions in which individuals use a behavior to reveal an otherwise hidden personal characteristic to others—what economists call *signaling*. The second involves decisions under risk, in which individuals weigh certain and uncertain outcomes before choosing.

These studies are especially useful as benchmarks because they are analytically simple yet comprehensive. Furthermore, both provide precise descriptions of profiles, available strategies, and contextual parameters, which makes it possible for Choice Engine to ingest well-specified inputs and generate predictions that can be compared with the original results, without the usual ambiguity that comes from underspecified real-world cases.

In terms of Spence's job-market signaling model (Spence, 1973), think of two job candidates who differ in their actual productivity, but employers cannot tell them apart just by looking. Getting a degree serves as proof of capability, but it costs more effort for some people than others, and because of this cost difference, the two types naturally end up choosing different education levels (i.e., separating equilibrium), making themselves distinguishable to employers. We test three scenarios. The baseline reproduces this outcome, where the two types make different choices. The subsidy scenario makes education cheaper for everyone, and as a result, both types pursue more education and become harder to tell apart (i.e., pooling)—a pattern also observed when real-world subsidies boost enrollment (Bedard, 2001). The credential fraud scenario works in reverse, in that when degrees can be faked, they no longer prove much, so the motivation to study drops for everyone (Hvide, 2003). Both types then settle on less education.

Kahneman and Tversky's Prospect Theory (Kahneman & Tversky, 1979) foundational paper presents 16 decision problems and shows that people systematically deviate from the predictions of classical expected utility theory when facing risky choices. We test three scenarios. The first two replicate Problem 3 of the original paper, where individuals choose between a sure outcome and a probabilistic one. In the gains version, individuals tend toward risk aversion, preferring a sure gain over a larger but uncertain one. In the losses version, with equivalent magnitudes, they tend toward risk-seeking, preferring the gamble over a sure loss. This reversal illustrates the *certainty effect* and the *reflection effect*. The third scenario operationalizes the broader idea of *loss aversion*, using a mixed gamble in which a potential gain and a potential loss are simultaneously in play.

Together, these scenarios complement the signaling family: where the Spence cases test type-conditional choices under strategic interaction, the Prospect Theory cases test how individuals weigh certainty, downside risk, and the asymmetric valuation of gains and losses.

Choice Engine Workflow

Choice Engine follows a 7-step process. Step 1 requires human validation. Steps 2 through 7 are automated. The key design principle is simple: the output of each step becomes the explicit input to the next.

> "The key design principle is simple: the output of each step becomes the explicit input to the next."

Put differently, Step 1 defines the playing field. The remaining steps map the strategic landscape, ground each profile's perspective, pin down the numbers, evaluate trade-offs, rank strategies, and present the result in a format that can be inspected and challenged.

Step 1: Input

All relevant data is collected through Retrieval Augmented Generation (RAG) and organized into three structured templates: *Profiles*, *Strategies*, and *Scenarios*. Profiles describe the target individuals, including their characteristics, internal constraints, and private information. Strategies define the available courses of action. Scenarios describe the wider context in which decisions are made, including cultural, economic, and regulatory conditions. Public information is separated from private or unknown

Table 1: Prospect Theory Scenarios, Types and Options

Scenario	Type	Option A	Option B / D
PT1	Gains	Gain USD 3,000 for sure	80% chance to gain USD 4,000; 20% chance of USD 0
PT2	Losses	Lose USD 3,000 for sure	80% chance to lose USD 4,000; 20% chance of USD 0
PT3	Mixed gamble	50% chance to gain USD 2,000; 50% chance to lose USD 1,000	USD 0 for sure (decline)

Note: PT = Prospect Theory

Figure 1: Choice Engine's 7-step workflow.

information at the time of choice, which is essential for preventing leakage into later predictions. A set of neutrality rules helps ensure that the templates do not inadvertently favor one strategy over another, and an LLM judge verifies compliance. The behavioral scientist reviews and validates this input before the automated pipeline begins.

The information on Profiles, Scenarios, and Strategies is compiled into different templates: two documents with the worker profile descriptions (i.e., high and low productivity), three scenario descriptions (i.e., base, fraud, and subsidy) and five education levels (ranging from 0 to 4). The result is a decomposition of the public (e.g., conventions) and the private profile-specific knowledge required to build a clear signaling context.

Step 2: Strategic Brief

Using only Scenario and Strategy information, and no Profile data yet, the engine maps the decision environment: who the actors are, what is publicly observable, how others are expected to respond, and which informational gaps remain. In other words, this step sketches the 'physics of the world' before any profile enters it. The step is grounded mainly in Game Theory and Information Economics. If critical information is missing, the engine produces its best single prediction for that gap and logs the assumptions it had to make.

In the Spence case, the Strategic Brief establishes that employers observe credentials but not productivity, that wages follow a public convention, and that credentials at or above a given cutoff yield higher wage categories. The output is a compact, profile-agnostic map of actors, information structure, expected reactions, and key uncertainties.

Step 3: Agent Grounding

This step translates the Strategic Brief into a profile-specific point of view. For each Profile, Choice Engine combines the Scenario with the individual's characteristics to determine what the individual can do, what is actually available in the Scenario, what the individual knows at the time of deciding, and which outcomes each Strategy is expected to generate. The aim is not yet to predict the final decision but to delimit the feasible and relevant options for each Profile. This is where COM-B does most of its work: Capability and Opportunity act as hard feasibility conditions, while personal characteristics, motivation-relevant factors, and information access determine how the same Scenario is experienced differently by different Profiles.

In the Spence replication, both profiles face the same credential options and the same wage convention, but they differ in their private education costs. The same credential level therefore produces a lower burden for the high-productivity profile than for the low-productivity one. That asymmetry is what later drives separation.

Step 4: Binding & Netting

This is the point where Choice Engine stops guessing when direct calculation is possible. The purpose of the step is to avoid asking the LLM to infer values that can be obtained deterministically. *Binding* applies the correct profile-specific adjustment to obtain each individual's actual values. *Netting* then collapses gains and losses that share the same units into single quantities. This reduces arithmetic noise, prevents double counting, and ensures that later evaluations operate on clean, comparable data.

In the Spence case, the same credential yields the same wage for both types, but with different education costs. Binding and netting therefore produce different profile-specific values before motivation is assessed.

Step 5: BOCR Evaluation

At this point, the engine has two key ingredients: the net consequences of each Strategy and the grounded, profile-specific conditions under which those consequences are evaluated. Step 5 turns those

Table 2: Benefits, Opportunities, Costs, and Risks Evaluation

	Positive	Negative
Certain payoffs	Benefits	Costs
Uncertain payoffs	Opportunities	Risks

ingredients into a BOCR analysis of each Strategy's Benefits, Opportunities, Costs, and Risks.

This separation is useful because not only classifies outcomes, but it also prepares the ground for later simulation. Risk aversion and loss aversion, for example, can be modeled as different ways of weighting these four dimensions. In the baseline Spence scenario, the classification is straightforward because there is no uncertainty: wage offers are coded as Benefits and private education burdens as Costs, while Opportunities and Risks remain inactive.

Step 6: AHP

The BOCR narratives are converted into a formal preference ranking through Saaty's pairwise comparison method (Saaty, 1980, 2008). The process unfolds in three parts. First, the decision hierarchy is defined by specifying the goal, the available alternatives, and the four BOCR criteria. Second, these criteria are compared pairwise to assess their relative importance for a given Profile, and the alternatives are then compared with respect to each criterion. Third, those comparisons are aggregated into an overall priority score for each alternative.

A key diagnostic in this step is the *Consistency Ratio* (CR), which measures whether the pairwise comparisons inside the AHP matrices are internally coherent. In standard AHP practice, a CR below 0.10 indicates acceptably consistent judgments (Saaty, 2005), which is especially useful when some comparisons are generated by an LLM, because it offers a formal and transparent quality check. CR does not guarantee correctness, but it does help identify unstable or noisy judgments. Across our replications, alternative-level comparisons consistently yielded CRs of zero or near zero, which is expected because those comparisons rely directly on quantitative data. Criteria-level comparisons showed more variability, particularly in the signaling scenarios where the LLM must assess the relative importance of BOCR dimensions, but most remained below the 0.10 threshold.

Step 7: Simulation Panel

The final step integrates everything into an interactive decision artifact that is both predictive and interpretable. The panel is organized around four layers of information, each one designed to let the analyst progressively unpack the decision.

First, an LLM-generated summary presents the case in plain language: who the profiles are, what strategies are available, and how the scenario is defined. Right below it, a *Strategy Feasibility Matrix* illustrates whether each Profile has the Capability and Opportunity to pursue each Strategy. These two flags act as hard constraints, so infeasible strategies are excluded from the ranking no matter how attractive they might otherwise look.

Second, the core of the panel displays the final preference ranking. A bar chart shows each Profile's global AHP scores, and error bars reveal how stable those scores are across simulated pairs. In the Spence illustration, this makes the separating pattern easy to see: the high-productivity type concentrates weight on higher credential levels, while the low-productivity type concentrates on lower ones.

Third, the panel includes sensitivity controls. A *risk aversion* slider amplifies Benefits and Costs relative to Opportunities and Risks. A *loss aversion* slider amplifies Risks and Costs relative to Benefits and Opportunities. Because every change triggers a new computation, the analyst can quickly see whether the predicted ranking is robust or whether small motivational shifts would invert preferences.

Fourth, an expanded detail view decomposes each strategy's global score into its four BOCR components and pairs that decomposition with a narrative rationale. This lets the analyst inspect not only which strategy wins, but also why it wins. Every number in the final ranking can be traced back to the scenario definition, the profile, the bound outcomes, and the pairwise comparisons that produced it. The result is not a black-box prediction but a fully auditable decision record.

That transparency is what makes the output practically useful. In the Spence case, the analyst can see why the high-productivity type prefers higher credentials and why the low-productivity type does not, and they can then stress-test those conclusions under alternative behavioral profiles.

Replication Results

To provide a compact but informative empirical test of the workflow, we deployed the full pipeline on six benchmark scenarios drawn from the two classical studies described above: three signaling scenarios based on Spence and three lottery scenarios based on Kahneman and Tversky. The deployment used `gpt-5.2-2025-12-11` with temperature set to 0.5 and was executed as six scenario-specific batches. Within each pair, the high-type and low-type agents shared the same model configuration and seed, so the only manipulated factor was the agent profile under a fixed narrative environment. Each run followed the same seven-step flow: Input, Strategic Brief, Agent Grounding, Binding & Netting, BOCR Evaluation, AHP, and Simulation Panel.

“The results reveal that the engine recovers the theoretical structure of the signaling environments with complete accuracy.”

The results reveal that the engine recovers the theoretical structure of the signaling environments with complete accuracy. In the baseline Spence treatment, all 50 pairs reproduced the expected separating outcome: the high-productivity type chose a high credential and the low-productivity type chose a low one. Under full subsidy, all 50 pairs pooled on high education. Under credential fraud, all 50 pooled on low education. In other words, the predicted pair pattern was recovered in 100% of cases across all three signaling scenarios, and all confirmatory tests rejected the preregistered null benchmark.

Performance remained strong in the lottery environments, although the losses condition showed more dispersion. In the gains scenario, 48 of 50 pairs matched the predicted pattern, corresponding to 96% pair-level accuracy. In the mixed-gamble scenario, 47 of 50 pairs matched the predicted pattern, corresponding to 94% accuracy. The losses scenario was the only case in which the stricter pair-level test did not reach significance: 39 of 50 pairs matched the predicted joint pattern, or 78%. Importantly, this shortfall reflects the difficulty of recovering the joint configuration rather than a failure at the individual level. This difficulty lies in the fact that Choice Engine assesses possible strategies and motivational factors separately, which in the PT2 scenario creates a double standard; specifically, the agent is asked to follow this logic: “When facing a sure loss, you may be willing to take risks to potentially avoid that loss”. This results in individuals’ motivations changing based on the available options. In any case, the engine still selected the theoretically expected option in 86% of runs for the high type and 92% for the low type. Across the full analysis, 17 of the 18 confirmatory tests were significant.

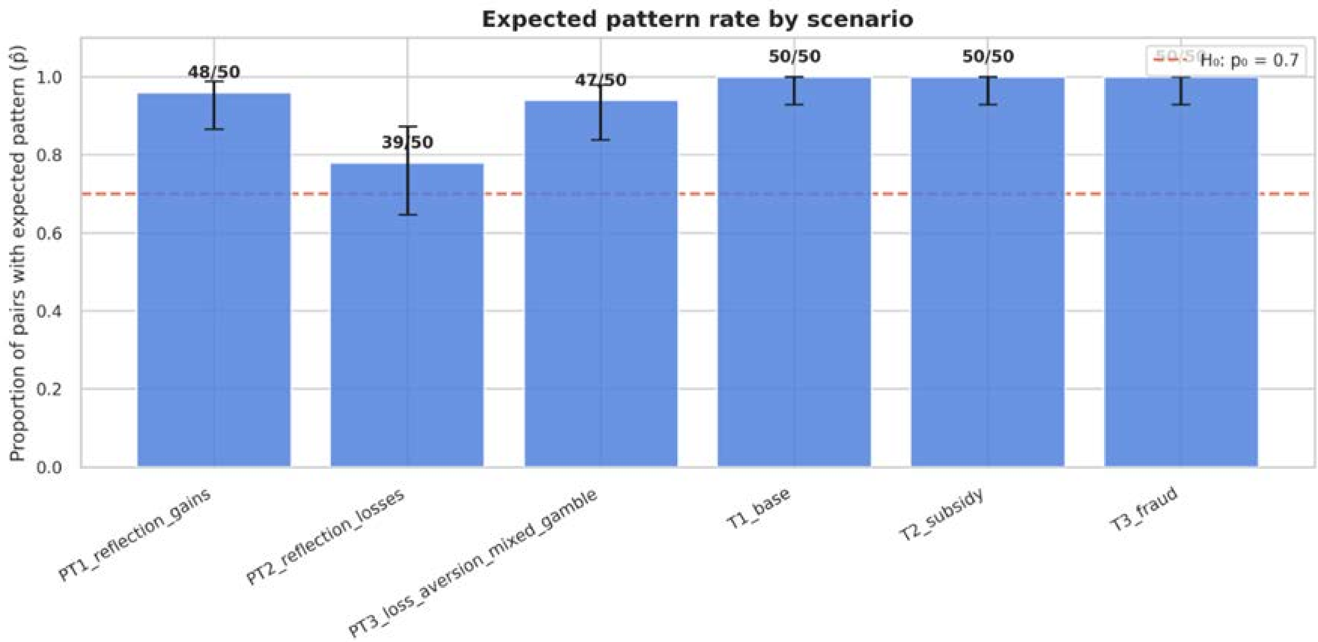


Figure 2: Expected pattern rate by scenario.

Table 3: Binomial Test Results

	Test	Scenario	Type	n	Hits	p̂	CI lower	CI upper	Cohen h	p
1	Pair pattern	PT1 Reflection gains	Both (H,L)	50	48	0.9600	0.8794	1.0000	0.7566	< .001
2	Pair pattern	PT2 Reflection losses	Both (H,L)	50	39	0.7800	0.6622	1.0000	0.1829	.139
3	Pair pattern	PT3 Loss aversion mixed gamble	Both (H,L)	50	47	0.9400	0.8522	1.0000	0.6643	< .001
4	Pair pattern	T1 Base	Both (H,L)	50	50	1.0000	0.9418	1.0000	1.1593	< .001
5	Pair pattern	T2 Subsidy	Both (H,L)	50	50	1.0000	0.9418	1.0000	1.1593	< .001
6	Pair pattern	T3 Fraud	Both (H,L)	50	50	1.0000	0.9418	1.0000	1.1593	< .001
7	Strategy by agent	PT1 Reflection gains	H	50	48	0.9600	0.8794	1.0000	0.7566	< .001
8	Strategy by agent	PT1 Reflection gains	L	50	50	1.0000	0.9418	1.0000	1.1593	< .001
9	Strategy by agent	PT2 Reflection losses	H	50	43	0.8600	0.7531	1.0000	0.3923	.007
10	Strategy by agent	PT2 Reflection losses	L	50	46	0.9200	0.8262	1.0000	0.5858	< .001
11	Strategy by agent	PT3 Loss aversion mixed gamble	H	50	47	0.9400	0.8522	1.0000	0.6643	< .001
12	Strategy by agent	PT3 Loss aversion mixed gamble	L	50	50	1.0000	0.9418	1.0000	1.1593	< .001
13	Strategy by agent	T1 Base	H	50	50	1.0000	0.9418	1.0000	1.1593	< .001
14	Strategy by agent	T1 Base	L	50	50	1.0000	0.9418	1.0000	1.1593	< .001
15	Strategy by agent	T2 Subsidy	H	50	50	1.0000	0.9418	1.0000	1.1593	< .001
16	Strategy by agent	T2 Subsidy	L	50	50	1.0000	0.9418	1.0000	1.1593	< .001
17	Strategy by agent	T3 Fraud	H	50	50	1.0000	0.9418	1.0000	1.1593	< .001
18	Strategy by agent	T3 Fraud	L	50	50	1.0000	0.9418	1.0000	1.1593	< .001

Note: PT = Prospect Theory scenarios from Kahneman & Tversky (1979); T = Title signaling scenarios from Spence (1973); H = High-type agent; L = Low-type agent. $\hat{p}$ = observed proportion of hits. Cohen's h = Effect size for proportions. p-values are one-tailed, testing against a chance level of .05.

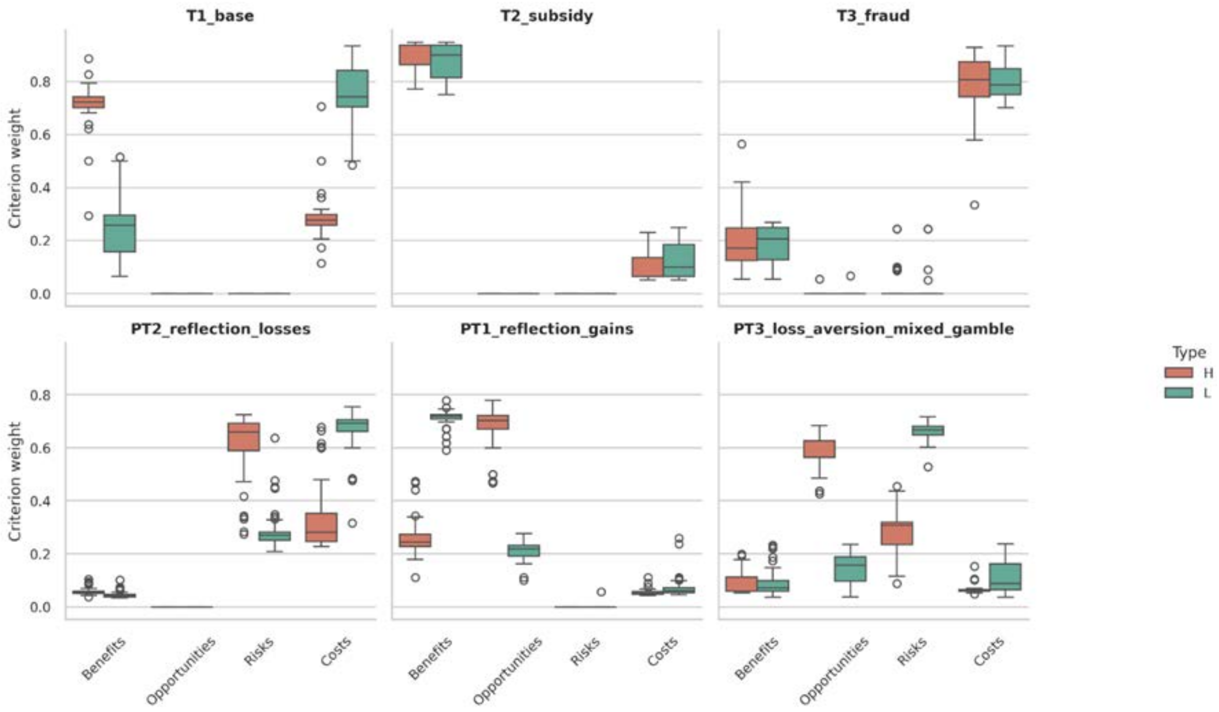


Figure 3: BOCR factor weighting by type. *Note:* PT = Prospect Theory scenarios from Kahneman & Tversky (1979); T = Title signaling scenarios from Spence (1973); H = High-type agent; L = Low-type agent.

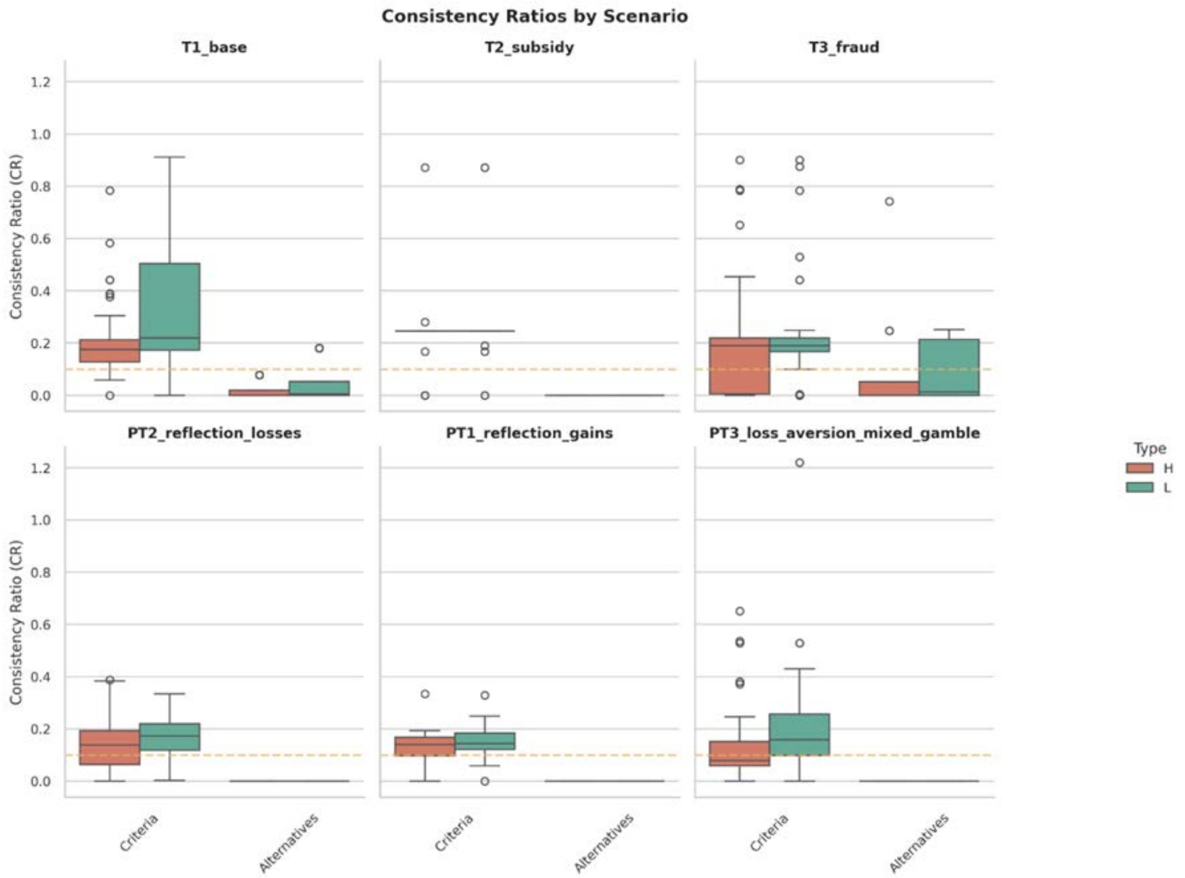


Figure 4: Consistency ratios by scenario. *Note:* T = Title signaling scenarios from Spence (1973); PT = Prospect Theory scenarios from Kahneman & Tversky (1979); H = High-type agent; L = Low-type agent.

Decision Factor Weights

The intermediate weights also moved in substantively meaningful directions. In the baseline Spence setting, the high type placed most weight on Benefits, whereas the low type placed most weight on Costs, reproducing the asymmetry that theoretically generates separation. In the gains lottery, the high type emphasized upside Opportunity, while the low type emphasized certain Benefits. In the losses lottery, the high type weighted Risk most heavily, whereas the low type placed more emphasis on Cost. This matters because it suggests the engine is not merely reproducing the correct labels at the end of the pipeline; it is reaching them through type-sensitive internal trade-offs.

Consistent with that reading, alternative-level AHP matrices were highly stable, with median consistency ratios of zero for both agent types, while criterion-level matrices were more variable. This is what we would expect: alternative comparisons use quantitative data directly, whereas criterion comparisons require qualitative judgments about the relative importance of BOCR dimensions. The fact that most criterion-level CRs still fell below the 0.10 threshold indicates that LLM-automated pairwise comparisons were generally reliable, with the consistency check acting as a built-in safeguard for cases deserving a closer review.

Discussion

This paper makes three contributions. First, it frames applied behavioral research in industry as a decision-engineering problem in which auditability is a first-class requirement. Second, it specifies an operational pipeline that combines LLM reasoning with deterministic stabilization through Binding & Netting and multi-criteria aggregation through AHP. Third, it demonstrates how type-conditional rankings and incentive-pattern diagnostics can be produced as inspectable artifacts rather than opaque recommendations.

At heart, Choice Engine is meant to simplify a part of the job that is usually messy, time-consuming, and difficult to explain to others. Applied behavioral scientists routinely work across client documentation, the academic literature, competitor case studies, qualitative interviews, and quantitative datasets before they can even begin to design an intervention. Choice Engine turns that early-stage complexity into a structured workflow. And because every step leaves an explicit trail, the tool produces not only a recommendation, but also a reasoning record that can be reviewed, challenged, updated, and defended.

> "Choice Engine produces not only a recommendation, but also a reasoning record that can be reviewed, challenged, updated, and defended."

For practitioners, the value of the workflow is upstream of experimentation. It supports high-throughput evidence digestion by turning documents into structured inputs rather than one-off summaries, offering a practical way to combine quantitative and qualitative evidence within the same decision framework. It also allows to make a pre-assessment and prediction of what may happen in a specific set up that can inform the design of the behavioral intervention and its probable outcome before launching it in the real world. This is an invaluable opportunity to tailor highly effective interventions with minimal risks. It is no coincidence that a growing body of work is proposing the use of agent-based simulations in many different fields tackling behavioral interventions such as health (Huang et al. 2021), workplace policy (Muñoz & Iglesias, 2021), and customer experience (Bell & Mgbemena, 2018), among others.

Incentive-pattern diagnostics, such as whether profiles separate or pool, also provide a useful segmentation lens. When type-conditional rankings diverge, interventions can be designed to attract or deter certain types selectively. When rankings converge, type-dependent levers are less likely to work. This connects the output of the workflow to concrete product and policy discussions around screening, fraud deterrence, trust-building, or compliance design. This approach is particularly useful for implementing behavioral segmentation strategies in which different individuals within the target population receive profile-specific interventions tailored to the same behavioral objective.

Choice Engine includes two distinct quality checks. The AHP Consistency Ratio in Step 6 provides a mathematical assessment of pairwise comparison coherence, while the LLM judge in Step 1 evaluates whether the input templates follow the workflow's neutrality rules. One check is formal, the other semantic. Together, they offer complementary safeguards against unreliable outputs.

Despite these checks, the limitations of Choice Engine should also be acknowledged. The engine has been validated with two seminal studies with well-defined constraints. Real-world conditions, however, tend to be less straightforward and much messier. One potential problem is including erroneous or incomplete data in the input that could be passed on the following steps of the process. Here, it is important to point out that Choice Engine's process can be iterated as many times as needed, making it possible to feed the initial input with data any time new information may become available. The role of the behavioral scientist in avoiding—or at least limiting—these problems is key. After all, Choice Engine is a tool that will require expert use. Behavioral scientists will have to be particularly careful with prompt leakage and informational gap assessment when making use of Choice Engine in real-life applications. Similarly, it should be noted that binding can create a false sense of precision if qualitative comparisons are treated as measurements rather than as structured judgments. LLM agents may also be inconsistent, overfit to prompt phrasing, or encode normative assumptions that require human review. AHP itself assumes independence among criteria, which may not hold in more complex product contexts. Most importantly, auditability does not guarantee correctness; it makes disagreements visible, but teams still need to decide which assumptions deserve more data.

Future work should evaluate the workflow in real private-sector settings, focusing on decision quality, time-to-insight, and updateability under changing evidence. Extending AHP to the Analytic Network Process (ANP), which models interdependencies among criteria, is a promising direction. Stronger calibration of quantitative bindings, better input governance through versioning and review workflows, and more principled uncertainty propagation from inputs to rankings are also clear opportunities for improvement. Finally, the further development of structured, auditable AI workflows that enhance, rather than replace, the judgment of applied behavioral scientists is warranted.

THE AUTHORS

Pablo Coello is a member of BeWay's Beway AI Team (BAIT), working on applied decision-support systems that combine behavioral economics, LLM-based reasoning, and multi-criteria decision analysis.

Pablo Villacañas is a software developer at BeWay's BeWay AI Team (BAIT), specializing in software architecture and Full Stack development. His research interests lie at the intersection of human behavior modeling and complex systems architecture, applying psychological frameworks to technical and logical environments.

Gonzalo Díaz is a member of BeWay's Beway AI Team (BAIT), specializing in implementing state-of-the-art AI techniques and RAG systems tailored for behavioral science applications and automated decision-making.

Sara Ramos-Cabo (PhD) is a behavioral scientist at BeWay. Her work centers on leveraging evidence from cognitive science, neuroscience, and psycholinguistics in combination with behavioral science frameworks to support organizations in making evidence-based human-centric interventions.

Antonio González is a behavioral data scientist in BeWay, holding a PhD in Health, Psychology, and Psychiatry, with a strong background in statistics and psychometrics. He works closely with consultants to deliver data-driven solutions for clients, designing rigorous experimental frameworks, building advanced statistical models, and translating behavioral insights into measurable business impacts. He also provides technical support and peer reviews across the team, ensuring methodological rigor and statistical validity throughout all client deliverables.

Francisco Ceballos is a member of BeWay's data department, where he leads the team responsible for ensuring quality in the design and analysis of experiments. He holds a PhD in genetics and statistics and works at the intersection of quantitative methods, experimental design, and behavioral science.

Juan Pablo Ospina Arboleda is a behavioral data scientist at BeWay, specializing in behavioral data analysis and the development of models that support evidence-based decision-making. His work focuses on applying principles of behavioral economics and decision-support systems to improve business processes in complex environments.

REFERENCES

Akerlof, G. A. (1978). The market for 'lemons': Quality uncertainty and the market mechanism. In P. Diamond & M. Rothschild (Eds.), *Uncertainty in economics* (pp. 235-251). Academic Press. https://doi.org/10.1016/B978-0-12-214850-7.50022-X

Bedard, K. (2001). Human capital versus signaling models: University access and high School drop-outs. *Journal of Political Economy, 109*(4), 749-775. https://doi.org/10.1086/322089

Bell, D., & Mgbemena, C. (2018). Data-driven agent-based exploration of customer behavior. *Simulation, 94*(3), 195-212. https://doi.org/10.1177/0037549717743106

Hauser, J. R., & Wernerfelt, B. (1990). An evaluation cost model of consideration sets. *Journal of Consumer Research*, *16*(4), 393-408. https://doi.org/10.1086/209225

Huang, W., Chang, C. H., Stuart, E. A., Daumit, G. L., Wang, N. Y., McGinty, E. E., ... & Igusa, T. (2021). Agent-based modeling for implementation research: An application to tobacco smoking cessation for persons with serious mental illness. *Implementation Research and Practice,* 2, 1-16. https://doi.org/10.1177/26334895211010664

Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., & Liu, T. (2025). A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, *43*(2), 42. https://doi.org/10.1145/3703155

Hvide, H. K. (2003). Education and the allocation of talent. *Journal of Labor Economics*, *21*(4), 945-976. https://doi.org/10.1086/377028

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, *47(2)*, 263-292. https://doi.org/10.1142/9789814417358_0006

Keeney, R. L., & Raiffa, H. (1993). *Decisions with multiple objectives: Preferences and value trade-offs*. Cambridge University Press.

Khot, T., Trivedi, H., Finlayson, M., Fu, Y., Richardson, K., Clark, P., & Sabharwal, A. (2023). Decomposed prompting: A modular approach for solving complex tasks. *arXiv*. https://doi.org/10.48550/arXiv.2210.02406

Michie, S., van Stralen, M. M., & West, R. (2011). The behaviour change wheel: A new method for characterising and designing behaviour change interventions. *Implementation Science*, *6*(1), 42. https://doi.org/10.1186/1748-5908-6-42

Muñoz, S., & Iglesias, C. A. (2021). An agent based simulation system for analyzing stress regulation policies at the workplace. *Journal of Computational Science, 51*, 101326. https://doi.org/10.1016/j.jocs.2021.101326

OpenAI. (2025). ChatGPT (5.2) [Large language model]. https://chat.openai.com/chat

Payne, J. W., Bettman, J. R., & Johnson, E. J. (1993). *The adaptive decision maker*. Cambridge University Press.

Saaty, T. L. (1980). *The analytic hierarchy process: Planning, priority setting, resource allocation*. McGraw-Hill International Book Company.

Saaty, T. L. (2005). *Theory and applications of the analytic network process: Decision making with benefits, opportunities, costs, and risks*. RWS Publications.

Saaty, T. L. (2008). Decision making with the analytic hierarchy process. *International Journal of Services Sciences*, *1*(1), 83-98. https://doi.org/10.1504/IJSSCI.2008.017590

Sclar, M., Choi, Y., Tsvetkov, Y., & Suhr, A. (2024). Quantifying language models' sensitivity to spurious features in prompt design or: How I learned to start worrying about prompt formatting. *arXiv*. https://doi.org/10.48550/arXiv.2310.11324

Simon, H. A. (1955). A behavioral model of rational choice. *The Quarterly Journal of Economics*, *69*(1), 99-118. https://doi.org/10.2307/1884852

Spence, M. (1973). Job market signaling. *The Quarterly Journal of Economics*, *87*(3), 355-374. https://doi.org/10.2307/1882010

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, *185*(4157), 1124-1131. https://doi.org/10.1126/science.185.4157.1124

West, R., & Michie, S. (2020). A brief introduction to the COM-B Model of behaviour and the PRIME Theory of motivation [v1]. *Qeios*. https://doi.org/10.32388/WW04E6

Building Behavioral Insights for International Development

RASHA ATTAR AND CHIARA CAPPELLINI

Behavioral Science Group

DEELAN MARU[1] AND DAVID HALPERN

Behavioral Insights Team

ZEINA AFIF

World Bank

Click or scan QR code
to watch video summary

Behavioral science has delivered measurable gains across development domains, yet the evidence base for international development remains thin and uneven. This article argues that the problem is not that behavioral tools don't work but that the field has consistently asked the wrong questions, in the wrong ways, and in the wrong institutional conditions. We identify three compounding failures. First, foundational behavioral assumptions are built on narrow, "WEIRD" samples, and development settings are treated as test beds for existing frameworks rather than sources of new theory. Second, research methods are poorly suited to development realities: excluding key populations, underestimating delivery volatility, and producing findings distorted by measurement problems. Third, the institutional conditions for generating, uncovering, and learning from behavioral evidence are frequently absent. For each failure, we propose a research agenda. The Global Behavioral Evidence Network for Development (GBEND) has been founded to coordinate action across all three domains.

A Research Agenda for the Global Behavioral Evidence Network for Development

Founded at the 2025 Behavioral Exchange conference in Abu Dhabi, the Global Behavioral Evidence Network for Development (GBEND) represents a landmark effort to expand and institutionalize behavioral insights in global development policy.

The network is stewarded by four chairing members: the UAE Office of Development Affairs, the Behavioral Science Group, the World Bank's Mind, Behavior, and Development Unit (eMBeD), and the Behavioral Insights Team. The network aims to ensure that the field of behavioral science produces evidence that serves people across different incomes, values, and contexts, and is fit for purpose when

1 Corresponding author: deelan.maru@bi.team

applied to development.

A core pillar of GBEND's mission is to provide rigorous thought leadership. This piece argues that the behavioral evidence base for international development is thin because the field has consistently asked the wrong questions, in the wrong ways, and in the wrong institutional conditions. It sets out a research agenda to address each in turn.

> "The field has consistently asked the wrong questions, in the wrong ways, and in the wrong institutional conditions."

The Challenge

Behavioral insights have delivered measurable gains across a range of development domains. Interventions grounded in behavior change have improved health outcomes, increased savings, raised school grades, and raised service uptake (OECD, 2017; DellaVigna & Linos, 2022), often at low cost and at scale. Such results have led to an expanding number of behavior change programs and specialized teams across governments, development agencies, NGOs, and multilaterals (Naru, 2024).

Despite these successes, however, substantive challenges remain in applying behavioral insights to international development[2]. Some evidence suggests that behavioral interventions have produced more uneven, heterogeneous, and frequently inconclusive results (Gvirtz & Sabherwal, 2024). Recent meta-research suggests that many findings across the social and behavioral sciences travel poorly beyond the contexts in which they were originally generated, particularly when evidence is built on narrow, high-income samples (Busara, 2024). In international development settings, this problem is compounded by cultural heterogeneity, linguistic diversity, data availability, institutional constraints, and research infrastructure—making it difficult to disentangle whether null or mixed results reflect genuine limits of behavioral mechanisms or failures of translation, design, and delivery.

A growing body of evidence suggests that some behavioral mechanisms do translate across settings, particularly when interventions are simple, address practical barriers, and align with local delivery systems (Hallsworth, 2023; DellaVigna & Linos, 2022). For example, a meta-analysis found that choice architecture interventions, on average, have meaningful effects on behavior, though effect sizes vary substantially across populations and contexts (Mertens et al., 2022). Even so, the evidence base is still too limited to draw broad conclusions about the effectiveness of behavioral science in international development.

This article diagnoses three compounding failures—in foundational assumptions, research methods, and institutional conditions—and proposes a research agenda to address them.

Research Question: How can we Optimize Behavioral Insights to Advance International Development Goals?

We identify three research questions to structure how behavioral insights could better serve international development:

- **Conceptual:** Which behavioral mechanisms travel across contexts, which depend on local norms and institutions, and when are adaptations or new concepts needed?
- **Methodological:** How can behavioral interventions be tested in ways that preserve meaning, rigor, and relevance in international development settings—and at low cost?
- **Institutional:** What institutional and political conditions are optimal for behavioral interventions to be implemented, sustained, and scaled in an international development program?

Section 1: Foundational Behavioral Assumptions Across Contexts

Many of our core behavioral assumptions and models are built on a narrow set of high-income settings (Busara, 2024). In international development, some mechanisms do travel, but others do not hold so readily.

Understanding How Behavioral Concepts Travel Across Settings

While the field often assumes that cognitive biases are universal human traits, recent meta-research

2 This piece uses a broad interpretation of international development. The underlying goal is human development, namely, advancing human abilities and creating the conditions for humans to use those abilities (UNDP, 2015).

suggests that mechanisms often leveraged in WEIRD[3] countries, such as salience, reminders, or social comparisons, operate differently across settings (Busara, 2024). To serve international development better, we must gain a firmer understanding of the interplay between universal behavioral processes and culturally defined ones, as outlined by Michael Hallsworth (2023, p. 78).

There are, however, a few large cross-country tests of core biases that start to answer this question. A multinational replication of Prospect Theory[4] choice patterns across 19 countries, for instance, found that patterns were detectable but often weaker than in original studies (Ruggeri et al., 2020). Thus, some biases may be detectable across contexts, but their strength, and therefore their policy relevance, may vary.

These cross-country tests confirm that basic cognitive patterns are broadly detectable but context-sensitive. A more operationally useful line of questioning would be to ask which classes of interventions are robust and which require adaptation? Evidence from multi-country implementations suggests a working distinction: interventions that reduce friction (simplification, reminders, removing procedural barriers) tend to transfer because they target universal cognitive constraints, while interventions that leverage culturally constructed meanings (social norms messaging, authority framing, identity appeals) show variation across settings (see Hallsworth, 2023). This distinction, between mechanism-level universality and meaning-level variability, provides a practical heuristic for prioritizing which interventions to deploy with confidence and which to test before scaling.

Social norms interventions assume that learning what "most people" do will shift behavior. But in settings where norms are enforced through close kinship networks or specific in-groups, or are conditional on normative expectations about what others approve of rather than what they actually do (Bicchieri, 2005), generic social comparisons may be irrelevant or counterproductive.

Using a Broader Framework of Human Behavior to Build a Better Evidence Base

Testing the transferability of behavioral concepts developed in WEIRD contexts is a necessary first step, but it is insufficient. The more challenging implication is that development settings should be not only test beds for existing frameworks, but also sources of new concepts, categories, and standards of explanation that reshape the field itself.

> "Testing the transferability of behavioral concepts developed in WEIRD contexts is a necessary first step, but it is insufficient."

The World Bank's "World Development Report 2015: Mind, Society, and Behavior" framework identifies three lenses of behavioral influence:

1. **Thinking automatically:** cognitive shortcuts, heuristics, and the contextual cues that shape them.
2. **Thinking socially:** how behavior is shaped by what others do and what others are perceived to expect.
3. **Thinking with mental models:** the shared categories, identities, narratives, and worldviews through which people interpret experience.

The field has drawn disproportionately on the first lens. A substantial share of rigorously evaluated interventions in low- and middle-income countries (reminders, feedback, micro-incentives, salience, and commitment devices) target automatic thinking (Booth et al., 2022). To build a broader evidence base better suited to international development, practitioners should expand their practice across the three lenses.

Understand the Link Between Human Cognition and Material Scarcity

Mani et al. (2013) showed that Indian farmers' cognitive scores were lower during periods of limited food supply (pre-harvest) than during the post-harvest period. A more recent meta-analytic synthesis reveals a more nuanced picture: the analysis finds a link between financial scarcity and lower

3 Western, Educated, Industrialized, Rich and Democratic countries, as defined by Henrich et al. (2010)

4 Prospect Theory is the idea that people judge gains and losses relative to a reference point, and they tend to feel losses more strongly than equivalent gains.

cognitive performance but attributes 60% of this effect to educational access (de Almeida et al., 2024). This suggests that the idea of cognitive scarcity is still rooted in structural challenges (access to education, cumulative disadvantage across the life course) as opposed to purely cognitive channels.

Account for the Role of Social Enforcements

Social dynamics and enforcements govern our behaviors in substantive ways beyond heuristics—and understanding how they shapeshift across societies is key. For example, behavioral interventions frequently assume that individuals can privately accumulate and reinvest returns. In many development contexts, income is subject to strong informal claims from family and community. Where earnings are observable, social obligations to share can operate like an informal tax on returns, creating incentives to conceal income even at a cost. In rural Kenya, for example, Jakiela and Ozier (2016) found that women were willing to forgo earnings to keep investment returns hidden from relatives. This is not a cognitive bias but a social norm operating through enforcement and observability (Lens 2).

Consider Local Mental Models to Design Relevant Interventions

Development practitioners seeking to influence behaviors must gain a deeper understanding of the reference points and narratives available within a given social context. Ray and Genicot (2017) theorize that when the gap between an individual's current status and attainable reference points is too wide, the result is rational disengagement, which they term "aspirations failure." Experimental evidence from Mexico demonstrates that showcasing attainable, local role models rather than distant successes can reactivate agency and improve economic outcomes more effectively than financial incentives alone (Lybbert & Wydick, 2018). Aspiration failure is a Lens 3 phenomenon: it operates through the mental models people hold about what is possible for someone like them.

These examples share a common structure. Each begins with a phenomenon that could be interpreted through Lens 1, but closer examination reveals that the behavioral driver sits in Lens 2 or Lens 3. Misidentifying the lens leads to misidentifying the intervention.

The three lenses are not equally actionable, however. Lens 1 interventions (reminders, defaults and simplification) map cleanly onto RCT designs with short feedback loops and measurable outcomes. Lens 2 and Lens 3 interventions require designs that can capture social dynamics and shifts in mental models over longer timeframes. These could be longitudinal studies or ethnographic research, which are more expensive and slower. This helps explain why the field so often defaults to Lens 1 regardless of whether it is the right diagnosis. Choosing a research method before identifying the relevant lens almost guarantees this bias will persist.

> "Choosing a research method before identifying the relevant lens almost guarantees this bias will persist."

Section 2: Research Tools to Build Behavioral Insights for International Development

Even where behavioral mechanisms hold, the evidence base for international development is weakened by the underlying methods and tools. We propose three methodological principles to address this issue.

Diagnose Before You Design

The dominant model in behavioral policy—identify a bias, design a corrective nudge, test it—assumes the behavioral bottleneck is already known. In high-income, high-capacity settings, this assumption is often reasonable: the institutional infrastructure works, the target population is reachable, and the binding constraint is plausibly cognitive. In international development, however, these assumptions frequently do not hold. The binding constraint may be structural, social, or institutional, and designing a solution before identifying it can waste resources and produce misleading null results. A more useful starting point is a diagnostic-first approach, as formalized by the World Bank's Mind, Behavior, and Development Unit (eMBeD), which maps the psychological, social, and structural factors shaping decisions for both target populations and implementers before intervention design (World Bank, 2015).

A good diagnostic should uncover the factors that most often undermine intervention design in international development settings: whether delivery systems are stable enough to support implementation, who actually delivers services on the ground, whether materials are linguistically and culturally appropriate, which populations standard methods miss, and what level of testing is feasible under real budget constraints.

Diagnosis also means designing research methods that include the populations development policy most needs to reach: rural women, people with low literacy, the digitally disconnected, and mobile populations such as refugees, migrants, and informal workers. They are excluded by research designs that assume urban residence, literacy, smartphone ownership, and stable addresses as default conditions for participation.

Standard methods can also miss the people the development policy is supposed to reach. Behavioral work often assumes that people can read and respond to written prompts, yet literacy rates remain very low in some settings (World Population Review, 2026). Alternatives exist—interactive voice response, audio-based nudges, and visual decision aids—but they are typically more expensive. Emerging AI-enabled tools may reduce these costs (Human Truths, n.d.).

Design for Instability

In many international development programs, implementation can be impacted by staff turnover, absenteeism, seasonality, and exogenous shocks, in which case research designs that assume stable delivery will produce misleading results. For example, in a handwashing trial in Bangladesh, observations were interrupted by a cholera outbreak, a heatwave, Eid-related camp closures, violence, and Cyclone Mocha (BIT, 2023a). The goal is not to control for instability but to build it into the design from the outset.

> "The goal is not to control for instability but to build it into the design from the outset."

Implementation science offers a useful way to design for this reality. Rather than trying to keep every component identical across sites, it recommends holding constant the core function of an intervention (the mechanism it must achieve) while allowing the form (how it is delivered) to adapt to local conditions (Hawe et al., 2004). Behavioral science practitioners are becoming increasingly interested in how to design trials that reflect implementation challenges (e.g., List, 2024). Some practical implications: specify the non-negotiables up front, offer a number of interventions, plan and track permitted adaptations during delivery, and measure implementation outcomes such as feasibility and fidelity alongside behavioral outcomes.

Adaptive trial designs can also preserve rigor while accommodating real-world instability, which they achieve by allowing prespecified modifications in implementation or data collection in reaction to interim data or changing circumstances. For example, outcomes can be analyzed at interim checkpoints, and, based on those results, the study might adjust allocation or drop ineffective treatment arms while the trial is ongoing (Lauffenburger et al., 2022). Wider adoption of adaptive designs would also reduce the risk that null results are misattributed to intervention failure when the underlying cause is measurement disruption.

Measure Behavior, Not Reports

Smart evaluation techniques can overcome the common problem of social desirability bias. Participants may report what they think the researcher wants to hear rather than what they actually do. This problem exists everywhere, but it may be more pronounced in development programs where the intervention administrator is perceived as more powerful or authoritative. In sanitation and hygiene, for example, self-reports often exceed observed behavior. One study in rural India found that reported latrine use was nearly double the level recorded by unobtrusive sensors (Sinha et al., 2016). Novel techniques have been used to overcome these issues. For example, BIT measured the number of soap pedal presses used at handwashing stations to compare the control and treatment arms (BIT, 2022).

The search for robust behavioral evidence in international development is frequently undermined by poor data quality. Furthermore, volatility in program delivery, driven by shocks, seasonality, and staff turnover, renders missing data a pervasive challenge. When attrition is non-random, even a rigorously designed randomized controlled trial (RCT) can be fatally compromised, with one survey finding that nearly 25% of clinical trials had more than 10% of primary outcome data missing (Hewitt et al., 2010). One response is to use intention-to-treat design in RCTs, which can disentangle the intervention's efficacy from the program's effectiveness (Duflo et al., 2007). Emerging technologies could expand what is measurable at lower cost without sacrificing rigor.

AI-assisted qualitative tools can speed up transcription, translation, coding, and analysis, making it far easier to work across large volumes of inputs. They may also support co-creation by rapidly turning participant ideas into visual or textual prototypes that can be tested and refined in real time. This enables informants to co-create messaging and concepts with behavioral scientists, and it can also bridge cultural divides (BIT, 2023b).

Voice-based tools such as interactive voice response can reduce reliance on text-based methods and may offer greater privacy for sensitive topics (Gibson et al., 2017; Greenleaf et al., 2017). AI-assisted prompting and summarization could also make diary studies and other longitudinal qualitative methods cheaper to run at scale by helping researchers send adaptive prompts, track incoming responses, and synthesize material over time (Moëll & Aronsson, 2025).

One emerging option is synthetic participants: AI-generated respondents built by conditioning language models on real demographic data. These may help in early-stage design, especially where conventional recruitment is expensive or where participation itself may be unsafe. Early results suggest that synthetic samples can track broad patterns in attitudes and opinions reasonably well, though performance is weaker for behavioral outcomes, and effects may be overstated (Hanna-Amodio et al., 2025). They may therefore be useful in limited early-stage applications—but not as a substitute for real-world testing.

These tools carry their own risks, however. Most AI systems are trained on data that skew toward WEIRD countries, which means they may encode and amplify the same representational biases identified in Section 1. In settings with pronounced power asymmetries and limited accountability mechanisms, the risks of misuse or misinterpretation are heightened. Practitioners should thus be explicit about what AI tools should not yet be used for in development contexts, particularly where outputs could influence consequential decisions about populations with little recourse.

Section 3: The Role of Institutions in Behavioral Insights for Development Policy

Behavioral science depends on the institutions around it: who sets priorities, who controls data and budgets, and whether there is capability and legitimacy to test, learn, and scale. If behavioral science is to serve international development, it requires institutions that can generate evidence, translate that evidence into delivery, and learn over time.

Creating the Right Institutions

Where a behavioral team sits can shape what it can achieve. Teams embedded in delivery ministries are usually closer to real service constraints, but they depend on the strength of the host ministry and whether their mandate survives political turnover (Ñañez et al., 2024). Capability matters, too. Effective behavioral work requires more than experimental skills alone; it also requires knowledge of delivery, data, and measurement capabilities, as well as political and partnership skills (Hernández-Agramonte, 2025).

Funding is part of this same institutional picture. In international development, much of the funding sits with bilateral donors, multilaterals, foundations, and NGOs outside governments. Strong results are more likely to be written up than null results (Franco et al., 2014), and local partners often conduct fieldwork without control over budgets or full overheads, which weakens local research capacity (Crane et al., 2018).

> "These are structural conditions that shape what evidence gets produced, who it serves, and whose knowledge counts."

These funding dynamics reflect deeper power asymmetries that run through the field. Global North researchers typically control study design, budgets, and publication decisions, while local implementers and research partners contribute the contextual knowledge, fieldwork, and access that make studies possible, often without authorship, overhead recovery, or influence over how findings are used. Moreover, donors shape what gets tested by setting priorities that may not reflect the most binding constraints facing governments or communities. And the relationship between researchers and participants is itself asymmetric, in that data is extracted, findings are published elsewhere, and communities rarely see results in a form they can use. These are structural conditions that shape what evidence gets produced, who it serves, and whose knowledge counts.

Assembling the Evidence

A further issue is evidence infrastructure. Even when good studies exist, they often do not contribute to cumulative learning, because synthesis is not up to date, locally relevant, or available in accessible formats and languages. Existing "what works" systems remain heavily WEIRD-centered, and many evidence products are too static or too narrow to support live policy decisions (Halpern & Maru, 2024). What is needed instead is a stronger evidence infrastructure for international development in the form of living evidence reviews, evidence gap maps, and synthesis products designed for real policy questions (see, e.g., Saran et al., 2020; The Global SDG Synthesis Coalition, 2024).

Findings from behavioral research should be published. Governments default to non-disclosure, as they lack technical capability, fear political backlash, or are capacity-constrained. The cumulative effect is a systematic bias toward silence, and so the behavioral studies most relevant to government implementation are the least likely to enter the public evidence base. The "what works" clearinghouses and evidence gap maps described above can only index what is visible.

Assessing the Political Economy of Nudging

The effectiveness of behavioral interventions is mediated by the political context and institutional environment within which they operate. A nudge that works in a high-trust, high-capacity system can fail or backfire when the state is absent or contested. Furthermore, many interventions assume systems will operate effectively; however, the "last mile" can often be the most difficult to overcome (Wills Silva, 2025). Some nudges work partly because enforcement is credible, but where the state is constrained, the enforcement mechanism may not work. In environments with low trust in institutions, messaging can

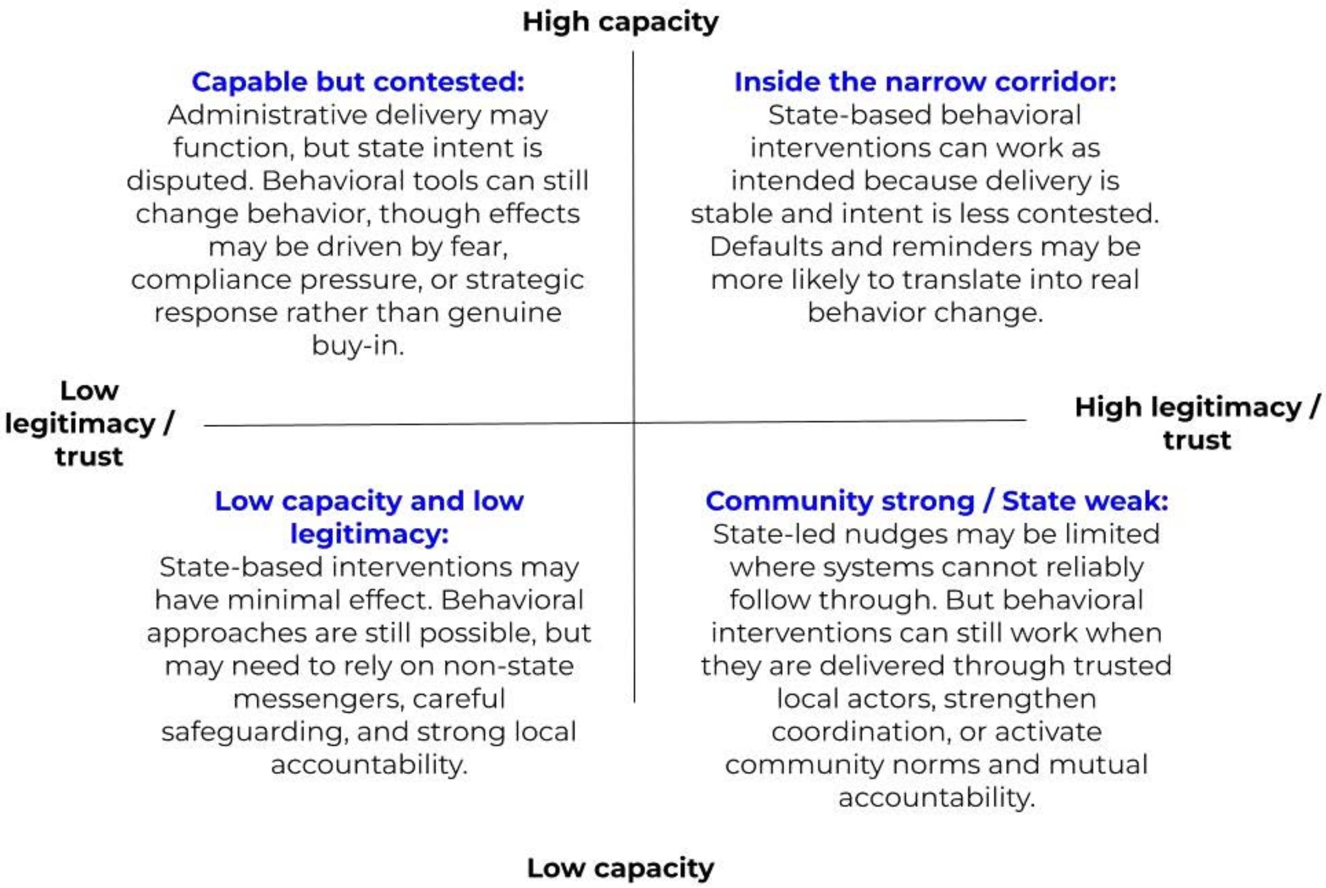


Figure 1: 2x2 matrix showing the intersection of state capacity and legitimacy/trust on behavioral interventions.

be interpreted as surveillance or overreach.

One way to understand the political context can be through the lens of the *Narrow Corridor* (Acemoglu & Robinson, 2019), i.e., effective governance sits in a corridor where the state is strong enough to provide services and enforce rules, but society is strong enough to constrain abuse. In development, it is important to understand how close or far the context is from this ideal corridor. Within the corridor, top-down behavioral policies at scale may translate more readily into wider impacts. Conversely, outside the corridor, those same tools may be weaker or interpreted differently, while society-based nudges may become more important.

This can be theorized as a 2x2 matrix, with state capacity on one axis and legitimacy or trust on the other. The matrix should be read not as showing where behavioral interventions can or cannot work but as indicating which behavioral channels are likely to be more effective in different political contexts.

Conclusion: The Next Chapter of Behavioral Insights for International Development

This research agenda reflects the collective expertise of practitioners who have spent years embedding behavioral science within international development. GBEND's aim is to ensure that behavioral insights are embedded from the outset, producing evidence that is as relevant to the global majority as it is rigorous.

GBEND welcomes new organizations to the network. Please email chiara.cappellini@bsg.ae if you would like to be involved.

THE AUTHORS

Rasha Attar is director of the Behavioral Science Group (BSG), the central behavioral science unit for the UAE government based in the Office of Development Affairs. She has over 18 years of experience in public service, most recently embedding applied behavioral science into strategy and policy design across health, education, sustainability, and social development. She is co-chair of the Global Behavioral Evidence Network for Development (GBEND), convening multi-disciplinary experts to strengthen development outcomes by incorporating applied behavioral science. Rasha holds a degree in Economics from the American University of Beirut, and an Executive MSc in Behavioural Science from the London School of Economics.

Zeina Afif is a senior social scientist with the World Bank's Mind, Behavior, and Development Unit (eMBeD). With over 15 years of experience, she designs and implements behavioral science interventions to address complex global challenges, including women's economic empowerment, youth unemployment, and social cohesion. Her work focuses on institutionalizing behavioral insights within government systems to improve public service delivery across the Middle East, Africa, and Latin America. Zeina is a visiting senior fellow at the London School of Economics and holds an MBA from George Washington University and an MSc in Behavioral Science from the LSE.

Deelan Maru is a senior advisor at the Behavioral Insights Team (BIT) and a member of the GBEND Secretariat. Deelan was policy advisor to David Halpern and has co-authored landmark publications, including the Global Evidence Report, which serves as a foundation for the network's research agenda. Deelan holds an MSc in International Public Policy from University College London.

David Halpern CBE is president emeritus of the Behavioral Insights Team (BIT) and a chairing member of the Global Behavioral Evidence Network for Development (GBEND). As the founding director of the UK's "Nudge Unit," he has been a central figure in the global movement to apply behavioral science to public policy. David has served as a key advisor to numerous behavioral units worldwide, supporting the institutionalization of the field across diverse governments and international organizations. A former National What Works advisor, he has held academic positions at Cambridge and Harvard and is a visiting professor at King's College London.

Chiara Cappellini is the principal behavioral scientist at the Behavioral Science Group within the Office of Development Affairs. As a member of the GBEND Secretariat, she leads the network's convening and strategic expansion, driving many of its foundational initiatives. Her past experience includes work bridging social development and innovation through emerging technologies delivery with major international organizations such as Meta, UNICEF, and Google. She holds an MSc in International Development and Research Methods from the University of Edinburgh and a BSc in Social Policy from the London School of Economics.

REFERENCES

Acemoglu, D., & Robinson, J. A. (2019). *The narrow corridor: States, societies, and the fate of liberty: Winners of the 2024 Nobel Prize in economics*. Penguin UK.

Bicchieri, C. (2005). *The grammar of society: The nature and dynamics of social norms*. Cambridge University Press.

BIT. (2022). *If you build it, will they come? Increasing handwashing in Bangladesh during a pandemic.* https://www.bi.team/blogs/if-you-build-it-will-they-come-increasing-handwashing-in-bangladesh-during-a-pandemic/

BIT. (2023a). *Improving hygiene behaviours within Rohingya settlements in Cox's Bazar*. https://www.bi.team/publications/improving-hygiene-behaviours-within-rohingya-settlements-in-coxs-bazar/

BIT. (2023b). *AI to BI: Using artificial intelligence in qualitative research*. Behavioural Insights Team. https://www.bi.team/blogs/ai-to-bi-using-artificial-intelligence-in-qualitative-research-in-the-middle-east/

Booth, S., Cavatassi, R., Curtis, B., Kim, D. S., Langer, L., Mapitsa, C. B., ... & Robertsen, J. (2022). *Behavioural science interventions within the development and environmental fields in developing countries: An evidence gap map*. https://ueaeprints.uea.ac.uk/id/eprint/91067

Busara. (2024). *A better how: Notes on developmental meta-research* (P. Forscher & M. Schmidt, Eds.). https://doi.org/10.62372/ISCI6112

Crane, J. T., Andia Biraro, I., Fouad, T. M., Boum, Y., & R. Bangsberg, D. (2018). The 'indirect costs' of underfunding foreign partners in global health research: A case study. *Global Public Health*, *13*(10), 1422-1429. https://doi.org/10.1080/17441692.2017.1372504

de Almeida, F., Scott, I. J., Soro, J. C., Fernandes, D., Amaral, A. R., Catarino, M. L., ... & Ferreira, M. B. (2024). Financial scarcity and cognitive performance: A meta-analysis. *Journal of Economic Psychology*, *101*, 102702. https://doi.org/10.1016/j.joep.2024.102702

DellaVigna, S., & Linos, E. (2022). RCTs to scale: Comprehensive evidence from two nudge units. *Econometrica*, *90*(1), 81-116. https://doi.org/10.3982/ECTA18709

Duflo, E., Glennerster, R., & Kremer, M. (2007). Using randomization in development economics research: A toolkit. *Handbook of Development Economics*, 4, 3895-3962. https://doi.org/10.1016/S1573-4471(07)04061-2

Franco, A., Malhotra, N., & Simonovits, G. (2014). Publication bias in the social sciences: Unlocking the file drawer. *Science*, *345*(6203), 1502-1505. https://doi.org/10.1126/science.1255484

Genicot, G., & Ray, D. (2017). Aspirations and inequality. *Econometrica*, *85*(2), 489-519. https://doi.org/10.3982/ECTA13865

Gibson, D. G., Farrenkopf, B. A., Pereira, A., Labrique, A. B., & Pariyo, G. W. (2017). The development of an interactive voice response survey for noncommunicable disease risk factor estimation: technical assessment and cognitive testing. *Journal of Medical Internet Research*, *19*(5), e112. https://doi.org/10.2196/jmir.7340

The Global SDG Synthesis Coalition. (2024). *US$74 million new investments in SDG evidence to be announced*. https://www.sdgsynthesiscoalition.org/news/us74-million-new-investments-sdg-evidence-be-announced

Greenleaf, A. R., Gibson, D. G., Khattar, C., Labrique, A. B., & Pariyo, G. W. (2017). Building the evidence base for remote data collection in low-and middle-income countries: Comparing reliability and accuracy across survey modalities. *Journal of Medical Internet Research*, *19*(5), e140. https://doi.org/10.2196/jmir.7331

Gvirtz, A., & Sabherwal, A. (2024). The limits of doing global, cross-cultural behavioral science research. *Proceedings of the National Academy of Sciences*, *121*(36), e2316690121. https://doi.org/10.1073/pnas.2316690121

Hallsworth, M. (2023). *A manifesto for applying behavioral science*. https://www.bi.team/publications/a-manifesto-for-applying-behavioral-science/

Halpern, D., & Maru, D. (2024). *A blueprint for better international collaboration on evidence.* https://www.bi.team/publications/international-collaboration-evidence/

Hanna-Amodio, A., Cappellini, C., Attar, R., & Finlayson, A. (2025). A new tool for behavioral scientists: Exploring the potential of synthetic participants to accelerate how we study attitudes and behaviors. In A. Samson (Ed.), *The Behavioral Economics Guide 2025* (pp. 16–25). https://www.behavioraleconomics.com/be-guide/the-behavioral-economics-guide-2025/

Hawe, P., Shiell, A., & Riley, T. (2004). Complex interventions: How "out of control" can a randomised controlled trial be?. *BMJ*, *328*(7455), 1561-1563. https://doi.org/10.1136/bmj.328.7455.1561

Hernández-Agramonte, J. (2025). *Turning evidence into routine practice: Lessons from embedded evidence labs.* Innovations for Poverty Action. https://poverty-action.org/turning-evidence-routine-practice-lessons-embedded-evidence-labs

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences*, *33*(2-3), 61-83. https://pubmed.ncbi.nlm.nih.gov/20550733/

Hewitt, C. E., Kumaravel, B., Dumville, J. C., Torgerson, D. J., & Trial Attrition Study Group. (2010). Assessing the impact of attrition in randomized controlled trials. *Journal of Clinical Epidemiology*, *63*(11), 1264-1270. https://doi.org/10.1016/j.jclinepi.2010.01.010

Human Truths. (n.d.). *Uncover human truths at scale.* https://humantruths.ai/

Jakiela, P., & Ozier, O. (2016). Does Africa need a rotten kin theorem? Experimental evidence from village economies. *The Review of Economic Studies*, *83*(1), 231-268. https://doi.org/10.1093/restud/rdv033

Lauffenburger, J. C., Choudhry, N. K., Russo, M., Glynn, R. J., Ventz, S., & Trippa, L. (2022). Designing and conducting adaptive trials to evaluate interventions in health services and implementation research: Practical considerations. *BMJ Medicine*, *1*(1), e000158. https://doi.org/10.1136/bmjmed-2022-000158

List, J. A. (2024). Optimally generate policy-based evidence before scaling. *Nature*, *626*(7999), 491-499. https://doi.org/10.1038/s41586-023-06972-y

Lybbert, T. J., & Wydick, B. (2018). Poverty, aspirations, and the economics of hope. *Economic Development and Cultural Change*, *66*(4), 709-753. https://doi.org/10.1086/696968

Mani, A., Mullainathan, S., Shafir, E., & Zhao, J. (2013). Poverty impedes cognitive function. *Science*, *341*(6149), 976-980. https://doi.org/10.1126/science.1238041

Mertens, S., Herberz, M., Hahnel, U. J., & Brosch, T. (2022). The effectiveness of nudging: A meta-analysis of choice architecture interventions across behavioral domains. *Proceedings of the National Academy of Sciences*, *119*(1), e2107346118. https://doi.org/10.1073/pnas.2107346118

Moëll, B., & Sand Aronsson, F. (2025). Journaling with large language models: A novel UX paradigm for AI-driven personal health management. *Frontiers in Artificial Intelligence*, *8*, 1567580. https://doi.org/10.3389/frai.2025.1567580

Ñañez, P., Decuypere, M., & Hartong, S. (2024). Unpacking the contextualities of behavioural public policy: A case study of the Peruvian Nudge unit MineduLAB. *Journal of Social Policy*, *55*(1), 304-320. https://doi.org/10.1017/S0047279424000187

Naru, F. (2024). Behavioral public policy bodies: New developments & lessons. *Behavioral Science & Policy*, *10*(1), 1-17. https://doi.org/10.1177/23794607241285614

OECD (2017), *Behavioural Insights and Public Policy: Lessons from Around the World*, OECD Publishing, Paris. https://doi.org/10.1787/9789264270480-en

Ruggeri, K., Alí, S., Berge, M. L., Bertoldo, G., Bjørndal, L. D., Cortijos-Bernabeu, A., ... & Folke, T. (2020). Replicating patterns of prospect theory for decision under risk. *Nature Human Behaviour*, *4*(6), 622-633. https://doi.org/10.1038/s41562-020-0886-x

Saran, A., White, H., Albright, K., & Adona, J. (2020). Mega-map of systematic reviews and evidence and gap maps on the interventions to improve child well-being in low- and middle-income countries. *Campbell Systematic Reviews*, *16*(4), e1116. https://doi.org/10.1002/cl2.1116

Sinha, A., Nagel, C. L., Thomas, E., Schmidt, W. P., Torondel, B., Boisson, S., & Clasen, T. F. (2016). Assessing latrine use in rural India: A cross-sectional study comparing reported use and passive latrine use monitors. *The American Journal of Tropical Medicine and Hygiene*, *95*(3), 720-727. https://doi.org/10.4269/ajtmh.16-0102

UNDP. (2015). What is human development? *Human Development Reports.* https://hdr.undp.org/content/what-human-development

Wills Silva, M. (2025). *Behavioural science and rigour at the last mile of aid.* BIT. https://www.bi.team/comment/behavioural-science-and-rigour-at-the-last-mile-of-aid/

World Bank. (2015). *World Development Report 2015: Mind, Society, and Behavior*. https://www.worldbank.org/en/publication/wdr2015

World Population Review. (2026). *Literacy Rate by Country 2026*. https://worldpopulationreview.com/country-rankings/literacy-rate-by-country

When Evidence Doesn't Travel: A Practitioner Case Study on Localizing Behavior Change Techniques

MARISA NOWICKI[1]

Lirio

Click or scan QR code
to watch video summary

Context matters in behavioral science. Yet, much of the field's evidence base is derived from Western, Educated, Industrialized, Rich, and Democratic (WEIRD) populations — even though more than 80% of people belong to other groups. While behavioral science has developed robust theories, taxonomies, and design tools, it offers less guidance on how to localize behavior change techniques for multicultural and non-Western contexts. This practitioner case study examines how one behavioral science team localized a digital health intervention for Costa Rica (CR) while preserving theoretical foundations. Rather than proposing a prescriptive framework, the paper illustrates how cultural frameworks and culturally informed audit activities were used to structure uncertainty, refine research questions, and guide localization decisions across strategy, copy, and visual design. The case highlights how practitioners working in real-world settings can complement existing evidence-based tools with structured processes to make localization considerations more visible, examinable, and actionable.

Introduction: When Evidence Doesn't Travel

Behavioral scientists know that *context matters*. One element of context, culture, must be considered when developing a new intervention or scaling an existing one.

Practitioners applying behavioral science in real-world settings often encounter a localization challenge: evidence-based insights and tools that are effective in one context can be difficult to interpret or apply in another. A growing body of evidence suggests that behavioral interventions can perform unevenly when applied across cultural contexts (Catoja & Crede, 2025; Hall et al., 2016; Talhelm, 2026). While the field has developed robust theories, taxonomies, and design tools, it offers less guidance on localizing behavior change techniques for multicultural and non-Western contexts (Castro et al, 2024, Henrich et al., 2010).

"While the field has developed robust theories, taxonomies, and design tools, it offers less guidance on localizing behavior change techniques for multicultural and non-Western contexts."

From Evidence to Application: A Practitioner Gap

A key contributor to this challenge is the uneven representation of contexts and populations within behavioral science research. Much of the

1 mnowicki@Lirio.com

existing evidence base has been generated in Western, Educated, Industrialized, Rich, and Democratic (WEIRD) societies (Henrich et al., 2010). Cultural norms, social structures, and institutional arrangements can shape how behavioral mechanisms are activated, influencing whether and how interventions translate across settings. And when insights derived from a relatively narrow set of contexts are applied elsewhere, outcomes may become less predictable—not because the underlying theories are invalid, but because contextual assumptions embedded in the evidence may not fully hold. This difficulty in scaling insights can often be traced to two common sources of error: lack of representativeness of the population and of the situation (List, 2021).

Evidence-based tools are critical to translating behavioral science research into practice. Resources like the Theory and Techniques Tool (TaTT) provide a shared language and structured starting point for intervention design (Johnston, 2021). However, practitioners may encounter limitations when interpreting these evidence mappings across contexts. Because such tools synthesize research that disproportionately represents WEIRD populations and settings, they can also inherit the assumptions and constraints of the underlying evidence base. As a result, these tools are often well suited to answer questions such as 'Which mechanisms or techniques are relevant to this behavioral barrier?' but are less supportive for the context-specific question 'To what extent does the evidence support using this technique to address this barrier *in this context*?'

Although all interventions require tailoring to the local context, practitioners working outside WEIRD contexts may find that widely used tools require a disproportionate amount of additional sensemaking to apply reliably. In the absence of explicit guidance for incorporating cultural nuance, practitioners are often left to rely on individual judgment and ad hoc processes to determine when and how to localize the intervention. Even behavioral scientists are prone to biases and blind spots when attempting to replicate interventions in unfamiliar cultural contexts.

This use case offers an illustrative example of how a US-based behavioral science team localized a digital health intervention in Costa Rica (CR) by using cultural frameworks and cultural audit activities to structure uncertainty, uncover assumptions, and guide adaptation decisions. It offers a reference point for practitioners facing similar challenges when evidence does not travel easily across contexts.

Towards a More Systematic Approach for Cultural Adaptation

Uncertainty Management

Cultural localization failures in behavioral science are often failures in uncertainty management rather than failures of underlying behavioral science theory. Teams frequently overlook the assumptions embedded in evidence bases, determinant categorization, and mechanism selection—particularly when interventions are applied across cultural contexts. The Rumsfeld Matrix provides a simple 2×2 lens for categorizing these risks, based on awareness and understanding (The Uncertainty Team, n.d.).

"Cultural localization failures in behavioral science are often failures in uncertainty management rather than failures of underlying behavioral science theory."

The Rumsfeld Matrix offers a useful framework for identifying how to mitigate risks of cultural mismatch in intervention design and scaling. Cultural blind spots often fall into the 'unknown knowns' category. However, many become visible by meaningfully engaging local expertise, namely, by identifying contextual and behavioral knowledge that exists locally but has not been recognized, valued, or integrated into intervention design.

The Rumsfeld Matrix reveals two levers to reduce risk: 1) Increase awareness of assumptions, tacit knowledge, and potential blind spots within the local context and 2) Increase understanding of the questions and hypotheses that support localization.

In the following case study, we outline the uncertainty reduction activities conducted by a behavioral science team to identify risks, improve hypotheses, and document tacit knowledge to improve localization efforts.

Use Case: Digital Health Intervention in Costa Rica

As Lirio continues to expand globally, we are experimenting with approaches for localizing behaviorally

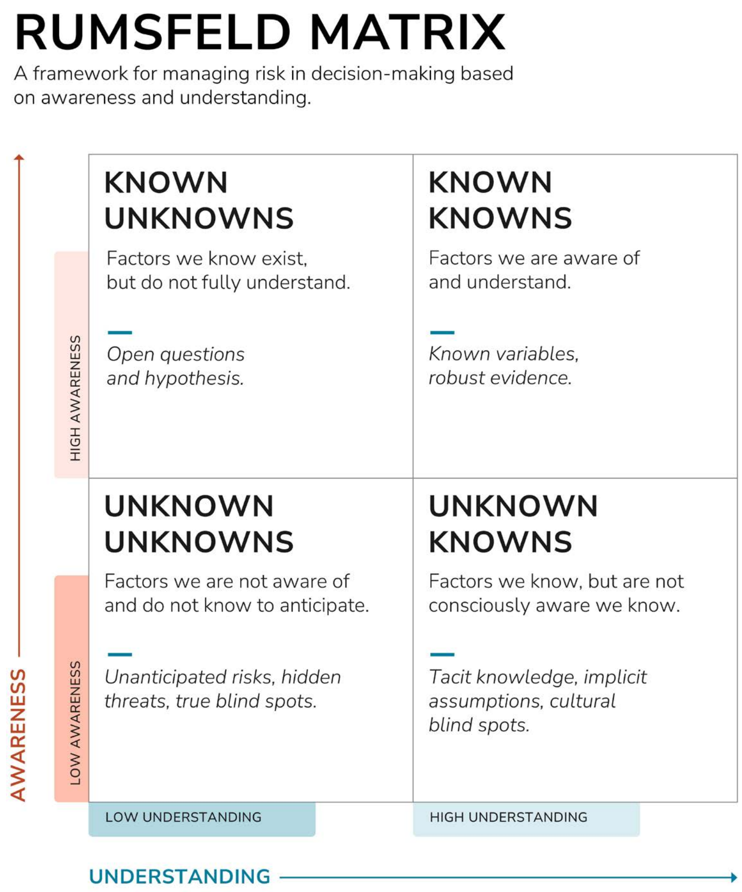


Figure 1: The Rumsfeld Quadrant framework for decision-making. *Note:* Designed by Michael Pancini, Visual Designer, Lirio.

infused digital health interventions. Our team recently conducted a digital health intervention in Costa Rica to increase uptake of a health and wellness program. We drew inspiration for this novel intervention from patterns observed and insights generated across our previously successful digital health interventions, including approaches to actualizing the behavior change techniques (BCTs) in visuals and copywriting.

This program includes a comprehensive health assessment, a 1:1 consultation with a doctor to outline individual health goals, a personalized health improvement roadmap, and discounts on additional healthcare services. Our task was to localize our Precision Nudging approach—which leverages behavioral science and agentic AI to send personalized nudges—to drive enrollment and engagement with this program. We aimed to localize three key components of the intervention: the behavioral science strategy matrix, visual design, and copywriting.

Activities: Increasing Awareness

First, our team conducted standard discovery activities to understand the behavior and context, including reviewing the action path and relevant literature. Next, we leveraged cultural frameworks and structured brainstorming activities to document knowns, identify risks of potential cultural mismatch, and prioritize areas for further investigation. These activities were intended to increase awareness by making implicit assumptions explicit and uncovering potential sources of uncertainty early in the design process.

Cultural Framework Tools

Using a cultural framework, our team compared the United States (where many intervention patterns and assets used for inspiration had previously been developed) and Costa Rica as the target deployment context. Cultural frameworks describe commonly shared patterns of meaning—such as beliefs, norms, values, and traditions—that tend to be more prevalent or salient within a group. Often operating implicitly, these patterns can influence what is perceived as appropriate, motivating, or credible within a given context.

These typologies provide high-level orientation for early-stage scoping, but they neither capture the full diversity of perspectives within a country nor determine individual behavior. A diverse society (such as the US) may encompass multiple cultural frameworks. Used appropriately, cultural frameworks can help reveal contextual features and cultural factors that may influence intervention effectiveness. Practitioners can use insights generated from a cultural framework lens to develop initial hypotheses about where and why an intervention may require tailoring, which can then be validated through local research and system realities.

> "Practitioners can use insights generated from a cultural framework lens to develop initial hypotheses about where and why an intervention may require tailoring."

There are a range of cultural framework tools, each with its own strengths and limitations (Yeganeh, 2025). The most appropriate framework depends on the intervention goals, target behavior, and sources of influence. In this case, our team used Hofstede's Cultural Dimensions Theory, which provides country-level comparisons across six dimensions: Individualism vs. Collectivism; Power Distance; Uncertainty Avoidance; Long- vs. Short-Term Orientation; Indulgence vs. Restraint; and Masculinity vs. Femininity.

Within this framework, the domains with the largest differences between the US and CR included:

Individualism-Collectivism: The US scores high on individualism, while CR scores higher on collectivism. To account for this difference, we evaluated BCT selection, message framing and tone, and visual depictions of people and their interactions. We also considered adjusting collaboration with local partners to incorporate more relational and collaborative working practices.

Masculinity vs. Femininity (Motivation toward Achievement vs. Quality of Life): The US places greater emphasis on goal orientation, achievement, and measurable success, while CR places relatively more value on relationships, collaboration, and work life balance. We therefore explored BCTs with a more prosocial or quality-of-life emphasis, and we reframed goals and success at the family or community level in imagery and copy.

Uncertainty avoidance: CR scores substantially higher than the US on uncertainty avoidance. To account for this, we evaluated BCTs that emphasize clarity, structure, and predictability and identified opportunities to frame the desired action as a simple, step-by-step process in visuals and messaging.

Pre-Mortem Activities

Next, our team conducted a pre-mortem activity to identify potential areas of cultural mismatch that might require additional risk mitigation. A pre-mortem is a strategic exercise in which a team imagines a project has already failed and works backward to identify plausible causes. Here, it was used to document known knowns and known unknowns related to cultural fit.

During the activity, we brainstormed potential failure risks associated with localizing for CR, including possible misalignments exposed through the cultural framework comparison. Using the framework added structure and helped unveil risks that may otherwise have remained implicit, thereby supporting the refinement of the hypotheses to be tested. We then conducted a risk-prioritization exercise to identify the most likely and highest-impact risks, after which we generated risk mitigation strategies for priority items.

Through this step, we identified assumptions requiring validation with local partners. For example, we determined that additional research was needed to localize imagery—particularly around abstract concepts, interpersonal interactions, and body language—and to understand better the perceptions of behaviorally-infused content relative to traditional marketing approaches.

Activities: Increasing Understanding

From the insights gained through the cultural framework lens, and based on the intervention elements we identify as potentially needing tailoring, we identified a series of cultural audit activities to gather additional information. By 'cultural audit activity', we refer to approaches for gathering and documenting collective tacit knowledge (unknown knowns) among those with lived experience in the local context. This included surveys, workshops, and structured activities to explore specific behavioral and intervention elements. The goal of cultural audit activities is to answer targeted questions about the local context to inform discovery and design decisions.

> "By 'cultural audit activity', we refer to approaches for gathering and documenting collective tacit knowledge (unknown knowns) among those with lived experience in the local context."

Image Audit

One cultural audit activity was an image audit, conducted to tailor our existing visual content strategy to the local context. In this activity, local stakeholders reviewed a curated set of images previously used in digital health interventions, including depictions of people, interpersonal interactions, emotional expressions, and abstract representations of health concepts. Stakeholders provided feedback on which images failed to translate, evoked unintended meanings, or lacked resonance.

The image audit helped validate and deepen our understanding of previously identified risk areas. For example, based on the cultural framework, we hypothesized that visuals depicting patient-doctor interactions would benefit from localization. Local stakeholders gravitated toward images in which the patient and doctor were closer in proximity, without a table between them, and in some cases engaging in light physical contact (e.g., handshake, hand on shoulder, shoulder-to-shoulder). They also noted that doctors typically wear scrubs rather than formal attire or white coats. These insights informed updates to better align the visual representations of care interactions with local expectations.

UX Research

In addition, our team conducted quantitative and qualitative UX research to support localization. This research focused on validating hypothesized behavioral determinants related to enrolling in the health and wellness program and identifying additional cultural nuances. Findings informed updates to the strategy matrix, including the prioritization of behavioral barriers, Mechanism of Action (MoA) categorization, and the selection of appropriate BCTs.

As part of this research, we tested behaviorally informed messaging against more traditional marketing approaches to understand perceptions of different framing strategies. We also assessed local communication norms, including perceptions of tone and formality in email communications. For example, respondents expressed high sensitivity to potential email scams, suggesting that hard-selling strategies—such as evoking urgency, emphasizing sales, or using individuals' names—could backfire in this context.

Localizing a Digital Health Intervention for CR

Localizing the Behavioral Science Strategy Matrix

Typically, our team develops a strategy matrix utilizing the TaTT. We categorize and prioritize behavioral determinants into MoAs, then pair them with evidence-linked BCTs. In many cases, a single determinant could plausibly be categorized under multiple MoAs, depending on how it is framed or experienced in context. The content is built with one nudge per MoA×BCT pairing.

To localize the strategy matrix, we leveraged insights from our cultural audits and UX research to add nuance—including reconsidering their classification and splitting broader determinants into more specific MoAs. For example, one identified barrier to enrollment was confusion about what the program is, how it works, and for whom it is intended. At face value, this barrier could be categorized as a knowledge deficit. However, cultural audit findings revealed that this confusion was shaped by prevailing local mental models of hospital-affiliated savings programs. One well-established program that dominated consumer awareness focused on routine health needs. By contrast, the new program focused on proactive wellness behaviors and was designed to complement existing services. Because these programs were often perceived as mutually exclusive rather than complementary, the barrier reflected not only a lack of information, but also a mismatch with existing norms and expectations. To account for this nuance, we incorporated MoAs related to both Knowledge and Norms.

Next, we leveraged cultural insights to inform the selection of MoA-BCT pairings. A single MoA is often linked to multiple evidence-supported BCTs, any of which may be theoretically appropriate. Cultural insights helped us prioritize techniques most likely to be effective in this context. For example, given the relatively collectivistic orientation of the setting, we prioritized socially-oriented BCTs, such as Social

Table 1: Example - Localizing Copy for US-based vs. Costa Rican Audiences

	US	CR
Individualism vs. Collectivism	Often emphasize personal benefits, autonomy, self-improvement, control, and convenience. Use "you," "your progress," "your goals." Highlight independence and personal achievements.	Emphasize family or community benefit, shared well-being, harmony, and collective care. Use "your family," "our community."
Motivation toward Achievement vs. Quality of Life	Use goal-oriented, results-driven, performance-enhancing language: "Reach your goals," "Take charge," "Achieve better health." Celebrate progress and personal wins.	Use warm, modest, quality-of-life language: "Feel better day-to-day," "Care for yourself and others," "Small steps toward well-being." Avoid competitive or aggressive framing.
Uncertainty Avoidance	Ambiguity acceptable. You can use flexible, exploratory language: "Try it," "See what works best for you." Tone can be light and autonomy-inviting.	Emphasize reassurance, clarity, and predictable steps. Use structured, stepwise instructions and signal safety: "Here's exactly what happens next," "Secure," "Approved," "Trusted by your clinic."

Support (Practical). In addition, because CR scores lower on the Long-Term Orientation dimension, we prioritized techniques that foreground short- to moderate-term benefits (e.g., Social Reward) rather than those emphasizing distant future outcomes (e.g., Comparative Imagining of Future Outcomes). This approach did not exclude specific BCTs but instead provided structure and guidance when selecting from multiple viable options.

Localizing Written Copy

We made several adjustments to the copywriting strategy to align better with the identified cultural insights. For example, UX research indicated that messaging framed in relational, supportive terms resonated more strongly than traditional marketing or sales-oriented language, aligning with our hypotheses about tailoring for a more collectivistic context. To account for these dynamics, we shifted toward a friendlier, lower-pressure tone that emphasized credibility, transparency, and approachability.

Our team applied a cultural lens to the copy, making moderate language adjustments. Below are examples of how the same copy might be localized for US-based and CR audiences.

In addition to adjusting message framing, we addressed barriers related to trust and credibility by shifting the call to action from direct enrollment to scheduling a consultation with a team member. This approach acknowledged that decision-making often occurs through interpersonal interaction and relationship building. We also listed the office location within a major hospital to signal credibility and support walk-in visits.

Additionally, we planned for transcreation by ensuring the content could be translated while preserving BCT intent, framing, reading level, and tone. To support this, we wrote in plain language, limited metaphors and idioms, and avoided double meanings. We also reviewed best practices for gender-neutral language in a grammatically gendered language. Finally, we worked with transcreation vendors and local partners to compare English and Spanish versions, using an 'intent to maintain' section to ensure the meaning carried through translation.

Together, these adjustments resulted in subtle but purposeful differences in messaging that felt locally appropriate without leaning into stereotypes. Although we cannot share the exact message for privacy reasons, if we were to rephrase the content for

Table 2: Example Precision Nudges - Localized Copy Across US and Costa Rican Contexts

US	CR
Hi [Name]! You're eligible for our new [Product Name] program. It's designed to help you feel stronger, more energized, and in control of your health. Ready to start your personal plan? Tap here: [link]	Hi [Name]! Your local clinic is offering a new [Product Name] Program to help you and your loved ones stay well. It's easy to follow, reliable, and supported by our care team. Here are the steps so you can get started with confidence: [link]

the same product, localized for these two audiences, they might be written as follows:

Localizing Visual Design

Our team made several adjustments to the visual strategy to align with branding guidelines and incorporate insights uncovered through the cultural audits. This brand is associated with motherly care, support, and playfulness, which the team leveraged through illustration style and color palette choices. Insights from cultural audits and stakeholder feedback helped refine the images. For example, our team noted a preference for visual proximity between the provider and patient, without a table or computer between them, as a signal of a closer, trusted relationship.

We also paid specific attention to ensuring clothing depicted in the visuals reflected CR norms. Patients were shown in a more classic style, rather than the more casual or trendy styles often used with US-based audiences. In addition, doctors were illustrated wearing scrubs rather than white coats to reflect local expectations. Lastly, we incorporated photographs taken within the hospital, featuring team members and the affiliated doctor, in select images to reinforce familiarity and trust.

Table 3 shows how we might think about localizing visuals for these two audiences. Figure 2 is an example of how we might adapt a single image for a US-based versus CR audience.

Discussion: Implications for Scalable, Culturally Nuanced Behavioral Design

As behavioral science tools are applied across increasingly diverse settings, practitioners face a growing challenge: designing interventions that are both scalable and responsive to cultural context. Although evidence-based tools make intervention design more systematic and replicable, they often require additional interpretation when applied in new cultural contexts. This gap can create both effectiveness and equity challenges, placing a disproportionate burden on practitioners working in contexts historically underrepresented in the literature.

This case study illustrates the value of structured approaches to support localization decisions while preserving theoretical foundations. In this example, cultural frameworks helped reveal potential cultural

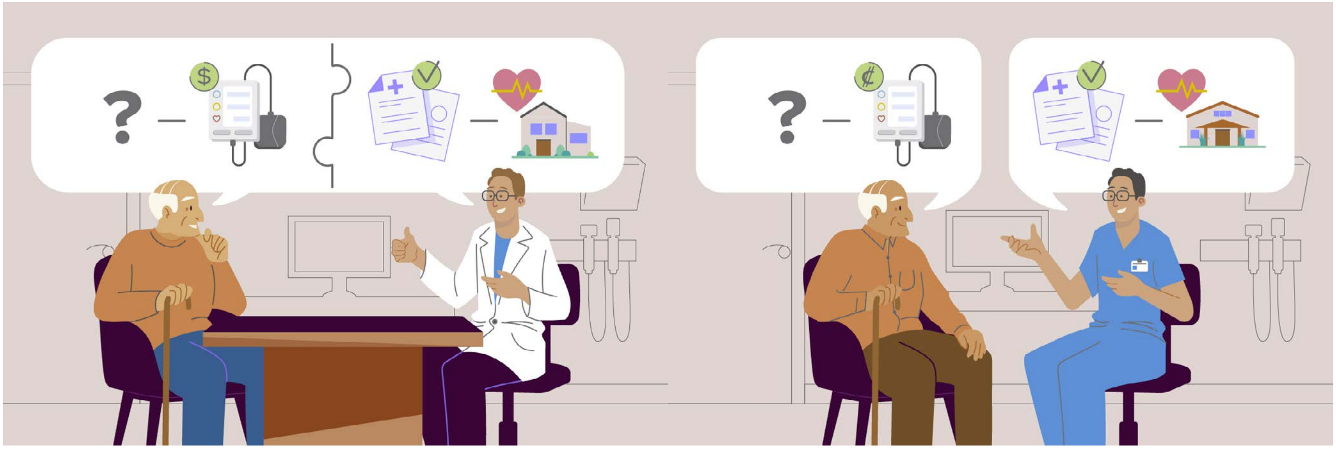


Figure 2: Example of localized digital health visuals. *Note:* Designed by Michael Pancini, Visual Designer, Lirio.

Table 3: Visual Content Strategy for US-Based and Costa Rican Audiences

	US	CR
Individualism vs. Collectivism	- Individuals performing the target behavior (e.g., one person exercising) - Often emphasize personal progress, self-care, and autonomy - Use solo portraits or close-ups that highlight the individual as the agent of change or decision-maker	- Small groups or families performing the behavior together - Depict belonging and togetherness (e.g., playing outside with family, a couple cooking) - Often emphasize shared well-being, support, and interdependence - Express relational warmth (e.g., proximity, eye contact, and gentle social cues)
Motivation toward Achievement vs. Quality of Life	- Dynamic, energetic imagery (e.g., finishing a run, celebrating) - "before/after" styles that suggest improvement or transformation. - Visual cues of progress (e.g., graphs, trackers, badges) - Colors and compositions that feel bold, active, and forward-moving - "Aspirational" and performance-oriented imagery	- Calm, balanced visuals emphasizing well-being (e.g., stretching, lifting child in air) - People engaging in everyday, sustainable activities (e.g., cooking, walking in nature) - Soft, warm color palettes - Natural settings that feel relatable - Depict process rather than outcome—emphasizing gentle progress and harmony
Uncertainty Avoidance	- Conceptual, abstract visuals - Symbolic or metaphorical imagery - Emphasis on novel or modern health technologies	- Imagery that reinforces predictability, such as clear process visuals (e.g., steps in a process) - Icons with clear, unambiguous meaning - Trusted figures (e.g., nurses, doctors, community health workers) - Familiar settings (clinics, homes, family environments).

differences that warranted further investigation, while cultural audit activities grounded these hypotheses in local tacit knowledge.

Building on this logic, future innovations may further account for cultural variability by combining group-level cultural insights with individual-level adaptation. Approaches that personalize communications at an individual level—such as Lirio's Precision Nudging, which applies behavioral science and agentic AI to create an N=1 messaging experience—may offer one way to address variation that exists both across and within cultural contexts.

"Localization risks cannot be eliminated, but they can be made more visible, examinable, and actionable."

Localization risks cannot be eliminated, but they can be made more visible, examinable, and actionable. This use case demonstrates how practitioners can complement existing evidence-based tools with structured risk management activities to localize behavioral science interventions in cases where evidence does not travel cleanly across contexts.

THE AUTHOR

Marisa Nowicki is a behavioral designer at Lirio, where she applies behavioral science to digital health interventions in real-world settings. She is a core contributor to Lirio's Precision Nudging approach, which combines behavioral science and agentic AI to deliver timely, personalized nudges. In her work, Marisa supports the translation of theory into practice by shaping intervention strategy and collaborating with cross-functional teams to create deployable solutions. Prior to Lirio, she worked at ideas42 as a behavioral designer on the Global Development team, with a portfolio spanning human rights, agriculture, and sustainability.

REFERENCES

Castro, O., Fajardo, G., Johnston, M., Laroze, D., Leiva-Pinto, E., Figueroa, O., Corker, E., Chacón-Candia, J. A., & Duarte, G. (2024). Translating the Behaviour Change Technique Taxonomy version 1 into Spanish: Methodology and validation. *Wellcome Open Research, 9*, 298. https://doi.org/10.12688/wellcomeopenres.21388.1

Catoja, S., & Crede, A.-K. (2025). The behaviorally-informed insurance company: Insights from the Allianz Global Center for Behavioral Economics. In A. Samson (Ed.), *The Behavioral Economics Guide 2025* (pp. 48-54). Behavioral Science Solutions. https://www.behavioraleconomics.com/be-guide/the-behavioral-economics-guide-2025/

Hall, G. C. N., Ibaraki, A. Y., Huang, E. R., Marti, C. N., & Stice, E. (2016). A meta-analysis of cultural adaptations of psychological interventions. *Behavior Therapy, 47*(6), 993-1014. https://doi.org/10.1016/j.beth.2016.09.005

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? (RatSWD Working Paper No. 139). *EconStor.* http://hdl.handle.net/10419/43616

Johnston, M., Carey, R. N., Connell Bohlen, L. E., Johnston, D. W., Rothman, A. J., De Bruin, M., ... & Michie, S. (2021). Development of an online tool for linking behavior change techniques and mechanisms of action based on triangulation of findings from literature synthesis and expert consensus. *Translational Behavioral Medicine, 11*(5), 1049-1065. https://doi.org/10.1093/tbm/ibaa050

List, J. (2021). The voltage effect in behavioral economics. In A. Samson (Ed.), *The Behavioral Economics Guide* 2021 (pp. VI–XI). Behavioral Science Solutions. https://www.behavioraleconomics.com/be-guide/the-behavioral-economics-guide-2021/

Schwartz, S. H. (2012). An overview of the Schwartz theory of basic values. *Online Readings in Psychology and Culture*, 2(1), 11. https://doi.org/10.9707/2307-0919.1116

Talhelm, T. (2026). Adapting interventions to culture can improve effectiveness and cost-efficiency. *Policy Insights from the Behavioral and Brain Sciences, 13*(1), 61-68. https://doi.org/10.1177/23727322251408076

The Uncertainty Team. (n.d.). *Rumsfeld matrix.* The Uncertainty Project. https://www.theuncertaintyproject.org/tools/rumsfeld-matrix

World Economic Forum. (2024, April 19). *How the 'NO, NO' Matrix can help professionals overcome failure.* https://www.weforum.org/stories/2024/04/how-the-no-matrix-can-help-professionals-overcome-failure/

Yeganeh, H. (2025). Hofstede, GLOBE, Schwartz and Inglehart's cultural dimensions: A critical examination and assessment. *European Business Review, 37*(6), 1073–1090. https://doi.org/10.1108/EBR-05-2025-0144

Extended Review: Misinformation in the Attention Economy

LIZA BERGMAN-GEONIE[1]

The Chicago School

Click or scan QR code
to watch video summary

Misinformation spreads and persists in part because digital environments are designed to prioritize engagement over accuracy. Social media platforms promote content that elicits quick reactions, often amplifying emotionally engaging, novel, or belief-aligned material, regardless of its credibility. In these settings, sharing is fast and easy and tends to produce immediate feedback, while verifying information takes more effort and often carries fewer immediate consequences. Research shows that false information can spread farther and faster than true information, and processes such as repetition and reliance on mental shortcuts contribute to its persistence over time. Although strategies like digital literacy training and fact-checking can improve people's ability to evaluate information, they do not always lead to changes in what people share. Addressing misinformation requires changes to the environments in which sharing occurs, including making verification easier and increasing the visibility and reinforcement of accurate information.

Introduction

According to a 2025 Pew Research study, 54% of U.S. adults get their news from social media, with similar figures reported in the United Kingdom (Milmo, 2025). These figures underscore the growing role of social media in shaping how news and information are consumed—a shift that has been linked to declining trust in journalism and investigative reporting (Suherman, 2025). Recent advances in artificial intelligence have intensified these concerns, as deepfakes and other computer-generated content are increasingly difficult to distinguish from authentic media (Giansiracusa, 2021). Together, these developments have created conditions in which misinformation spreads rapidly across digital platforms (Vosoughi et al., 2018).

Researchers define misinformation as the unintentional spread of false information, whereas disinformation is intentionally misleading content (Diepeveen & Pinet, 2022; McKay & Tenove, 2020). Although misinformation is not new, the conditions under which it spreads have changed substantially in recent times; before the internet, television, newspapers, and radio shaped how information was produced and distributed, and while sensational content did exist, its spread was constrained by time, cost, and access. Historical developments such as the printing press and the rise of the penny press expanded access to information while also increasing the circulation of sensational and misleading content (Soll, 2016; Watson, 2018). However, these systems imposed structural friction, including editorial oversight, limited distribution channels, physical production constraints, and restricted audience participation, all of which slowed the spread of misinformation (Posetti & Matthews, 2018). Social media has removed many of these constraints, allowing individuals to publish

1 lgeonie@gmail.com

and distribute content globally with minimal effort (Epstein et al., 2023). These distinctions are important, but both misinformation and disinformation disperse within the same environments that shape how information is shared and reinforced.

Misinformation, along with related concepts such as disinformation and malinformation—defined as the intentional use of truthful information in misleading ways (Santos-d'Amorim & Fernandes de Oliveira Miranda, 2021)—has become a central focus of research. Early theoretical work anticipated current conditions. Arendt (1972), for instance, described "defactualization" as the erosion of the boundary between fact and fiction, while Rothkopf (2003) coined the term "infodemic" to capture the rapid spread of mixed-quality information during public health crises. These patterns are now amplified in digital environments, where information circulates quickly across interconnected networks with few barriers to sharing. Users encounter large volumes of competing content, including emotionally charged and non-news material, making it harder to distinguish credible information from misinformation (Epstein et al., 2023). For example, a user may encounter a compelling headline, share it immediately, and receive instant feedback through likes or replies before any verification occurs. These platforms also foster echo chambers, where users are more likely to encounter belief-consistent content, thus reducing opportunities for verification and increasing vulnerability to misleading information (Adams et al., 2023). At the same time, platform design features such as algorithmic amplification, ease of content creation, and global connectivity increase the visibility of engaging content regardless of accuracy (Posetti & Matthews, 2018). In some cases, verification has shifted from centralized systems to decentralized approaches, where corrections occur only after content has already spread widely, thereby limiting their effectiveness (Lewandowsky et al., 2012; Epstein et al., 2023). Furthermore, during breaking news events, inaccurate claims can circulate rapidly before verification is complete, limiting the impact of later corrections and contributing to continued distrust in news sources.

These patterns operate within an "attention economy" (Browne & Watzl, 2026), which serves as the broader context shaping how information is selected, amplified, and shared. Digital platforms are designed to capture and sustain attention by prioritizing content based on engagement signals such as likes, shares, comments, and viewing time. Features such as infinite scrolling, push notifications, and social feedback loops continuously draw attention to new content, and because attention is monetized, content that generates rapid, sustained engagement is more likely to be promoted. Under these conditions, misinformation is well-positioned to spread; moreover, novel, emotionally engaging, or belief-aligned content is more likely to attract attention and be shared quickly. As a result, digital environments systematically favor low-effort, attention-grabbing content over accuracy. Within this system, behaviors that produce immediate engagement are consistently selected, while more effortful accuracy-based behaviors are less likely to occur. These findings indicate that misinformation is less a failure of individual judgment and more a predictable outcome of environments that reward speed, engagement, and low-effort responding over accuracy.

Why Misinformation Spreads

Misinformation spreads quickly online because it aligns with how people process, share, and reinforce information. In digital environments, users are constantly exposed to fast-moving streams of content, prompting rapid decisions about what to attend to and what to ignore. To manage this volume, individuals rely on mental shortcuts, or heuristics (Chen et al., 2021). These shortcuts prioritize speed over accuracy, heightening sensitivity to cues such as repetition, social consensus, and emotional appeal. At the same time, digital platforms are structured to reward engagement. Algorithms prioritize content that generates likes, shares, comments, and longer viewing times, creating an environment in which visibility is determined by interaction rather than accuracy.

As shown in Figure 1, a substantial share of adults regularly get news from social media platforms such as Facebook, YouTube, and TikTok, reflecting a broader shift in how information is encountered in digital environments (Atske, 2025). This shift places news consumption within systems that reward rapid engagement, where emotionally salient and easily

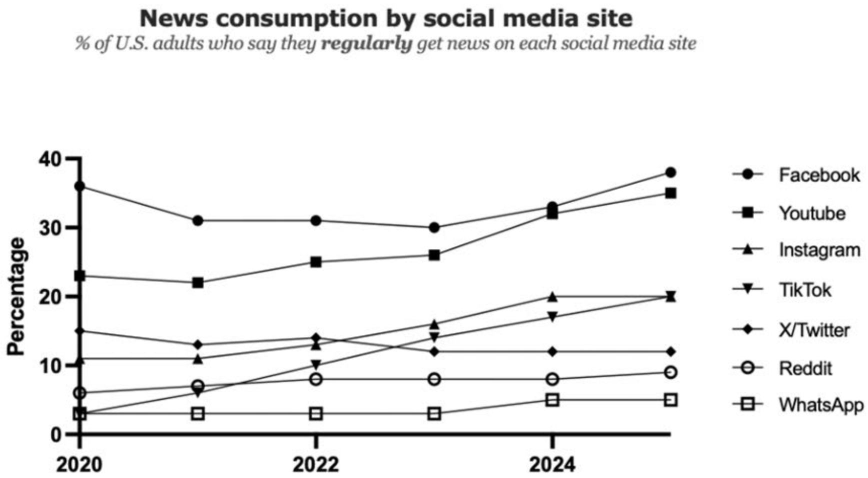


Figure 1: News consumption by social media platform among U.S. adults (2020–2025). A growing proportion of adults report regularly obtaining news from social media, particularly on Facebook, YouTube, and TikTok (Atske, 2025). This shift reflects the increasing role of engagement-driven environments in shaping how information is encountered, placing news consumption within systems that prioritize speed, visibility, and user interaction over verification. *Source:* Pew Research Center (2025).

processed content is more likely to be shared than carefully evaluated information. As a result, the same environments that structure everyday information use also create conditions that facilitate the spread of misinformation.

> "In digital environments, users are constantly exposed to fast-moving streams of content, prompting rapid decisions about what to attend to and what to ignore."

Because misinformation is often novel, emotionally compelling, or aligned with existing beliefs, it is well positioned to capture attention and spread rapidly. In some cases, content is deliberately designed to provoke strong reactions, particularly outrage—a strategy often referred to as "rage-baiting" (Washburn, 2024). Importantly, engagement does not require agreement. Content that challenges or contradicts beliefs can still generate high levels of interaction as users respond with criticism, correction, or disapproval. These responses serve as engagement signals, increasing visibility within platform algorithms. As illustrated in Figure 2, this creates a reinforcement loop in which engagement increases visibility, and increased visibility produces further engagement. This pattern is evident in real-world cases. During the spread of the "Pizzagate" conspiracy theory, for example, sensational claims circulated rapidly across online forums and social media. Users shared the content widely and received immediate social reinforcement through likes, comments, and continued dissemination, while the speed and volume of these interactions allowed the claims to circulate broadly before verification could occur. In this case, the consequences extended beyond online engagement, culminating in an individual entering the restaurant with a rifle to investigate the claims, demonstrating how rapidly reinforced misinformation can translate into real-world behavior (Robb & Robb, 2017).

Empirical research supports these patterns. Vosoughi et al. (2018) found that false information spreads farther, faster, and more broadly than true information, driven in part by novelty and emotional engagement. Users also frequently engage with headlines rather than full articles—a pattern known as "sharing without clicks," whereby content is shared based on superficial cues rather than verified understanding (Sundar et al., 2024). In addition, deciding whether to share content can reduce attention to accuracy. Epstein et al. (2023) found that individuals are less able to distinguish true from false information when making sharing decisions than when explicitly evaluating accuracy. This shift in attention toward engagement-related features increases the likelihood that misleading content will be shared. Emotional and social dynamics further accelerate these processes. Additionally, content that evokes strong reactions, such as fear, outrage, or humor, is more likely to be shared as a form of identity expression and social connection. These patterns are reinforced within networks, especially when content aligns with group norms or beliefs. Algorithmic systems amplify this effect by promoting

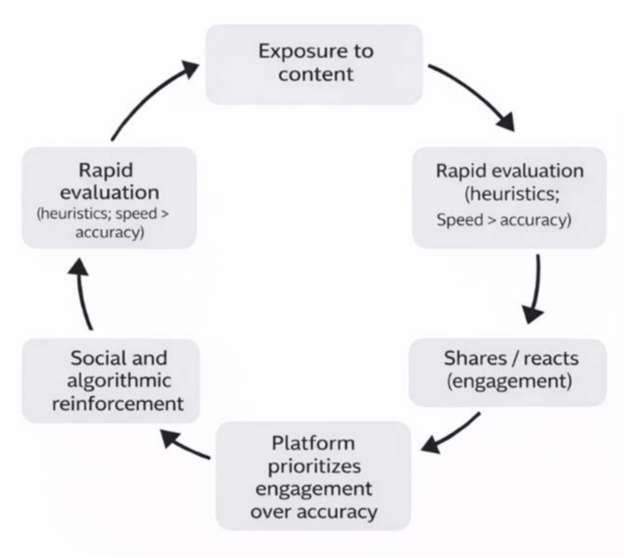


Figure 2: Reinforcement loop of information-sharing in digital environments. Exposure to content leads to rapid, heuristic-based evaluation, followed by engagement behaviors such as sharing and reacting. These behaviors generate immediate social and algorithmic reinforcement, increasing visibility and exposure, and creating a self-reinforcing cycle in which engagement is prioritized over accuracy.

highly engaging content to broader audiences, thereby extending its reach. At the same time, cognitive processes contribute to persistence. Individuals tend to favor belief-consistent information and discount contradictory evidence (Westerwick et al., 2017; Zhou & Shen, 2021), repetition increases perceived truth through the illusory truth effect, and misinformation can continue to influence reasoning even after correction—a phenomenon known as the "continued influence effect" (Swire-Thompson et al., 2021).

> "Even after correction, misinformation often persists because it becomes embedded in how people process and remember information, reinforced by mental shortcuts and repeated exposure."

At a structural level, misinformation is often easier to produce than accurate information. Barriers to entry in digital media are low, allowing individuals to create and distribute content quickly without fact-checking or editorial oversight (Allcott & Gentzkow, 2017). Misleading content can be generated by modifying headlines, exaggerating claims, or presenting opinion as fact. These tactics often require minimal time and expertise while maintaining an appearance of credibility. Coordinated amplification strategies, such as using multiple accounts to promote the same message, can further extend reach without increasing production effort. Advances in generative technologies have additionally reduced production effort by enabling the rapid creation of realistic but false text, images, and videos (Arribas et al., 2025; Waldrop, 2023). In contrast, producing accurate information typically requires greater time, expertise, and resources. These differences in effort and reinforcement shape behavior. Sharing content requires minimal effort and produces immediate, visible feedback, whereas verifying accuracy demands additional time and cognitive resources and offers less immediate reinforcement. As a result, sharing is more likely than careful evaluation. Finally, digital environments prioritize speed, engagement, and emotional impact, creating conditions in which sharing is consistently reinforced over accuracy. The processes summarized in Table 1 operate within and strengthen the reinforcement cycle shown in Figure 2.

Why Misinformation Sticks

Correcting misinformation is more challenging than simply providing accurate facts. Even after correction, misinformation often persists because it becomes embedded in how people process and remember information, reinforced by mental shortcuts and repeated exposure. A well-established example of this persistence is the continued influence effect. Research shows that individuals often continue to rely on misinformation even after it has been corrected (Lewandowsky et al., 2012; Swire-Thompson et al., 2021), in that once information becomes part of an individual's mental model of an event, it continues to shape reasoning and interpretation. For example, early public health messaging emphasized dietary fat as the primary cause of cardiovascular disease. Although later evidence highlighted the role of sugar and refined carbohydrates, the initial explanation continued to influence public understanding and behavior (Kearns et al., 2016). This illustrates how initial information can persist within mental models and continue to guide interpretation even after correction.

Another process that sustains misinformation is the illusory truth effect. Repeated exposure to a claim increases the likelihood that it will be perceived as true (Fazio et al., 2015). The brain uses processing fluency as a heuristic, so information that feels familiar or easy to process is more likely to be judged as accurate. For example, repeated claims linking egg consumption to cardiovascular risk contributed to widespread public belief, even though later research showed a more limited effect for most individuals (Rong et al., 2013). This pattern demonstrates that repetition alone can increase perceived accuracy, even in the absence of strong supporting evidence.

Credibility judgments are often shaped by these surface-level cues, especially when individuals rely on heuristics rather than detailed evaluation. Importantly, knowledge alone does not fully protect against these effects, and research reveals that even individuals with domain expertise remain susceptible to repeated misinformation, likely because familiarity and fluency can override analytic processing (Fazio et al., 2015). These effects are amplified in digital environments, where algorithms, echo chambers, and continuous scrolling increase repeated exposure to the same claims, thereby strengthening familiarity and perceived credibility.

These dynamics also shape how individuals rely on external information. People may depend on information generated or summarized by automated systems without independent verification—a pattern consistent with automation bias (Dorrell & Heal, 2024). When information is presented in formats associated with expertise, such as scientific writing or authoritative AI-generated outputs, perceived credibility can increase, reflecting authority bias. A recent example illustrates this process. Almira Osmanovic Thunström created a fictional medical condition, "Bixonimania," and introduced it into online and academic spaces (Stokel-Walker, 2026). AI systems subsequently generated responses that treated the condition as legitimate, and these outputs were later cited in peer-reviewed research. This case demonstrates how reliance on automated systems as authoritative sources can lead to the reproduction and dissemination of false information across both AI-generated and scientific channels. In this way, automation bias and authority cues contribute to the persistence of misinformation by reinforcing trust in system outputs, even when those outputs are inaccurate.

Taken together, these processes reflect a set of cognitive and behavioral mechanisms that sustain misinformation over time. As summarized in Table 1, they interact with platform features and reinforcement structures, shaping how information is evaluated, accepted, and maintained in digital

Precommitment Pledge:

I pledge my earnest efforts to share truth by:

- Verifying: fact-check information to confirm it is true before accepting and sharing it
- Balancing: share the whole truth, even if some aspects do not support my opinion
- Citing: share my sources so that others can verify my information
- Clarifying: distinguish between my opinion and the facts

Signature

Figure 3: Pre-commitment pledge adapted from Gleb Tsipursky's Pro-Truth Pledge, used to encourage participants to verify information and engage in responsible sharing prior to posting on social media.

Table 1: Cognitive and Behavioral Processes Contributing to the Spread and Persistence of Misinformation in Digital Environments

Process	What it Does	Where it Appears in Digital Environments	Effect on Misinformation
Confirmation bias	Skews attention toward belief-consistent information	Echo chambers, selective exposure, group-aligned content	Increases acceptance and sharing of belief-consistent misinformation
Illusory truth effect	Increases perceived accuracy through repetition	Algorithmic repetition, viral reposting, continuous feeds	Familiar misinformation is judged as more credible
Continued influence effect	Maintains influence of misinformation after correction	Delayed corrections, ongoing exposure, weak retractions	False information continues to shape reasoning
Heuristic processing	Promotes fast, low-effort judgments over careful evaluation	High-volume feeds, time pressure, headline scanning	Reduces attention to accuracy and increases susceptibility to misleading content
Automation bias	Shifts trust toward automated or AI-generated outputs	AI summaries, search engines, generated content	Users accept inaccurate information without verification
Authority bias	Increases reliance on perceived expertise or credibility cues	Scientific tone, professional formatting, AI outputs	Increases acceptance of misleading information
Social reinforcement	Strengthens sharing through immediate social feedback	Likes, shares, comments, visibility metrics	Encourages continued sharing regardless of accuracy
Algorithmic amplification	Increases exposure based on engagement signals	Ranking systems, recommendation engines	Expands reach of high-engagement misinformation

Note: These processes interact with platform features and reinforcement structures to shape how information is evaluated, shared, and sustained.

environments.

Evidence From Applied Research

Applied research on the spread of misinformation on social media platforms remains limited, particularly in behavior-analytic literature. Interventions such as pre-commitment strategies and direct instruction in digital literacy aim to promote responsible sharing by increasing awareness, reinforcing rules, and improving evaluation skills. Although these approaches can influence certain behaviors, their impact is often constrained by the environments in which they are applied.

A study conducted on Facebook (Bergman-Geonie et al., 2025) required participants to engage with social media in ways that promoted accuracy verification when sharing news content. The pre-commitment component, adapted from Gleb Tsipursky's Pro-Truth Pledge and shown in Figure 3, required participants to pledge to verify information before posting (Tsipursky et al., 2018). The direct-instruction component provided explicit training in evaluating sources using

Figure 4: The SIFT method (Stop, Investigate, Find, Trace), a structured approach to evaluating online information that prompts users to pause, assess sources, seek better coverage, and trace claims back to their original context before sharing (Caulfield, 2019).

structured strategies.

These interventions targeted rule-based behavior and aimed to increase the likelihood that participants would select and share more reputable information. A nonconcurrent multiple-baseline design across participants was used to examine changes in posting behavior across the baseline, pre-commitment, training, post-training, and follow-up phases. The direct-instruction phase, shown in Figure 4, included a five-module training program based on the SIFT method—Stop, Investigate, Find, and Trace—which prompts users to pause, evaluate sources, seek better coverage, and trace claims back to their original context before sharing (Caulfield, 2019; Brodsky et al., 2021).

Results indicated that direct instruction improved participants' ability to distinguish between higher- and lower-quality sources, demonstrating acquisition of evaluation skills. Participants also reported increased confidence in identifying reputable sources, reflected in consistently high social validity ratings; however, these improvements did not translate into changes in actual posting behavior, and they continued to share content in line with their typical engagement patterns, even though they were able to identify more credible information. Reputable posts represented only a small proportion of total posts, with minor and variable increases across phases. Similarly, the pre-commitment strategy did not produce consistent behavioral change. Although participants agreed to follow accuracy-based rules, these commitments were insufficient to override the immediate and reinforcing consequences of typical social media use. The pre-commitment phase produced modest short-term shifts in posting behavior, but these effects were not sustained. The direct-instruction phase was also associated with a decrease in overall posting and did not increase the proportion of reputable content shared.

> "Individuals can learn to evaluate information accurately and endorse rules that support responsible sharing, yet they still fail to act on them when the environment continues to reinforce low-effort, high-engagement behaviors."

These findings highlight a clear gap between discrimination and performance. Participants learned to evaluate information and expressed intentions to share responsibly, yet their behavior remained largely unchanged in real-world settings. One explanation is that the intervention did not alter the environmental contingencies that surround sharing. Evaluating sources requires additional steps, whereas sharing produces immediate social and algorithmic reinforcement. Under these conditions, behavior continues to favor rapid responding over accuracy. More broadly, these findings highlight a gap between knowledge and action. Individuals can learn to evaluate information accurately and endorse rules that support responsible sharing, yet they still fail to act on them when the environment continues to reinforce low-effort, high-engagement behaviors. In digital systems, likes, visibility, and social feedback serve as immediate reinforcers, while the consequences of sharing inaccurate information are delayed or less salient. As a result, engagement-based reinforcement outweighs accuracy-based outcomes, making rapid sharing more likely than careful evaluation.

Applied research suggests that interventions focused on knowledge, rules, or commitment are unlikely to produce lasting behavioral change unless they also modify the decision-making environment. This includes platform features such as easily accessible sharing functions, visible engagement metrics, and continuous content feeds. Gains in knowledge and rule adherence do not consistently translate into sustained behavior change when these features continue to reinforce rapid sharing. Research further indicates that the social media context itself can interfere with accuracy judgments, as attention is drawn to engagement-related features rather than to evaluation (Epstein et al., 2023). Design elements such as infinite scrolling and algorithmic systems that prioritize highly engaging content increase exposure to attention-grabbing material, regardless of accuracy, while limiting opportunities for reflection or verification (Pennycook et al., 2020).

When Strategies Break Down

Many commonly recommended strategies for reducing misinformation focus on improving individual decision-making to promote more deliberate processing and the better assessment of credibility. Although these approaches are effective in controlled settings, their impact is often diminished in real-world digital environments, where conditions favor rapid, low-effort responses.

One reason these strategies often fail is the effort they require. Evaluating information involves additional steps, such as opening new sources, cross-referencing claims, and reading beyond headlines (Brodsky et al., 2021). In contrast, sharing content requires minimal effort and can be done with a single click. Research shows that behaviors requiring greater effort are less likely to occur, especially when easier alternatives are available (Friman & Poling, 1995; Perry & Fisher, 2001; Wilder et al., 2020). On social media, users often process content superficially, with over 75% of shared posts distributed without clicking to read the full article (Sundar et al., 2024). In these environments, accuracy-based strategies must compete with faster, lower-effort responses.

“Evaluating information involves additional steps, such as opening new sources, cross-referencing claims, and reading beyond headlines. In contrast, sharing content requires minimal effort and can be done with a single click.”

A second factor is the timing and strength of consequences. Sharing content yields immediate reinforcement through likes, shares, and comments, whereas accurate evaluation produces delayed or less visible outcomes. This imbalance favors behaviors that yield immediate feedback. Platform features such as social feedback loops, push notifications, and infinite scrolling further reinforce these patterns by continuously prompting engagement (Giansiracusa, 2021). In an attention economy that rewards speed, visibility, and reaction, accuracy becomes the less reinforced option.

A third limitation concerns the durability of rule-based strategies. Instructions such as “verify before sharing” serve as rules that guide behavior, but their influence weakens when they compete with responses that are consistently reinforced. Research indicates that initial rules can persist even when more accurate rules are introduced, especially when those rules are supported by social reinforcement (Dixon et al., 2000; Harte et al., 2020). In online environments, belief-consistent content is repeatedly reinforced, increasing the likelihood that individuals follow socially supported patterns rather than accuracy-based rules. These dynamics operate within the reinforcement cycle shown in Figure 2, where immediate feedback and algorithmic amplification strengthen rapid responding over accuracy.

These patterns are embedded in broader platform systems, where algorithms prioritize engagement metrics such as clicks, shares, and watch time. As a result, emotionally engaging or belief-consistent content is amplified regardless of accuracy (Ahadzadeh et al., 2021; Bourne, 2023). Effective solutions therefore extend beyond individual users and require changes to platform design, industry practices, and policy frameworks that shape these reinforcement contingencies. Furthermore, because these systems are intentionally designed and maintained by social media companies, responsibility for meaningful change lies largely with those who control the structure of these environments and with policymakers shaping regulatory constraints (Lazer et al., 2018; Pennycook & Rand, 2022). Without modifications to these environmental conditions, interventions targeting individual behavior alone are unlikely to produce sustained effects, particularly given evidence that platform contexts can interfere with accuracy-based judgments (Epstein et al., 2023).

Even platform-level interventions exhibit limited effects when implemented without broader changes to the underlying engagement structures. Third-party fact-checking programs, including those previously used by Meta, can reduce belief in false information under specific conditions but have more limited and inconsistent effects on sharing behavior in large-scale environments (Nyhan & Reifler, 2015; Pennycook et al., 2020). More recent approaches, such as crowd-based systems like Community Notes on X (formerly Twitter), aim to provide contextual corrections within the platform. Research indicates that aggregated user input can reliably identify misleading content and reduce its perceived credibility (Allen et al., 2024). However, these interventions produce more limited and context-dependent effects on behavior, including sharing, due to constraints related to visibility, timing, and the broader platform environment (Allen et al., 2024; Pennycook et al., 2020; Epstein et al., 2023). Together, these findings suggest that while such interventions can influence evaluation, they do not consistently alter behavior when reinforcement structures continue to prioritize engagement.

Under conditions of high cognitive load, time pressure, and emotional arousal, individuals are more likely to rely on rapid, low-effort responses rather than on careful evaluation. And even when individuals possess the skills to assess information, these behaviors must compete with faster, more strongly reinforced alternatives. Social media platforms, designed for continuous interaction, are often used passively and within limited time frames, further increasing reliance on rapid responses. As a result, strategies that depend on deliberate processing are less likely to be applied consistently in real-world settings.

This pattern aligns with Lewin's (1951) formulation of behavior as a function of the individual and the environment (B = f(P, E)). Lewin's model emphasizes that behavior emerges from the interaction between personal abilities and environmental conditions at the moment of action. In digital environments, behavior is shaped by embedded cues such as notifications, alerts, message indicators, and algorithmically timed prompts that serve as immediate triggers for engagement. Even when individuals can accurately evaluate information, these behaviors must compete with a continuous stream of salient, low-effort prompts that prioritize speed and interaction. Under these conditions, environmental variables exert strong control over behavior, often overshadowing more effortful, accuracy-oriented responses. Without changes to these environmental conditions, sustained behavior change is unlikely.

> "Under conditions of high cognitive load, time pressure, and emotional arousal, individuals are more likely to rely on rapid, low-effort responses rather than on careful evaluation"

Social Reinforcement and Identity

Social reinforcement plays a central role in the spread of misinformation, especially in digital environments where peer feedback and group dynamics continuously shape behavior. On social media, actions such as posting, liking, sharing, and commenting produce immediate social effects, including approval, visibility, and validation. These effects are amplified by platform design features that prioritize engagement, as well as by content strategies used by creators, influencers, and media outlets to attract and maintain attention. For example, posts are often structured to increase interaction by using ambiguous or incomplete information, emotionally charged language, or formatting that requires additional effort to interpret. These strategies increase time spent on content and the likelihood of engagement. As a result, engagement is not simply driven by user preference but is actively shaped by platform design and content practices that promote interaction, sometimes at the expense of accuracy or clarity. These processes operate within the reinforcement cycle shown in Figure 2, where social feedback and algorithmic amplification increase the visibility of highly engaging content.

Algorithmic systems further amplify these patterns by promoting content that drives high engagement.

Posts that provoke strong reactions, especially those that evoke emotion, are more likely to be boosted and shared with broader audiences (Ahadzadeh et al., 2021), which creates a feedback loop in which socially reinforced content gains visibility, leading to further reinforcement as exposure increases. Internal platform data illustrate this effect. For example, Facebook's introduction of reaction features such as "love," "wow," and "angry" increased the algorithmic weighting of emotionally charged engagement, contributing to the spread of divisive and lower-quality content (Oremus et al., 2021; Bakir & McStay, 2023). These forms of engagement serve as reinforcement signals within the system, increasing visibility and further exposure (see Figure 2).

These dynamics intensify within belief-aligned communities, where engagement is shaped by both user preferences and platform-level amplification. Although individuals may be less likely to intentionally seek out information that conflicts with their beliefs, they are frequently exposed to such content through algorithmic distribution and network structures. Importantly, exposure does not require agreement to generate interaction; users often engage with opposing content through comments, shares, or reactions such as anger, all of which increase visibility regardless of accuracy (Westerwick et al., 2017; Zhou & Shen, 2021). Within these networks, repeated exposure increases perceived credibility, while social signals indicate that content is accepted and endorsed by others. Over time, these patterns strengthen group norms and reduce the likelihood that alternative viewpoints will be considered. They also reinforce identity expression, as engagement signals indicate which beliefs and positions are supported within a group.

On a broader scale, these processes produce network-level effects in which engagement-driven reinforcement increases the visibility of belief-consistent content while limiting the spread of opposing viewpoints. As a result, content aligned with group norms receives more interaction and amplification, thereby contributing to polarization and the continued spread of misinformation across digital networks.

Designing Better Systems

Misinformation persists because digital environments reinforce rapid sharing, emotional salience, and identity signaling more reliably than accuracy checking (Vosoughi et al., 2018; Pennycook & Rand, 2019). Under these conditions, interventions that rely solely on awareness or instruction compete with systems that reward speed and engagement. These systems prioritize clicks, shares, and comments, amplifying emotionally salient and novel content regardless of accuracy (Vosoughi et al., 2018; Lewandowsky et al., 2017). When engagement is the primary reinforcement signal, low-effort sharing becomes an efficient behavior in the attention economy (Simon, 1971; Bago et al., 2020).

As illustrated in Figure 5, effective approaches must alter the contingencies surrounding sharing decisions by reducing the effort required to verify, increasing the salience of credibility cues, and introducing more immediate consequences for accuracy-oriented responses (Pennycook et al., 2021; Ecker et al., 2022). These implications apply directly to platform architecture, where modifying effort, feedback, and visibility can shift behavior from rapid sharing to more accurate information use.

Platforms can support accuracy-focused responses by implementing safeguards before sharing, promoting source evaluation, and emphasizing credibility indicators. Following major political events such as the 2016 U.S. presidential election and the Brexit referendum, platforms such as Meta (Facebook and Instagram) implemented and expanded third-party fact-checking services amid increased pressure to address misinformation (Lazer et al., 2018; Swire-Thompson & Lazer, 2020). In response, companies introduced warning labels, disclaimers, links to third-party verifications, and the removal of clearly false content. These efforts initially suggested that fact-checking could become a central feature of major platforms; however, they produced mixed and often short-lived effects. Concerns about censorship, perceived bias, and platform governance fueled resistance, and the focus gradually shifted from centralized fact-checking to user-driven evaluation and algorithm-mediated corrections, culminating in Meta's January 2025 decision to end third-party fact-checking partnerships (Marshall, 2025). Even before these changes, fact-checking interventions had stronger effects on belief accuracy than on sharing behavior. The volume and speed of online content often exceed the capacity of professional fact-checkers, allowing false claims to spread widely before corrections occur. Additionally, distrust of third-party verification sources and concerns about bias further limit their effectiveness.

Consequently, platforms have shifted toward distributed models that rely on user input and algorithmic systems to shape information flow. Examples include Community Notes on X (formerly Twitter), ranking systems, and visibility adjustments that incorporate crowd-sourced evaluations. These approaches provide contextual corrections directly on the platform and draw on aggregated user judgments. Early evidence suggests they can reduce the perceived credibility of misleading content when it is visible, but their effects on sharing behavior remain limited and depend on factors such as timing, visibility, and user engagement (Allen et al., 2024; Pennycook et al., 2020). These patterns highlight the continued influence of reinforcement structures on information-sharing behavior. Quick reposting can be reduced by delaying the display of engagement metrics or by prioritizing verified information in ranking systems. Social feedback processes can also be adjusted to emphasize credibility signals, shifting the relative advantage toward verification behaviors (Pennycook & Rand, 2019; Ecker et al., 2022). When credibility cues are weak or inconsistent, engagement-driven outcomes continue to dominate, and misleading content remains strongly reinforced.

System-level changes alter the conditions under which information is evaluated and shared. For example, Gupta et al. (2026) examined real-time, body-integrated interventions that deliver immediate verification cues when users encounter potentially misleading claims. Their wearable system used subtle haptic feedback paired with brief explanations to prompt reconsideration. These interventions introduce immediate consequences for misinformation processing and reduce the effort required to verify, thus disrupting automatic responses and encouraging more deliberate evaluation. Empirical findings suggest that such real-time nudges can improve truth discernment and increase verification behaviors without substantially increasing cognitive load, though they also carry risks of over-reliance when system feedback is inaccurate. When accuracy is supported by reduced effort, immediate feedback, and consistent reinforcement, it becomes more competitive with rapid, engagement-driven responses. Moreover, systems that align incentives with credibility are more likely to reduce misinformation than approaches that rely solely on individual motivation or education.

Conclusion

Misinformation continues to spread because digital platforms are designed to prioritize and monetize engagement through comments, likes, and reactions, rather than accuracy and verification, which require more time and effort. Social media sites capture and sustain attention by promoting content that drives engagement through speed, emotional appeal, and social relevance (Ahadzadeh et al., 2021; Bourne, 2023). In these environments, sharing is easy, immediate, and consistently reinforced, while ensuring accuracy demands additional effort and often yields delayed or less visible outcomes. This imbalance creates a predictable pattern in which misinformation spreads quickly and persists over time. Research across fields supports this pattern, and false information spreads farther and faster than true information,

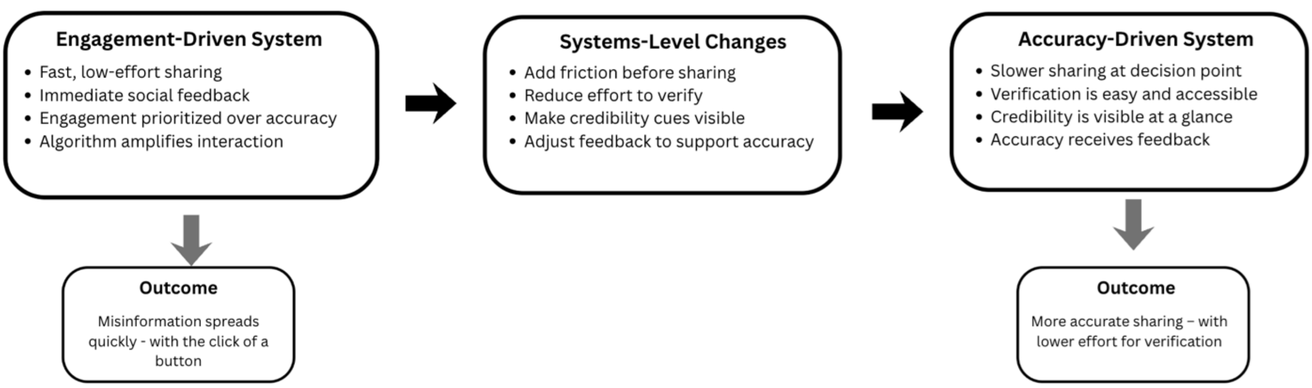


Figure 5: Engagement-driven versus accuracy-supporting systems. Engagement-driven environments promote rapid, low-effort sharing, whereas system-level changes that adjust effort, feedback, and visibility can make verification easier and improve the accuracy of sharing.

partly because of its novelty and emotional impact (Vosoughi et al., 2018). Individuals rely on low-effort processing strategies in high-volume information environments, increasing susceptibility to misleading content (Chen et al., 2021). At the same time, social and algorithmic reinforcement of belief-consistent information contributes to polarization and the persistence of misinformation within networks (Westerwick et al., 2017; Zhou & Shen, 2021; Ecker et al., 2024).

"Misinformation continues to spread because digital platforms are designed to prioritize and monetize engagement through comments, likes, and reactions, rather than accuracy and verification, which require more time and effort."

Earlier information systems imposed constraints that slowed the spread of misinformation. Editorial oversight, limited distribution channels, physical production barriers, and restricted audience participation created friction that slowed information flow and gave more time for verification. Although misinformation still occurred, these structural constraints limited its reach and reduced the likelihood of rapid, large-scale spread (Posetti & Matthews, 2018). In contrast, contemporary digital environments have removed many of these constraints, and so information can now be created, shared, and amplified instantly across global platforms with minimal effort. Algorithmic systems increase the visibility of engaging content, lower barriers to content creation, and reinforce rapid sharing (Giansiracusa, 2021). In this high-speed environment, verification has shifted from centralized systems to decentralized ones, such as community-driven corrections, which often occur after content has already circulated widely. As a result, corrective information competes with existing sharing patterns rather than preventing initial spread.

Traditional responses focus on improving individual decision-making through digital literacy, fact-checking, and accuracy-focused strategies. While these approaches can enhance knowledge and evaluation skills, they often fail to produce sustained behavioral change. Individuals may recognize credible information yet still share lower-quality content when the environment prioritizes speed and engagement. This reveals a persistent gap between knowledge and behavior. Addressing misinformation, therefore, requires shifting from individual-level solutions to system-level design changes. These system-level shifts are illustrated in Figure 5, which illustrates how modifying effort, feedback, and visibility can shift behavior from rapid sharing to more accurate use of information. Interventions that reduce the effort required to verify information, increase the salience of credibility cues, and introduce more immediate consequences for accuracy-oriented behavior can alter the conditions under which information is shared. For example, requiring interaction with content before sharing or displaying credibility indicators can reintroduce friction previously embedded in traditional information systems while maintaining broad access to information.

Misinformation persists because digital reinforcement systems favor rapid sharing over accuracy, amplifying attention, reinforcing identity-aligned content, and prioritizing engagement. When environments are structured so that accuracy is supported by low effort, immediate feedback, and visible consequences, more reliable sharing behaviors are more likely. Designing systems that make accurate sharing easier and more strongly reinforced offers a practical path to reducing the spread of misinformation. Without these changes, accurate information will continue to compete with stronger, more immediate incentives for sharing. In the attention economy, selection favors what is most engaging; when effort and reinforcement remain misaligned, accuracy is systematically outcompeted, and the spread of misinformation becomes an expected outcome rather than an exception.

THE AUTHOR

Liza Bergman-Geonie is a doctoral-level Board Certified Behavior Analyst with over 25 years of experience. She holds a PhD in Behavior Analysis from the Chicago School and has presented internationally on topics at the intersection of behavioral science, digital literacy, and misinformation. Her dissertation investigated how individuals engage with and share information on social media, with a focus on the behavioral and environmental factors that

influence the spread of misinformation. This work revealed the limits of individual-level interventions and underscored the importance of system-level influences, including platform design, response effort, and reinforcement, in shaping online behavior. She is passionate about expanding the role of behavior analysts beyond traditional settings and equipping practitioners with tools to navigate and address misinformation—especially in high-stakes areas such as news, politics, health, and education.

REFERENCES

Adams, Z., Osman, M., Bechlivanidis, C., & Meder, B. (2023, February 16). (Why) is misinformation a problem? *Perspectives on Psychological Science, 18*(6), 1436-1463. https://doi.org/10.1177/17456916221141344

Ahadzadeh, A., Ong, F., & Wu, S. (2021). Social media skepticism and belief in conspiracy theories about COVID-19: The moderating role of the dark triad. *Current Psychology, 42*, 8874-8886. https://doi.org/10.1007/s12144-021-02198-1

Allcott, H., & Gentzkow, M. (2017). Social media and fake news in the 2016 election. *Journal of Economic Perspectives, 31*(2), 211-236. https://doi.org/10.1257/jep.31.2.211

Allen, J., Arechar, A. A., Pennycook, G., & Rand, D. G. (2021). Scaling up fact-checking using the wisdom of crowds. *Science Advances*, 7(36), Article eabf4393. https://doi.org/10.1126/sciadv.abf4393

Arendt, H. (1972). Lying in politics: Reflections on the pentagon papers. *The New York Review of Books.* https://www.nybooks.com/articles/1971/11/18/lying-in-politics-reflections-on-the-pentagon-pape/

Arribas, C. M., Gertrudix, M., & Arcos, R. (2025). Preventive strategies against disinformation: A study on digital and information literacy activities led by fact-checking organisations. *Open Research Europe, 5*, 122. https://doi.org/10.12688/openreseurope.20160.1

Atske, S. (2025, September 25). *Social media and news fact sheet.* Pew Research Center. https://www.pewresearch.org/journalism/fact-sheet/social-media-and-news-fact-sheet/

Bago, B., Rand, D. G., & Pennycook, G. (2020). Fake news, fast and slow: Deliberation reduces belief in false (but not true) news headlines. *Journal of Experimental Psychology, 149*(8), 1608-1613. https://doi.org/10.1037/xge0000729

Bakir, V., & McStay, A. (2023). Core incubators of false information online. In *Optimising emotions, incubating falsehoods* (pp. 29-52). Springer International Publishing. https://doi.org/10.1007/978-3-031-13551-4_2

Bergman-Geonie, L., Ré, T., Griffith, A., & Flynn, S. (2025). Implementation of precommitment and direct instruction and its effect on sharing news stories and information on social media (31769221) [Doctoral dissertation, The Chicago School]. ProQuest. https://www.proquest.com/docview/3167866901?sourcetype=Dissertations%20&%20Theses

Bourne, P. (2023). Social media use, power, and money: Influence and economic implications. *International Journal on Transformations of Media, Journalism & Mass Communication, 10*(2), 15-29. https://www.eurekajournals.com/media.html

Brodsky, J. E., Brooks, P. J., Scimeca, D., Galati, P., Todorova, R., & Caulfield, M. (2021). Associations between online instruction in lateral reading strategies and fact-checking covid-19 news among college students. AERA Open, 7, 233285842110389. https://doi.org/10.1177/23328584211038937

Browne, K., & Watzl, S. (2026, January). The attention market—and what is wrong with it. *Philosophical Studies, 183*(1), 227-257. https://doi.org/10.1007/s11098-025-02436-3

Caulfield, M. (2019, June 19). SIFT (the four moves). *Hapgood.* https://hapgood.us/2019/06/19/sift-the-four-moves/

Chen, S., Xiao, L., & Mao, J. (2021). Persuasion strategies of misinformation-containing posts in the social media. *Information Processing & Management, 58*(5), 102665. https://doi.org/10.1016/j.ipm.2021.102665

Diepeveen, S., & Pinet, M. (2022). User perspectives on digital literacy as a response to misinformation. *Development Policy Review, 40*(S2). https://doi.org/10.1111/dpr.12671

Dorrell, D., & Heal, S. (2024, July). Responsible generative AI: Both humans and algorithms have biases. *Frontier Economics.* https://www.frontier-economics.com/uk/en/news-and-insights/articles/article-i20814-responsible-generative-ai-both-humans-and-algorithms-have-biases/

Ecker, U. K. H., Lewandowsky, S., Cook, J., Schmid, P., Fazio, L. K., Brashier, N., Kendeou, P., Vraga, E. K., & Amazeen, M. A. (2022, January 12). The psychological drivers of misinformation belief and its resistance to correction. *Nature Reviews Psychology, 1*(1), 13-29. https://doi.org/10.1038/s44159-021-00006-y

Ecker, U. K. H., Tay, L., Roozenbeek, J., Linden, S. V. D., Cook, J., Oreskes, N., & Lewandowsky, S. (2024). Why misinformation must not be ignored. *American Psychologist, 80*(6), 867-878. https://doi.org/10.1037/amp0001448

Epstein, Z., Sirlin, N., Arechar, A., Pennycook, G., & Rand, D. (2023). The social media context interferes with truth discernment. *Science Advances, 9*(9). https://doi.org/10.1126/sciadv.abo6169

Fazio, L. K., Brashier, N. M., Payne, B. K., & Marsh, E. J. (2015). Knowledge does not protect against illusory truth. *Journal of Experimental Psychology: General, 144*(5), 993-1002. https://doi.org/10.1037/xge0000098

Giansiracusa, N. (2021). How algorithms create and prevent fake news. *Apress.* https://doi.org/10.1007/978-1-4842-7155-1

Gupta, C., Aritonang, N. V., Rajendran, D. P. D., Danry, V., Maes, P., & Nanayakarra, S. (2026). Feeling the facts: Real-time wearable fact-checkers can use nudges to reduce user belief in false information. *CHI '26: Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems.* https://doi.org/10.1145/3772318.3791679

Kearns, C. E., Schmidt, L. A., & Glantz, S. A. (2016). Sugar industry and coronary heart disease research: A historical analysis of internal industry documents. *JAMA Internal Medicine, 176*(11), 1680. https://doi.org/10.1001/jamainternmed.2016.5394

Lazer, D. M. J., Baum, M. A., Benkler, Y., Berinsky, A. J., Greenhill, K. M., Menczer, F., Metzger, M. J., Nyhan, B., Pennycook, G., Rothschild, D., Schudson, M., Sloman, S. A., Sunstein, C. R., Thorson, E. A., Watts, D. J., & Zittrain, J. L. (2018). The science of fake news. *Science, 359*(6380), 1094-1096. https://doi.org/10.1126/science.aao2998

Lewandowsky, S., Ecker, U. H., & Cook, J. (2017). Beyond misinformation: Understanding and coping with the "post-truth" era. *Journal of Applied Research in Memory and Cognition, 6*(4), 353-369. https://doi.org/10.1016/j.jarmac.2017.07.008

Lewin, K. (1951). Field theory in social science: Selected theoretical papers (edited by Dorwin Cartwright). Harpers. https://psycnet.apa.org/record/1951-06769-000

Marshall, L. (2025, January 10). Why Meta fired its fact-checkers and what you can do about it. *CU Boulder Today.* https://www.colorado.edu/today/2025/01/10/why-meta-fired-its-fact-checkers-and-what-you-can-do-about-it

McKay, S., & Tenove, C. (2020). Disinformation as a threat to deliberative democracy. *Political Research Quarterly, 74*(3), 703-717. https://doi.org/10.1177/1065912920938143

Milmo, D. (2025, April 2). Internet replaces TV as UK's most popular news source for first time. *The Guardian.* https://www.theguardian.com/media/article/2024/sep/10/internet-tv-uk-most-popular-news-source-first-time

Nyhan, B., & Reifler, J. (2015). Estimating fact-checkings effects: Evidence from a long-term experiment during campaign 2014. *American Press Institute (FactChecking Project Report).* http://www.americanpressinstitute.org/wp-content/uploads/2015/ 04/Estimating-Fact-Checkings-Effect.pdf

Oremus, W., Alcantara, C., Merrill, J. B., & Galocha, A. (2021, October 26). How Facebook shapes your feed. *Washington Post.* https://www.washingtonpost.com/technology/interactive/2021/how-facebook-algorithm-works/

Pennycook, G., McPhetres, J., Zhang, Y., Lu, J. G., & Rand, D. G. (2020). Fighting COVID-19 misinformation on social media: Experimental evidence for a scalable accuracy-nudge intervention. *Psychological Science, 31*(7), 770-780. https://doi.org/10.1177/0956797620939054

Pennycook, G., & Rand, D. G. (2019). Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning. *Cognition, 188,* 39-50. https://doi.org/10.1016/j.cognition.2018.06.011

Pew Research Center. (2025, November 20). *Social media fact sheet.* https://www.pewresearch.org/internet/fact-sheet/social-media/

Posetti, J., & Matthews, A. (2018, July). *A short guide to the history of 'fake news' and disinformation.* International Center for Journalists. https://www.icfj.org/news/short-guide-history-fake-news-and-disinformation-new-icfj-learning-module

Robb, A., & Robb, A. (2017, November 16). Pizzagate: Anatomy of a fake news scandal. *Rolling Stone.* https://www.rollingstone.com/feature/anatomy-of-a-fake-news-scandal-125877/

Rong, Y., Chen, L., Zhu, T., Song, Y., Yu, M., Shan, Z., Sands, A., Hu, F. B., & Liu, L. (2013). Egg consumption and risk of coronary heart disease and stroke: Dose-response meta-analysis of prospective cohort studies. *BMJ, 346*(jan07 2), e8539-e8539. https://doi.org/10.1136/bmj.e8539

Rothkopf, D. J. (2003, May 11). When the buzz bites back. *Washington Post.* https://www.washingtonpost.com/archive/opinions/2003/05/11/when-the-buzz-bites-back/bc8cd84f-cab6-4648-bf58-0277261af6cd/

Santos-d'Amorim, K., & Fernandes de Oliveira Miranda, M. (2021). Informação incorreta, desinformação e má informação: Esclarecendo definições e exemplos em tempos de desinfodemia. *Encontros Bibli: revista eletrônica de biblioteconomia e ciência da informação, 26,* 01-23. https://doi.org/10.5007/1518-2924.2021.e76900

Soll, J. (2016, December 18). The long and brutal history of fake news. *Politico Magazine.* https://www.politico.com/magazine/story/2016/12/fake-news-history-long-violent-214535/

Stokel-Walker, C. (2026, April 6). Scientists invented a fake disease. AI told people it was real. *Nature.* https://www.nature.com/articles/d41586-026-01100-y

Suherman, A. (2025, July 11). The algorithmic trap: How social media monetization undermines investigative journalism in local media. *Frontiers in Communication, 10,* Article 1619367. https://doi.org/10.3389/fcomm.2025.1619367

Sundar, S., Snyder, E., Liao, M., Yin, J., Wang, J., & Chi, G. (2024). Sharing without clicking on news in social media. *Nature Human Behaviour. 9,* 156-168. https://doi.org/10.1038/s41562-024-02067-4

Swire-Thompson, B., Cook, J., Butler, L. H., Sanderson, J. A., Lewandowsky, S., & Ecker, U. K. H. (2021, December). Correction format has a limited role when debunking misinformation. *Cognitive Research: Principles and Implications, 6*(1), Article 83. https://doi.org/10.1186/s41235-021-00346-6

Tsipursky, G., Votta, F., & Mulick, J. A. (2018). A psychological approach to promoting truth in politics: The pro-truth pledge. *Journal of Social and Political Psychology, 6*(2), 271-290. https://doi.org/10.5964/jspp.v6i2.856

Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science, 359*(6380), 1146-1151. https://doi.org/10.1126/science.aap9559

Washburn, E. (2024, August 9). Don't fall for social media rage-bait: Here's what to look out for. *Daily Citizen.* https://dailycitizen.focusonthefamily.com/dont-fall-for-social-media-rage-bait-heres-what-to-look-out-for/

Watson, C. A. (2018). Information literacy in a fake/false news world: An overview of the characteristics of fake news and its historical development. *International Journal of Legal Information, 46*(2), 93-96. https://doi.org/10.1017/jli.2018.25

Westerwick, A., Johnson, B. K., & Knobloch-Westerwick, S. (2017). Confirmation biases in selective exposure to political online information: Source bias vs. content bias. *Communication Monographs, 84*(3), 343-364. https://doi.org/10.1080/03637751.2016.1272761

Zhou, Y., & Shen, L. (2021). Confirmation bias and the persistence of misinformation on climate change. *Communication Research,* 009365022110280. https://doi.org/10.1177/00936502211028049

A Practical, Multi-Dimensional Framework for Classifying AI Agents in the Investment Industry

IVANA ZILIC

Department of Management and Entrepreneurship,
Richard H. Driehaus College of Business, DePaul University

PATRICK WIERCKX AND MICHIEL KUHN

Panthera Institute

MARKUS SCHULLER[1]

EDHEC Business School, IE Business School, and Panthera Institute

Click or scan QR code
to watch video summary

Artificial Intelligence (AI) is increasingly embedded across the investment industry, influencing how information is processed, decisions are made and value is created. Asset managers, asset owners and regulators are experimenting with AI-enabled tools, yet many firms struggle to evaluate these applications consistently or assess their strategic importance within the investment process. As a result, AI adoption often remains fragmented, incremental and reactive rather than deliberate and outcome-driven. This article takes a practical approach to this challenge by combining existing research from behavioural economics, organisational management, investment theory and decision-making with real-world industry experience. It introduces a structured way to think about AI in the investment process, offering a clear framework to help classify and evaluate different types of AI tools. While the individual dimensions are not new, their combination provides a practical lens for comparative assessment across AI tools and use cases. The framework is intended to help investment professionals clarify where AI is most likely to add economic and organisational value, where its benefits may be limited and how different AI applications relate to one another strategically. Additionally, it supports the more informed prioritisation of AI initiatives, better alignment with business objectives and clearer communication across investment, technology and governance functions. Rather than presenting empirical results, this article aims to equip practitioners and policy stakeholders with a usable decision-making aid – one that moves beyond abstract discussions of AI and toward more systematic integration in real-world investment settings.

1 Corresponding author: mhschuller@panthera.mc

From Hype to Taxonomy: A Practical Framework for AI Agents in Investment Management

Professional investors are tasked with maximising the asymmetry between risk and return – a process that increasingly relies on advanced analytical tools. AI agents play an increasingly central role throughout the investment value chain, from data analysis to augmenting decision-making capabilities. For example, Sutiene et al. (2024) explore various AI and machine learning (ML) techniques for tasks such as asset allocation, including methods for selecting predictors and forecasting time series. The expanding adoption of AI agents within the investment industry has consequently created a need for a systematic and structured framework to understand their roles, capabilities and overall impact.

To establish a theoretical foundation for this framework, this study draws on an extensive literature review, revealing several critical gaps. First, existing research provides limited clarity regarding the classification of AI systems and agents in financial contexts, particularly with respect to issues of scalability and model interpretability (Joshi, 2025). Second, there remains insufficient understanding of how AI can be strategically deployed to generate business value and enhance decision-making processes (Borges et al., 2021). Third, the literature inadequately addresses how data-driven optimisation by AI agents may reshape market coordination mechanisms and organisational architectures (Lennart, 2024).

As a result, no comprehensive framework currently exists to guide the deployment of AI agents across the investment process to optimise the risk-return trade-off.

> "No comprehensive framework currently exists to guide the deployment of AI agents across the investment process to optimise the risk-return trade-off."

To address this deficiency, this article proposes a novel classification system to categorise AI agents in the investment industry. The objective is to establish a common conceptual framework for analysing and discussing the role of AI across the investment domain to improve decision-making and, in turn, drive better investment outcomes.

Theoretical Background

Definition of Artificial Intelligence and Large Language Models

The term 'artificial intelligence' (AI) was created by emeritus Stanford Professor John McCarthy in 1955. It was defined as "the science and engineering of making intelligent machines" (Rajaraman, 2014) and refers to a set of methodologies that enable robots to carry out complex tasks that typically require human intellect (Hicham et al., 2023).

Large Language Models (LLMs) are a prominent area within the AI domain and can be described as 'generative' language models, in that they can generate human-like language and perform a range of language-processing tasks due to their training on extensive textual datasets, including publicly available data and data licensed from third parties (Sarker, 2022).

Taken together, these perspectives suggest that the effective professional use of LLMs requires more than scalable text generation. Value emerges when generative capabilities are paired with domain-grounded, externally verifiable knowledge.

> "Value emerges when generative capabilities are paired with domain-grounded, externally verifiable knowledge."

Lack of Applicable Classification Systems

Despite the growing intent to adopt AI-based applications, their integration within the investment industry remains fragmented and unstructured (ESMA, 2026). To date, considerable attention has been devoted to the potential risks and biases associated with AI, particularly in relation to ethical concerns. Addressing these challenges depends critically on the development and application of Explainable AI (XAI), which helps demystify 'black-box' models and enhance transparency and trust in investment decision-making processes (Vuković et al., 2025). Sutiene et al. (2024) present the role of XAI in addressing the 'black-box' nature of specific AI models, thereby building trust and transparency in investment processes, particularly considering the increasing regulatory focus on the investment industry. Furthermore, robust governance frameworks are crucial for effectively addressing issues such as algorithmic biases and transparency. The aim is to strike a balanced approach that promotes innovation while ensuring the responsible and sustainable integration of AI within the financial ecosystem.

A potential yet underexplored driver of the fragmented implementation of AI in the investment industry is the absence of a common classification system to categorise systematically the growing range of AI applications. Mökander et al. (2023) explore methods for classifying AI systems to support the practical implementation of AI ethics and governance. They review existing approaches to classifying AI systems and identify three main 'mental models': the Switch, a simple binary classification; the Ladder, a risk-based approach; and the Matrix, a multi-dimensional classification. The article examines the strengths and weaknesses of each model in practical terms by assessing suitability for purpose, simplicity, clarity and stability over time. The aim is to help organisations and regulators define clear and workable boundaries for AI governance. However, the article's classification hardly helps investment firms seeking to determine where AI can generate the greatest value. This limitation primarily stems from the framework's lack of alignment with the typical investment processes employed within such organisations.

The OECD (OECD Framework for the Classification of AI Systems, 2022) introduced its Framework for the Classification of AI Systems, a tool designed to assist policymakers and others in understanding and evaluating the characteristics of AI systems in specific contexts. The OECD framework categorises AI systems across five dimensions: people and planet, economic context, data and input, AI model and task and output. It highlights the relationship between AI system characteristics and the AI system lifecycle, provides guidance on contributions for risk management and identifies relevant actors involved at different stages. However, it does not align with the typical workflow of investment firms and therefore provides limited practical guidance for the optimal integration of AI within processes. Consequently, the effective implementation of AI in the investment industry depends on the development of more sophisticated categorisation frameworks.

Design of the Classification System

Financial markets are complex, adaptive systems (Lo, 2004), the primary purpose of which is to facilitate the efficient allocation of resources among their participants. The relationship between investors and financial markets reflects an interaction between endogenous and exogenous forces. Investors act as endogenous agents whose decisions shape capital allocation, whilst financial markets, in turn, form the exogenous environment in which those decisions play out, with varying degrees of market efficiency influencing outcomes. This article

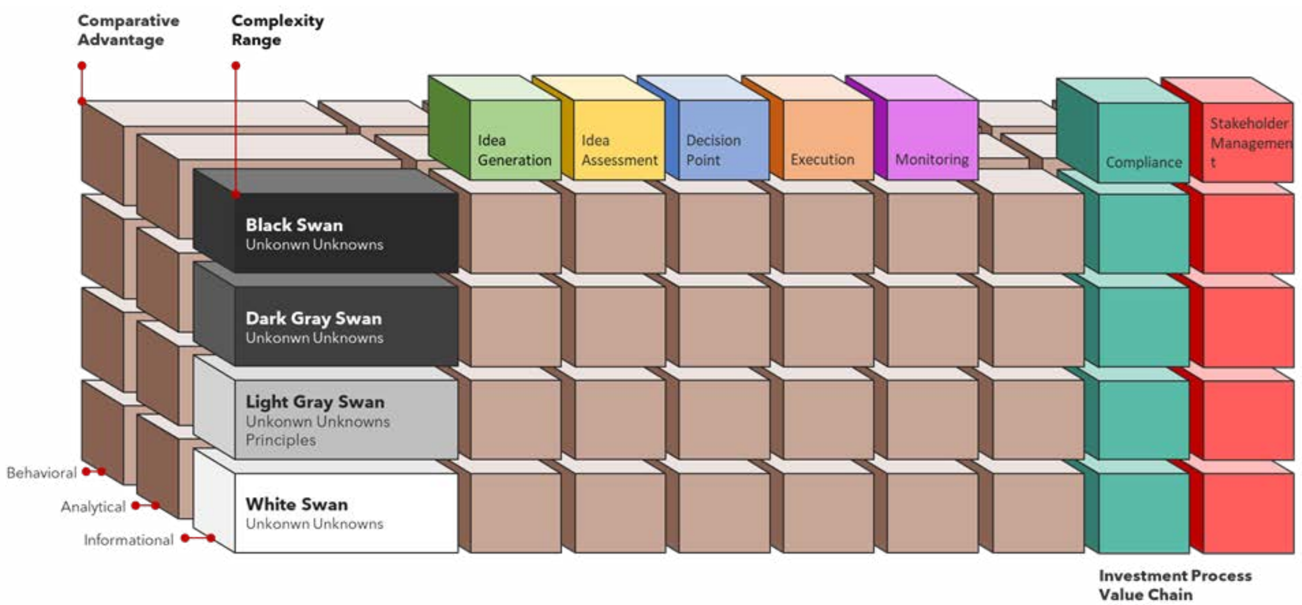


Figure 1: The multi-dimensional framework.

proposes a framework that integrates both sets of forces and the mechanisms linking them. The framework identifies three dimensions that jointly capture this system. The first is the investment process, which reflects the endogenous side, i.e., the sequence of activities undertaken by investors, including idea generation, analysis, decision making, execution and monitoring. The second is the range of complexity in financial markets, reflecting the exogenous side, i.e., the structural properties that make markets more or less predictable. The third dimension, comparative advantage, functions as a bridge between endogenous and exogenous drivers and specifies how investors interact with financial markets and the basis upon which they seek to establish an edge. Because markets are highly competitive, sustained success requires investors to identify and exploit distinctive advantages.

> “Because markets are highly competitive, sustained success requires investors to identify and exploit distinctive advantages.”

Taken together, the three dimensions form a conceptual model that offers a systematic way to categorise the growing range of AI solutions in investment management (see Figure 1). By mapping these tools onto the framework, practitioners can compare applications more clearly. The framework also highlights where AI is likely to deliver the greatest transformative impact, as well as where its effects may be more limited or incremental.

Classification System Dimensions

Three Types of Comparative Advantages in the Investment Industry

Financial markets are highly competitive ecosystems in which countless participants compete to achieve excess returns. Just as companies aim to outperform their rivals to achieve above-average profitability, many investors aim to outperform other investors, many of whom are equally skilled and well-resourced. This makes a comparative edge essential for success.

Industrial organisation theory provides a useful lens for understanding this dynamic: investment strategies can be seen as competitive strategies. As Michael Porter (1980/1998) emphasised, companies can only earn sustainable excess returns if they possess competitive advantages (i.e., enduring traits that allow them to outperform competitors consistently). In capital markets, the same principle applies. Sustained outperformance requires durable advantages.

The typology introduced by Fuller (1998) and expanded in the CAB framework of Wierckx (2021) categorises these advantages into three broad types: informational, analytical and behavioural (see Figure 2). These differ in origin, ease of acquisition and vulnerability to erosion.

An informational advantage arises when a market participant possesses knowledge that is materially relevant and not yet reflected in market prices. This can come from earlier access to news, unique data sources or overlooked public disclosures. The

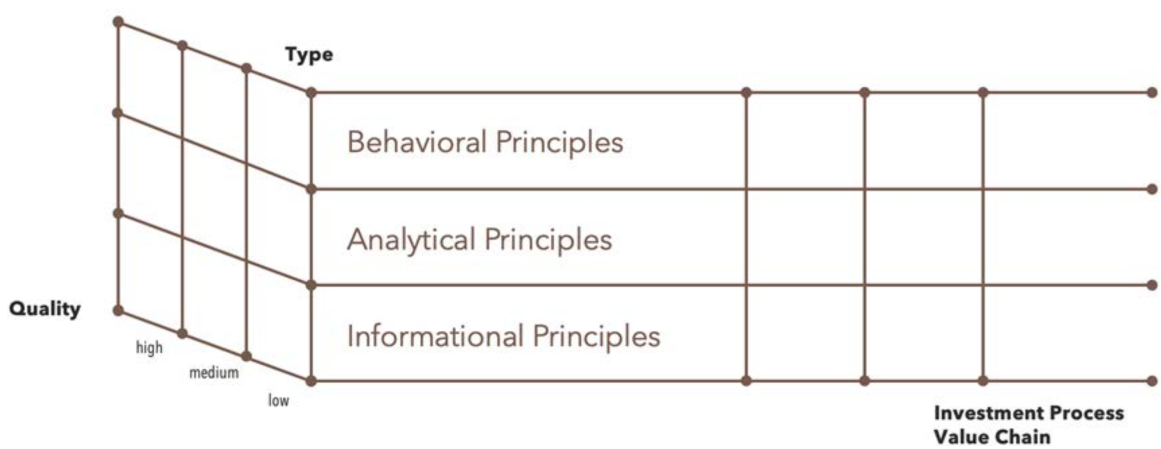


Figure 2: The investment industry’s comparative advantages.

barriers to acquiring such information have recently dropped dramatically, thanks to the availability of technological tools like web scraping, alternative data and faster data feeds.

However, the benefits that may be derived from these advantages tend to be short-lived. Once the information is disseminated or absorbed by the market, such as after an earnings release or regulatory filing, the advantage disappears. Moreover, regulations such as Regulation Fair Disclosure and the Sarbanes-Oxley Act, combined with the widespread availability of the Internet and now AI-driven tools, have dramatically reduced the exclusivity and shelf life of information-based advantages.

Analytical advantage often comes from interpreting public information better than competitors, and it arises when investors can synthesise complex or fragmented data more effectively. Gaining this capability typically requires sustained investment in expertise, research and human capital. Limits on human cognition create barriers to entry, making analytical advantages relatively resilient. Over time, though, these advantages often weaken as AI and machine learning increasingly automate synthesis, pattern recognition and inference. A behavioural advantage capitalises on the persistent irrationalities and biases of other market participants. Rooted in the pioneering work of Kahneman and Tversky (1974), behavioural finance identifies predictable deviations from rational expectations, including overreaction (De Bondt & Thaler, 1985; 1987), anchoring (Kahneman & Tversky, 1974) and loss aversion (Kahneman & Tversky, 1979). Other biases that contribute to systematic mispricing in financial markets include overconfidence (Odean, 1998), limited attention (Hirshleifer & Teoh, 2003) and representativeness-driven extrapolation (Barberis et al., 1998).

> “Behavioural advantages tend to be more durable than those derived from informational or analytical advantages.”

Because cognitive biases are deeply ingrained and evolve only gradually over time, the inefficiencies they produce are relatively persistent. Consequently, behavioural advantages tend to be more durable than those derived from informational or analytical advantages. Although awareness of behavioural biases has increased significantly, such awareness does not necessarily translate into an enhanced ability to mitigate their effects. As Kahneman (2011) argues, intuitive cognitive processes are often resistant to correction, even when individuals are explicitly aware of their biases. Especially in environments that are complex and uncertain, such as most investment markets, debiasing is much harder because feedback is delayed, noisy and often ambiguous (Fischhoff, 1982).

In contrast, informational and analytical advantages are far more susceptible to erosion: once information is disseminated or an analytical insight becomes widespread, the associated opportunity is typically arbitraged away. Behavioural advantages, by comparison, exhibit greater resilience precisely because knowing about a bias does not equate to overcoming it. As a result, behavioural advantages are the most durable and resilient of the three.

Investment Process

In practice, these comparative advantages manifest at different points in the investment workflow. The following section outlines the common stages through which investors apply and operationalise these advantages. Although the uses of specific comparative advantages and their corresponding investment strategies differ widely across market participants, the investment process generally follows a common sequence (e.g., Maginn & Tuttle, 1990). The initial stage, namely the Idea Generation Phase, involves identifying potential investment opportunities that align with the participant's comparative advantage. Idea generation can be pursued through a top-down process, in which a broad investment universe is screened to isolate securities with favourable risk-return characteristics, or through a bottom-up process, in which opportunities emerge from external sources, such as sell-side brokers, or from the internal analysis of the investment team.

The second stage, the Idea Assessment Phase, involves evaluating the opportunities identified in the preceding step. The objective is to determine which investments exhibit the most attractive risk-return characteristics, conditional on the existing portfolio and mandate constraints. This evaluation typically follows a set of predetermined criteria that reflect the market participant's analytical framework and comparative advantage. Since AI can analyse large datasets and accelerate time-consuming tasks, such as extracting information from multiple annual reports into spreadsheets, it is expected to have a significant impact in this area.

If an opportunity is deemed attractive, the process advances to the Decision Phase, where a formal decision to commit capital is made. Decision-making practices vary substantially across market participants and depend on factors such as investment team size, organisational hierarchy and client- or mandate-specific guidelines. Crucially, because society generally expects humans to remain accountable for decisions, it is unlikely that AI will fully take over at this stage; rather, it is expected to augment human decision-makers. For instance, it can serve as a valuable partner by providing critical feedback, ensuring that decisions are made only after careful consideration.

Once an investment decision has been taken, the Execution Phase follows, in which the position is established through trading activity. This step entails selecting appropriate trading venues and strategies while balancing objectives to minimise transaction costs, market impact and execution risk.

During the Monitoring Phase, market participants continually assess whether the investment remains aligned with its intended objectives. This includes evaluating the position's ongoing risk-return profile, its interaction with other portfolio holdings and any changes in the underlying thesis. The monitoring process ensures that deviations from expectations are addressed through rebalancing, hedging or divestment.

In parallel with monitoring, the Compliance Phase ensures that both regulatory obligations and client-specific guidelines are adhered to throughout the investment's life cycle. Compliance functions serve as a safeguard, ensuring that investment activities remain consistent with legal requirements, fiduciary duties and contractual mandates.

Finally, the Stakeholder Management Phase encompasses communicating investment rationales, portfolio positioning and performance outcomes to clients and other stakeholders. Transparent and consistent communication is essential for maintaining trust, aligning expectations and demonstrating adherence to fiduciary responsibilities. This phase closes the loop by ensuring that investment processes are not only internally consistent, but also externally validated through accountability to clients and beneficiaries.

Taken together, these stages define a general investment process. They include idea generation, idea assessment, decision-making, execution, monitoring, compliance and stakeholder management. While the substance of each stage differs across firms and contexts, the overall sequence is broadly consistent throughout the industry.

Complexity Range

Across the investment process, decisions are made under very different conditions of uncertainty. Understanding these differences is essential for evaluating not only where AI systems add value, but also which types of AI agents are appropriate in different settings.

Figure 3: Summary of the various uncertainty environments, using the methods discussed above.

Taleb (2012) classifies risk and uncertainty into three distinct categories: White Swans, Grey Swans and Black Swans. White Swans pertain to contexts in which risks are well-defined, meaning that the probability distribution of outcomes is known, even though the specific outcome remains uncertain. In practice, however, the acquisition of additional data can introduce more noise than an informative signal (Dawes, 1971), thereby complicating the precise estimation of probabilities despite theoretical knowledge of the distribution. These situations are often referred to as 'Light Grey Swans'. Dark Grey Swans relate to environments when information is limited; here, identifying relevant factors and causal relationships becomes more critical than accurately assigning quantitative weights (Dawes & Corrigan, 1974). Finally, Black Swans represent contexts of maximal uncertainty, encompassing 'unknown unknowns'. In these environments, Gigerenzer (2011) argues that simpler strategies, such as those that rely on logic and heuristics, may be more effective than complex models.

These distinctions have direct implications for the design and deployment of AI agents in investment contexts. In White Swan and Light Grey Swan contexts (where probability distributions are known or can be reasonably approximated), AI tools tend to perform strongly. Machine learning models, particularly those optimised for pattern recognition and prediction, can efficiently process large volumes of structured financial data, identify statistical regularities and support forecasting tasks such as asset pricing, risk estimation and portfolio optimisation. However, even in these settings, the marginal benefit of increasingly complex models may diminish when additional data primarily introduces noise rather than a signal, implying that model governance and feature selection remain critical.

As uncertainty increases in Dark Grey and Black Swan environments, the limitations of AI become more pronounced. Model performance becomes constrained by the availability and quality of inputs, making it difficult for purely data-driven approaches to infer causal relationships reliably. In extreme uncertainty, characteristic of Black Swans, AI models trained on historical data are inherently ill-equipped to anticipate unprecedented events. As a result, adoption in these contexts tends to shift toward hybrid approaches, where AI augments rather than replaces human judgment. In this case, AI can still add value through scenario analysis, stress testing and the identification of weak signals, but decision-making relies more heavily on human intuition, domain expertise and heuristic reasoning.

These differences in uncertainty imply that the suitability of AI systems is context-dependent – a point that directly motivates the classification of AI agents introduced in the next section.

Integrating the Dimensions: A Practical Illustration

The interaction of the three dimensions can be illustrated using an NLP-based earnings-call sentiment analyser, which may generate an analytical advantage by converting public, unstructured information into systematic signals.

Across the investment process, the analyser is most effective in the idea generation and idea assessment phases, but it can also support the monitoring phase by enabling the longitudinal tracking of management communication. Its role in execution and final decision-making remains limited, with accountability retained by human decision-makers.

In terms of complexity, the tool performs best in White Swan and Light Grey Swan environments, where the effects of certain corporate communication on investment returns are stable and historical data is available.

Integrated Classification System

Taken together, the three dimensions, namely comparative advantage, investment process stage and complexity range, form an integrated classification system for AI applications in investment management. Their joint consideration allows AI tools to be positioned more precisely and compared more systematically.

> "AI effectiveness depends on the interaction of three dimensions: comparative advantage, investment process stage and complexity range."

The central insight is that AI effectiveness depends on the interaction of these dimensions. AI agents that generate informational or analytical advantages may be well-suited to specific investment process stages, albeit under particular uncertainty conditions. Moreover, tools that support behavioural advantages tend to rely more on interpretability and human oversight and are less dependent on data richness. Mapping AI agents across all three dimensions therefore clarifies both where they are best deployed and the conditions under which they add value.

Discussion

Implications for Practice

The proposed framework offers investment managers, software developers and regulators a structured approach to classifying AI agents. By providing a shared taxonomy, it improves decision-making and governance while clarifying when AI should act as an augmenting copilot. This systematic classification reduces the risk of misapplication, promotes responsible adoption and helps align AI deployment with a firm-specific comparative advantage. For developers and regulators, the framework supports fit-for-purpose design and differentiated governance, reducing misapplication and aligning AI use with firm-specific strengths and fiduciary responsibilities.

Strengths, Implications and Future Directions

A central contribution of this work is its integration of behavioural economics, organisational management theory, investment theory and decision-making theory into a coherent classification system. By moving beyond purely technical AI perspectives, the framework captures the interplay of design, human interaction and the financial market environment, thereby bridging the fragmented literature.

Future research should expand empirical validation of the classification system through case studies, simulations and practitioner feedback. Longitudinal studies could shed light on the effects of AI agent adoption on investment outcomes, whilst ethical considerations, particularly regarding transparency, accountability and bias, require deeper investigation. Finally, cross-sectoral comparisons would help assess the framework's transferability beyond investment management, thereby reinforcing its broader relevance.

Conclusions

This study examines the application of AI in investment management through a multi-dimensional classification framework. By integrating insights from behavioural economics, organisational management, investment theory and decision-making research, the framework brings practical conceptual clarity to a field where inconsistent taxonomies can result in both missed opportunities and the inappropriate use of AI. By offering a structured way to compare AI applications, the framework helps investment professionals, developers and regulators in aligning AI adoption with both organisational strengths and governance requirements.

THE AUTHORS

Ivana Zilic, Ph.D., is a clinical professor in the Department of Management and Entrepreneurship at DePaul University's Driehaus College of Business. Her research interests range from CSR issues to business ethics. Dr. Zilic has published in *Employee Relations*, *Industrial Relations Journal*, Business and Society Review and the *Journal of Communication Management*, among others. She has presented her work at numerous conferences and meetings, including the Academy of Management, the Eastern Academy of Management, the Society for Business Ethics and other organisations. Dr. Zilic teaches courses in international business, management and innovation. She has taught internationally in Europe. She recently received MIT Course Certificates in the Executive Program in Artificial Intelligence and in Leadership.

Patrick J. Wierckx, MSc, CFA, is a Dutch investment professional with more than 30 years of experience in managing institutional equity portfolios. He has held senior positions, including head of equities at one of the Netherlands' largest asset managers, and he is the author of "Investing in Hidden Monopolies: Why Customer Loyalty Creates Superior Moats and How You Can Profit." Wierckx is a CFA Charterholder, an ESMA-registered Institutional Investment Advisor and a member of the CFA Institute. He holds a master's degree in Business Economics.

Michiel Kühn is a quantitative research analyst at Panthera Solutions, where he researches investor decision-intelligence, spanning agentic AI, climate

risk and quantitative methods, and he regularly collaborates with portfolio managers and CIOs. Kühn is also a graduate student (MiM/MSc in Financial Engineering) at EDHEC Business School, where he is on the Dean's List (top 5%) and served as a 2024/25 class representative. An EDHEC Academic Excellence and France-South Africa scholar, he previously served clients and financial advisors at Allan Gray, Africa's largest privately owned asset manager.

Dr. Markus Schuller is the founder and managing partner of Panthera Group, a multi-award-winning firm recognised for pioneering the practical integration of Human and Artificial Intelligence into institutional investment processes. Panthera's proprietary Decision GPS platform blends behavioural science and quantitative finance with machine intelligence to optimise investment decision-making. Alongside his entrepreneurial work, Schuller serves as an adjunct and visiting professor at EDHEC Business School, IE Business School and the International University of Monaco, where he teaches in top-ranked Master in Finance programs. A former hedge fund manager, equity trader and derivatives trader, he is a published researcher, keynote speaker and regular contributor to leading academic and professional journals.

REFERENCES

Barberis, N., Shleifer, A., & Vishny, R. (1998). A model of investor sentiment. *Journal of Financial Economics, 49*(3), 307-343. https://doi.org/10.1016/S0304-405X(98)00027-0

Borges, A. F. S, Laurindo, F. J. B., Spínola, M. M., Gonçalves, R. F., & Mattos, C. A. (2021). The strategic use of artificial intelligence in the digital era: Systematic literature review and future research direction. *International Journal of Information Management, 57*, 102225. https://doi.org/10.1016/j.ijinfomgt.2020.102225

Dawes, R. M. (1971). A case study of graduate admissions: Application of three principles of human decision making. *American Psychologist, 26*(2), 180-188. https://doi.org/10.1037/h0030868

Dawes, R. M. , & Corrigan, B. (1974). Linear models in decision making. *Psychological Bulletin, 81*(2), 95-106. https://doi.org/10.1037/h0037613

De Bondt, W. F. M., & Thaler, R. (1985). Does the stock market overreact? *Journal of Finance, 40*(3), 793-805. https://doi.org/10.1111/j.1540-6261.1985.tb05004.x

De Bondt, W. F. M., & Thaler, R. (1987). Further evidence on investor overreaction and stock market seasonality. *Journal of Finance, 42*(3), 557-581. https://doi.org/10.1111/j.1540-6261.1987.tb04569.x

ESMA (2026). AI adoption and trends in securities markets: EU evidence. *ESMA TRV Risk Analysis.* https://www.esma.europa.eu/document/ai-adoption-and-trends-securities-markets-eu-evidence-trv-article

Fischhoff, B. (1982). Debiasing. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 422-444). Cambridge University Press.

Fuller, R. J. (1998). Behavioral finance and the sources of alpha. *Journal of Pension Plan Investing,* 2(3). https://www.acsu.buffalo.edu/~keechung/Collection%20of%20Papers%20for%20courses/Behavioral%20Finance%20and%20Sources%20of%20Alpha.pdf

Gigerenzer, G., & Gaissmaier, W. (2011). Heuristic decision making. *Annual Review of Psychology, 62*, 451-482. https://doi.org/10.1146/annurev-psych-120709-145346

Hicham, N., Habbat, N., & Karim, S. (2023). Strategic framework for leveraging artificial intelligence in future marketing decision-making. *Journal of Intelligent Management Decision,* 2(3), 139-150. https://doi.org/10.56578/jimd020304

Hirshleifer, D., & Teoh, S. H. (2003). Limited attention, information disclosure, and financial reporting. *Journal of Accounting and Economics, 36*(1-3), 337-386. https://doi.org/10.1016/j.jacceco.2003.10.002

Joshi, S. (2025). A literature review of gen AI agents in financial applications: Models and implementations. *Social Science Research Network.* https://doi.org/10.2139/ssrn.5133985

Kahneman, D. (2011). *Thinking, fast and slow.* Farrar, Straus and Giroux.

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica,* 47(2), 263-291. https://doi.org/10.2307/1914185

Kahneman, D., & Tversky, A. (1974). Judgment under uncertainty: Heuristics and biases. *Science,* *185*(4157), 1124-1131. https://doi.org/10.1126/science.185.4157.1124

Lennart, A. (2024). Autonomous AI agents in decentralized finance: Market dynamics, application areas, and theoretical implications. *Social Science Research Network.* https://doi.org/10.2139/ssrn.5055677

Lo, A. W. (2010, February 25). *The adaptive markets hypothesis and the new investment paradigm*. CFA Institute Research and Policy Center. https://rpc.cfainstitute.org/research/multimedia/2010/the-adaptive-markets-hypothesis-and-the-new-investment-paradigm

Lo, A. W. (2004). The adaptive markets hypothesis: Market efficiency from an evolutionary perspective. *Journal of Portfolio Management, 30*(5), 15-29. https://doi.org/10.3905/jpm.2004.442611

Maginn, J. L. (Ed.) (2007). *Managing investment portfolios: a dynamic process (3rd. ed.)*. Wiley (CFA Institute investment series).

Mökander, J., Sheth, M., Watson, D. S., & Floridi, L. (2023). The switch, the ladder, and the matrix: Models for classifying AI systems. *Minds and Machines, 33*(1), 221-248. https://doi.org/10.1007/s11023-022-09620-y

Odean, T. (1998). Volume, volatility, price, and profit when all traders are above average. *Journal of Finance, 53*(6), 1887-1934. https://doi.org/10.1111/0022-1082.00078

OECD. (2022a). OECD framework for the classification of AI systems. https://www.oecd.org/en/publications/oecd-framework-for-the-classification-of-ai-systems_cb6d9eca-en.html

Porter, M. E. (1998). *Competitive strategy: Techniques for analyzing industries and competitors: with a new introduction.* Free Press.

Rajaraman, V. (2014). *John McCarthy: Father of artificial intelligence.* Resonance.

Sutiene, K. et al. (2024). Enhancing portfolio management using artificial intelligence: Literature review. *Frontiers in Artificial Intelligence, 7.* https://doi.org/10.3389/frai.2024.1371502

Sarker, I. H. (2022). AI-based modeling: Techniques, applications and research issues towards automation, intelligent and smart systems. *SN Computer Science, 3*(2), 158-158. https://doi.org/10.1007/s42979-022-01043-x

Taleb, N. N. (2012). *Antifragile: Things that gain from disorder.* Random House.

Vuković, D. B., Dekpo-Adza, S., & Matović, S. (2025). AI integration in financial services: A systematic review of trends and regulatory challenges. *Humanities and Social Sciences Communications, 12*(1), 1-29. https://doi.org/10.1057/s41599-025-04850-8

Wierckx, P. J. (2021). Thinking beyond the hiring and firing of asset managers: A new framework truly aligning asset owners with asset managers. *Journal of Accounting and Finance, 21*(2). https://doi.org/10.33423/jaf.v21i2.4244

Bridging the Gap between Saving and Income in Retirement: A Field Experiment to Drive the Adoption of Retiree-Friendly Installment Features

CHRISTELLE T. NGNOUMEN[1], CONEY H. LU, AND SEAN R. ASBURY

Voya Financial

Click or scan QR code
to watch video summary

Workers in United States defined contribution retirement plans spend decades focused on accumulating their savings for retirement. As retirement approaches, they face a fundamentally different challenge: converting their accumulated assets into sustainable income. Behavioral science offers valuable tools for easing this shift. Installment-based withdrawals are automated mechanisms that can convert retirement portfolios into scheduled, recurring payments and reduce cognitive load, mitigate behavioral biases, and mitigate risk. This chapter examines the role of behaviorally informed design and communication in connecting older investors to retiree-friendly decumulation features, such as installment-based withdrawals, on digital retirement platforms. Drawing on evidence from a field experiment, it evaluates targeted messaging designed to increase the awareness and adoption of installment-based withdrawal features. The findings suggest that messaging framing may increase engagement with installment-based withdrawal features among plan participants approaching retirement.

Introduction

The United States' aging population is rapidly increasing and is projected to reach 82 million by 2050 (World Bank, 2026). These demographic shifts have highlighted the need to support a swiftly growing cohort of Americans turning 67 each year with their financial planning needs. Over the past 50 years, the United States retirement system has shifted from defined benefit (DB) pensions—which are funded by an employer and automatically guarantee an income stream in retirement—to defined contribution (DC) retirement plans like the 401(k), which are funded and managed by individual employees. Today, roughly half of private-sector workers participate in a DC plan, and more than 730,000 DC plans cover over 86 million active savers (Cerulli Associates, 2025; Employee Benefit Research Institute, 2025; Investment Company Institute, 2025). This transition has placed the responsibility for key retirement decisions, such as how much to save, how to invest, and ultimately when and how to draw down those savings, squarely on individuals.

Behavioral design and behaviorally informed communications can play a critical role in bridging the gap between accumulation and decumulation by connecting older investors to retiree-friendly features like installment-based withdrawals on online retirement platforms. These features allow investors to withdraw a fixed amount from their retirement investments at regular intervals. Withdrawal amounts

1 Corresponding Author: christelle.ngnoumen@voya.com

and schedules are determined by the investor, and the feature helps schedule and generate recurring payments while a remaining balance in investments continues to grow and sustain income levels over time through the return on these investments. As structured decumulation tools, installment-based withdrawals can help reduce cognitive load, bias in decision-making, and risk.

> "Behavioral design and behaviorally informed communications can play a critical role in bridging the gap between accumulation and decumulation."

The following examines the importance of connecting older investors to retiree-friendly features that support the transition from accumulation to decumulation, with particular attention paid to installment-based withdrawals offered on digital retirement platforms. It presents findings from a field experiment that tested targeted, behaviorally informed messaging designed to increase the awareness and adoption of these tools. By evaluating how targeted communication strategies can reduce behavioral barriers and encourage participation, the chapter highlights the role of behavioral science in facilitating the connection to sustainable retirement income options for investors both nearing and in retirement.

Easing the Transition From Work to Retirement: Advantages of Installment-Based Withdrawal Features for Older Investors

For decades, participants in DC retirement plans have been conditioned to accumulate assets. Retirement disrupts this mindset, requiring a shift from saving to spending. Installment-based withdrawals can ease this transition by recreating the familiar cadence of earned income. Regular paycheck-like payments provide continuity and structure at a time when retirees may feel unanchored due to changes in professional identity and cognitive processing.

Reducing Cognitive Load

Decumulation requires pre-retirees and retirees alike to make repeated judgments about withdrawal timing (i.e., when to take out their money), withdrawal amounts (i.e., how much money to take out), market conditions (i.e., how current and future economic environments may affect their portfolio), tax implications (i.e., how different withdrawal strategies influence their tax burden), and long-term sustainability (i.e., whether their savings will support their spending needs throughout retirement), all of which are tasks that demand working memory, numerical reasoning, and future-oriented planning. These demands can be especially taxing for older adults, who often experience age-related declines in processing speed and executive functioning (Karjalainen, 2025; Strough et al., 2021), making financial decision-making more effortful and more susceptible to bias (Keane & Thorp, 2016). Atshan, Angrisani, and Lee (2025) found that individuals in the U.S. experience the largest decline in wealth following the onset of cognitive decline compared to individuals from other countries—a disparity that seems largely driven by the structure of DC retirement funding in the U.S. that requires individuals to make more complex decisions about the decumulation of retirement assets.

Installment-based withdrawals reduce cognitive burden by automating key decisions. Once a withdrawal schedule is set, participants are no longer required to reassess repeatedly how much or when to withdraw, thus minimizing decision points, choice overload, and decision fatigue.

Reducing Bias in Decision-Making

Installment-based withdrawals also limit the influence of well-documented cognitive and emotional tendencies that become more pronounced with age. Many pre-retirees and retirees enter decumulation with deeply engrained accumulation habits, heightened loss aversion, and strong preferences for the status quo. These tendencies can distort withdrawal behavior by making them reluctant to draw down assets (Olafsson & Pagel, 2018), to view withdrawals from savings as a loss (Brown et al., 2013; Shu & Shu, 2018), to be overly reactive to market fluctuations (Lal et al., 2024; Korniotis & Kumar, 2011), or to rely on recent events in a manner that prevents long-term thinking.

Related to the reluctance to draw down assets, spending from qualified accounts like 401(k)s and IRAs only increases once Required Minimum Distributions (RMDs)—minimum amounts individuals must withdraw from their retirement accounts each year, starting when they reach age 73 —begin, suggesting that retirees avoid drawing down assets unless, or until, forced by institutional rules. Even still, the effect of RMDs on decumulation behavior is modest (Mortenson et al., 2019). Blanchett and Finke (2025) analyzed how Americans spend their money in retirement and found that whereas retirees consume around 80% of guaranteed lifetime income sources such as social security, pensions, and private annuities, they only consume about half (50%) of investment savings from their retirement accounts, leading to significant underspending of their total savings. This suggests that mental accounting leads pre-retirees and retirees to treat 'income' and 'savings' differently—and in a manner that impacts decumulation behavior.

By introducing a regular and pre-committed withdrawal structure, installment-based withdrawal features promote greater behavioral discipline. Empirical evidence links these features to fewer discretionary withdrawals, more stable withdrawal rates, and a lower risk of premature asset depletion (IRALOGIX, 2024; Pang & Warshawsky, 2025). Pre-commitment helps investors avoid impulsive changes and reduces the tendency to overweight recent market performance.

Mitigating Risk

A key risk in retirement is sequence-of-returns risk, where market downturns early in retirement can permanently reduce income if withdrawals during downturns are large or reactive (Clare et al., 2025). Installment-based withdrawals can mitigate this risk by promoting consistent, measured withdrawals and discouraging large drawdowns during market declines, thereby supporting more sustainable income over retirement.

Barriers to Adoption

Despite the known advantages of installment-based withdrawal features, and despite being offered by a majority (70%) of retirement plans, adoption of these features remains low (Institutional Retirement Income Council, 2024). Most large retirement recordkeepers and industry researchers report that only about 15 - 40% of retirement participants who have access to in-plan withdrawal features like installments use them, even though they are designed to function as structured decumulation pathways that provide retirees with a predictable and sustainable method for generating retirement income while preserving an opportunity for tax-deferred or tax-free growth.

> "Only about 15 – 40% of retirement participants who have access to in-plan withdrawal features like installments use them."

There are several barriers to the adoption of installment-based withdrawal features, including cognitive, motivational, and environmental ones. Many investors in retirement plans don't realize that they have access to installment-based withdrawal features through their retirement plan. Moreover, when there is awareness of access to these features, many still do not fully understand what they are, how they work, or their value. Low awareness and low understanding are two related cognitive barriers to the adoption of installment-based withdrawal features. Research tracking pre-retirees' and retirees' perceptions and behaviors found low familiarity with retirement income features, including installment-based withdrawals, confusion about withdrawal strategies, and a preference for lump sum withdrawals even when more tax-efficient and structured withdrawal options were available (LIMRA, 2025; LIMRA & Ipsos, 2025). Using real-world data, Horan (2024) found that most retirees taking taxable distributions withdraw lump sums or annually and rarely use more frequent withdrawal schedules.

Installment-based withdrawals are often poorly placed in digital flow, making the features difficult to find; furthermore, poor information architecture that hinders findability and the ability to take action are two additional environmental barriers to their adoption.

Taken together, these obstacles to adoption raise a natural question: how can behaviorally informed messaging increase engagement with installment-based withdrawal features?

A Field Experiment: Testing Behaviorally informed Messages to Improve the Adoption of Installment-Based Withdrawal Features

A field experiment was conducted with Voya retirement plan participants aged 55 and older to test and identify which behavioral messaging framing could most effectively increase engagement with and the adoption of installment-based withdrawal features.

Experiment Design

In total, 2,043 participants aged 55 and older were randomly assigned to view one of three message variations while visiting a page about withdrawals on their retirement plan website. To ensure visibility and salience, the messages were placed at the top of the page and in a consistent place across all three conditions. An additional group of 630 participants of matching characteristics (age; gender; income) who visited the same page without messaging during the same period the experiment ran were included as a control for analysis (Figure 1).

The experiment targeted individuals aged 55 and older in plans with online installment-based withdrawals features who were visiting a page to explore their withdrawal options on the retirement website to ensure that the behavioral messages they received would be both timely and directly aligned with their actions. Prior research indicates that similarly aged individuals visited the same page the most for reasons including exploring their options while preparing for an upcoming retirement decision, considering full lump sum withdrawals, exploring penalty-free withdrawal rules, or satisfying general curiosities while browsing. By engaging participants at the precise moment they were considering their retirement options, each messaging intervention addressed a specific barrier to installment adoption and was connected to a clear action path to facilitate engagement with and the adoption of the feature.

The independent variable included the messaging shown to participants on top of the webpage. Messaging consisted of three behaviorally informed messaging interventions designed to address cognitive and motivational barriers to installment-based withdrawal adoption. Across all three conditions, the messaging was placed at the top of the page to maintain consistency in the experimental design and to reduce environmental barriers to adoption by improving the findability of the feature on the page and within the plan experience (Figure 1).

Outcomes were measured at the participant level and defined based on whether a participant was exposed to the messaging treatment at any point during the study window. Two primary dependent variables were examined. Participant *Interest* was assessed based on the proportion of participants who viewed each of the message variations and clicked on the respective call-to-action (CTA) button to learn more or get additional information about installment-based withdrawals. *Adoption* of installment-based withdrawal features was assessed based on the number of clicks-to-enroll, or by the number of participants who viewed each of the message variations and subsequently signed up for installment-based withdrawals. A secondary dependent variable, *Retirement Income Resource Engagement,* was assessed based on the proportion of participants who viewed each of the message variations, clicked on their respective CTAs, and then clicked on an additional link to access a resource center with information about how to manage retirement income and withdrawals strategically (e.g., factoring in longevity, inflation, taxes, medical expenses, market downturns).

Experimental Conditions and Behaviorally Informed Messaging

Education-Focused Messaging: This message variation focused on explaining the nature of installment-based withdrawals, and how they work, to educate retirement participants about the feature as well as raise awareness of the fact that their plan offered it. This message variation was included to address cognitive barriers to installment-based withdrawal adoption.

Benefits-Focused Messaging: This message variation highlighted the benefits of installment-based withdrawals, including their ability to help participants reduce the risk of outliving their savings by setting up regular retirement paychecks at safe withdrawal rates. Inflation and not having enough money to support long-term living expenses are two top factors that cause the most financial stress for active and retired retirement participants (Cerulli Associates, 2025). They were therefore highlighted in the benefits-focused messaging. This message variation was included to address motivational barriers to installment-based withdrawal feature adoption.

Ownership-Focused Framing: This message variation applied the behavioral principle of endowment to position installment-based withdrawals as an active benefit that participants already had access

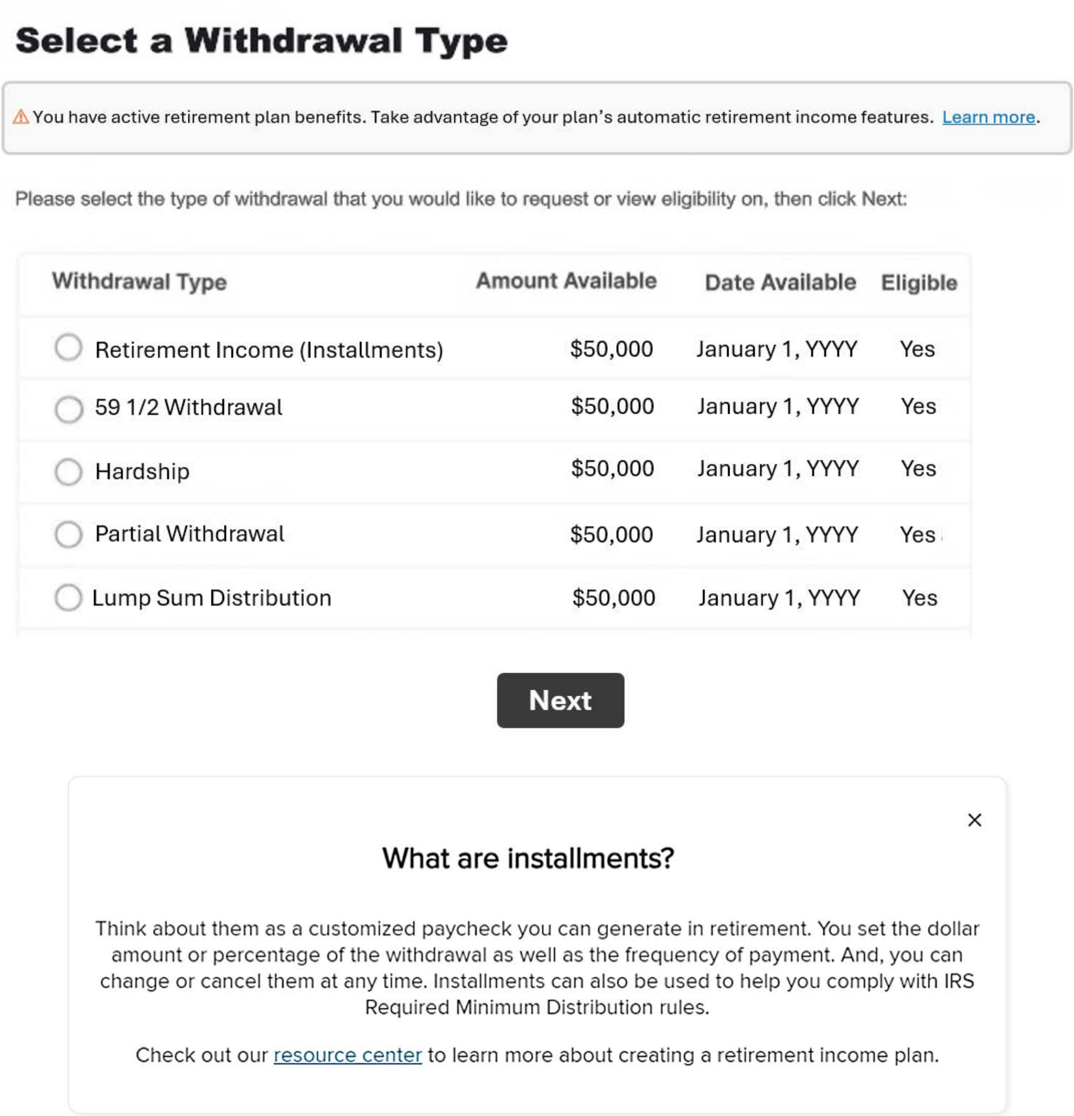


Figure 1: Representation of the online retirement plan webpage and the placement of behaviorally informed messaging interventions. The example shown features the ownership-focused message (top). A pop-up message with additional information about installment-based withdrawal features appears after a participant clicks on the "Learn More" link (bottom).

to but were simply not using. Research has shown that people place higher value on things they feel they already own (Kahneman et al., 1990). By framing installment-based withdrawal features as an asset participants possessed, the message aimed to increase its perceived value and participants' engagement with it. This message variation was included to address motivational barriers to installment-based withdrawal feature adoption (Figure 2).

Participants

A total of 2,673 retirement plan participants aged 55 and older were recruited (671 were randomly assigned to the Education-Based Messaging condition; 648 to the Benefits-Based Messaging condition; 724 to the Ownership-Based Messaging condition; and 630 participants matched on age, gender, and income were assigned to a control condition). Participant ages ranged from 55 to over 65 years (35% were aged 50 – 59; 31% were aged 60 – 65; 34% were aged 65 and older). The sample was 49% female, 41% male, and 10% had unknown gender identification. Demographics (e.g., age, gender, income) did not differ by condition. Participants' behavioral data were collected from August 14th, 2024, to November 13th, 2024.

Results

Participants in the Ownership-Focused Messaging condition ($p_{ownership}$ = 5.5%, N = 724) showed significantly higher levels of interest in installment-based withdrawals compared to those in the Education-Focused

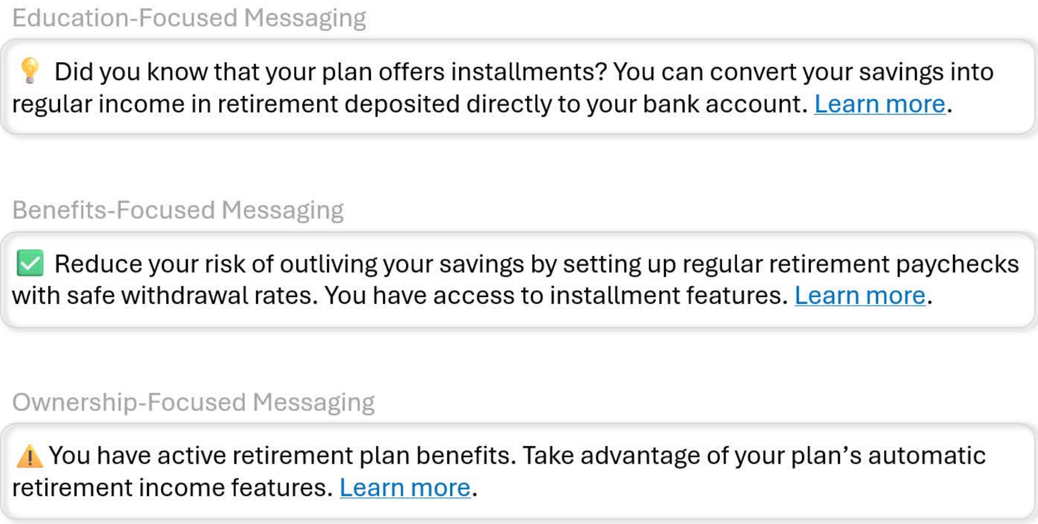


Figure 2: The three messaging interventions.

Messaging condition ($p_{education}$ = 3.3%, N = 671), Z = 2.03, *p* = .040, *h* = .11, and compared to those in the Benefits-Focused Messaging condition ($p_{benefits}$ = 2.9%, N = 648), Z = 2.36, *p* = .018, *h* = .13. Education-Focused Messaging drove greater interest in installment-based withdrawals compared to Benefits-Focused Messaging, Z = 1.97, *p* = .048, *h* = .02.

About 45% of participants in the Ownership-Focused Messaging condition who clicked on the *Learn More* CTA also engaged with additional resources about retirement income planning (i.e., *Retirement Income Resource Engagement)* provided in the pop-up modal. This compares to about 59% of participants in the Education-Focused Messaging condition who engaged with retirement income planning resources, and to about 32% of participants in the Benefits-Focused Messaging condition who engaged with retirement income planning resources.

Participants in the Ownership-Focused Messaging condition were significantly more likely to sign up for installment-based withdrawals within the digital system ($p_{ownership}$ = 4.1%, N = 724) compared to those in the Education-Focused Messaging condition ($p_{education}$ = 2.0%, N = 671), Z = 2.39, *p* = .017, *h* = .12, and compared to those in the Benefits-Focused Messaging condition ($p_{benefits}$ = 2.0%, N = 648), Z = 2.27, *p* = .023, *h* = .12. The difference in installment-based withdrawal signups between Education-Focused Messaging and Benefits-Focused Messaging conditions was not statistically significant. The signup rate for installment-based withdrawals among the control group of matched individuals was 3% ($p_{ownership}$ = 3.0%, N = 630). Although statistically nonsignificant, a marginal trend was observed involving greater signups for installment-based withdrawals among participants in the Ownership-Focused Messaging condition compared to the control group of matched participants.

"Participants in the Ownership-Focused Messaging condition were significantly more likely to sign up for installment-based withdrawals within the digital system."

An interaction between messaging frame and gender on engagement with installment resources was statistically significant, χ^2(2, *N* = 41) = 6.72, *p* = .035, *w* = .41. Women were more likely than expected to engage with installment resources under the Benefits-Focused frame, whereas men were more likely to engage with installment resources under the Ownership-Focused Frame. Education-Focused framing generated similar levels of engagement for both genders.

Discussion

As the retirement landscape evolves, installment-based withdrawal features offer pre-retirees and retirees flexibility, liquidity, and the potential for continued market growth to sustain income levels over time. Behaviorally informed messaging on digital retirement platforms can be used to connect older investors to these features. While the observed effects were small, experimental evidence suggests that Ownership-Focused Messaging was more effective at increasing engagement with installment-based withdrawal features and adoption compared to Education-Focused and Benefits-Focused Messaging. Several behavioral mechanisms might explain these results. The interventions tested were designed to address specific barriers to the adoption of installment-based withdrawal features. It is possible that Ownership-Focused Messaging rooted in the endowment effect was more effective at driving engagement among older investors by virtue of tapping into their greater tendency toward loss aversion. Prior research has established that people become reluctant to part with something once they feel it belongs to them (Ericson & Fuster, 2014; Kahneman et al., 1990; Knetsch, 1989). Drawing participants' attention to active plan benefits they already possessed by virtue of being in the plan, but were not yet using, possibly triggered concern about losing those valuable plan features.

> "Prior research has established that people become reluctant to part with something once they feel it belongs to them."

It is also possible that the framing of installment-based withdrawals as 'retirement income' helped boost engagement with the features. These results align with the prior empirical literature suggesting that mental accounting leads retirees to treat 'income' differently from 'savings', and in a manner that supports and hinders decumulation behavior, respectively (Blanchett & Finke, 2025). Retirees are more apt to engage with installment-based withdrawal features when they emphasize income rather than savings being drawn down from.

The gender effects observed align with prior research in psychology and behavioral economics that has found that women respond more to messaging that emphasizes reduced risk and security, while men respond more strongly to messaging emphasizing winning or maximizing gains. Relatedly, Ownership-Focused Messaging might have resonated more for men, given the nature of the endowment effect and how it transforms financial decisions into opportunities to maintain control, avoid loss, and optimize outcomes (Croson & Gneezy, 2009; Barber & Odean, 2001).

The interventions tested in this research increased the adoption of installment-based withdrawal features from 2% to 4%. While this was a statistically significant doubling, it still left 96% of participants not adopting the features. While there are genuine behavioral barriers the interventions likely partially overcame, it is also likely that a portion of older investors prefer not to use installments for other valid and rational reasons. Many retirees prefer to keep full control over when and how much they withdraw to allow them to adjust withdrawals in response to unexpected expenses like medical bills. Moreover, retirees with substantial Social Security benefits, pensions, rental income, or spousal income may not need regular withdrawals from their retirement accounts. Because their essential expenses are already covered, they may prefer to leave tax-deferred assets untouched to continue compounding or to preserve them for later-life needs or legacy goals. And finally, some retirees intentionally avoid installment-based withdrawals because they manage their tax brackets year-to-year, such as waiting until RMDs begin, taking larger withdrawals in low-income years, or avoiding withdrawals that would push them into a higher tax bracket.

Bridging the Gap Between Accumulation and Decumulation

This research offers several practical insights for benefits leaders and organizations seeking to support older investors and pre-retirees in their journey both to and throughout retirement. Employers can make retirement plans more retiree-friendly by helping workers engage with features that allow them to easily generate income in retirement. While comprehensive retirement income solutions that automatically calculate and translate savings into sustainable income streams are most ideal, such advanced features are not always available within employer-sponsored plans. In these cases, employers can still make their retirement plans more retiree-friendly by improving how existing features are presented and communicated. Highlighting available installment-based withdrawal options through clearer information architecture and communications may help workers more effectively engage with tools that support income generation in retirement.

Behavioral design and behaviorally informed messaging can help older investors access available tools to guide and optimize their decision-making in retirement, bridging a critical gap between accumulation and decumulation. When combined with intuitive information architecture, simple educational materials, and targeted messaging, these approaches can help pre-retirees stay on track by helping them better understand their options, overcome financial literacy barriers, and make more confident decisions.

Intuitive Information Architecture

Messaging about installment-based withdrawals and related retiree-friendly features that is embedded directly into participants' action paths is more likely to influence behavior at the exact moment a decision is being made. For example, communications or messaging about installment-based withdrawals that are positioned on or at the top of a page participants visit to browse their retirement resources, explore their withdrawal options, and/or allocate their investments is going to be more effective than emails or printed materials operating outside of the decision environment.

Simple Educational Materials

Providing clear, easy-to-understand educational materials alongside withdrawal-based installment options can play a crucial role in overcoming financial literacy. Research shows that retirees often struggle with concepts like sequencing risk, withdrawal rates, and longevity planning. Plain-language explanations, visual examples, and step-by-step guidance tailored to the installment-based withdrawal feature itself can reduce the cognitive effort required to understand how the option works and how to use it to generate income sustainably.

Behavioral Targeting and Personalization

Behavioral targeting and personalized outreach can help older investors engage with retiree-friendly features, such as withdrawal-based installments, at moments when they are most receptive. Many individuals approaching retirement may not actively seek out information about withdrawals or income strategies, even when they need it. Behavioral signals—such as participants who are 55 years and older and have recently visited a withdrawal-related page—can be used to deliver timely and relevant digital nudges or email campaigns that meet people exactly where they are in their decision-making process.

> "Behavioral targeting and personalized outreach can help older investors engage with retiree-friendly features, such as withdrawal-based installments, at moments when they are most receptive."

The observed differential responses to messaging by gender (e.g., women responded more to Benefits-Focused Messaging; men responded more to Ownership-Focused Messaging) also suggest that personalized messaging strategies could be substantially more effective than one-size-fits-all approaches. Plans offering installment features should consider offering multiple entry points as well as different framings to engage a more diverse set of participants.

THE AUTHORS

Christelle Ngnoumen is AVP of Behavioral Science Research at Voya Financial, where she leads research and experimentation efforts focused on helping people boost their savings and benefits utilization and on optimizing the customer experience. As a behavioral scientist and research leader, she leverages behavioral science to design digital experiences and communications that help people make better decisions and improve their wealth and health outcomes. Prior to Voya, she led behavioral science at Headspace, a leading provider of wellbeing and mental health solutions. She has also worked with companies and government agencies on digital experiences and services to improve social outcomes. Christelle holds a Ph.D. in Experimental Psychology from Harvard University.

Coney Lu is a staff data analyst specializing in experimentation, product analytics, and data-driven decision-making in the fintech industry. She has extensive experience designing and analyzing large-scale A/B tests across millions of users to optimize financial product engagement, retention, and conversion. Her work focuses on KPI development, risk-aware metric design, and scalable analytics solutions. She partners cross-functionally with product, engineering, and risk teams to translate complex data into actionable insights for financial services. Her interests include causal inference, user behavior modeling, and experimentation frameworks for improving financial product performance.

Sean Asbury is a senior data analyst at Voya Financial, where he focuses on behavioral and retirement analytics to help improve participant outcomes, client health insights, and data-driven decision-making across the business. Sean brings deep analytical rigor to applied business problems. He designs and maintains enterprise-scale analytics and executive dashboards that integrate data from multiple sources to inform strategy around savings behavior, engagement, and asset trends. He holds a Master of Science in Biostatistics from the University of California, Los Angeles, and began his career in the medical field, where he supported a kidney transplant team through statistical analysis and outcomes research. He has also worked with consulting hospitals to evaluate performance and design data-driven approaches to improve patient outcomes.

Acknowledgements

We would like to thank Tom Armstrong, Daniella Listro, Taylor Leung, and Jennifer Chenail for their feedback and collaborative support.

REFERENCES

Atshan, S., Angrisani, M., & Lee, J. (2025). Cognitive function and household financial decisions at older ages: A cross-country analysis (Wharton Pension Research Council Working Paper no. 2025-05). *SSRN*. https://dx.doi.org/10.2139/ssrn.5331809

Barber, B., & Odean, T. (2001). Boys will be boys: Gender, overconfidence, and common stock investment. *Quarterly Journal of Economics*, *116*(1), 261-292. https://doi.org/10.1162/003355301556400

Blanchett, D., & Finke, M. (2025). Retirees spend lifetime income, not savings. *Financial Planning Review*, *8*(3), e70010. https://doi.org/10.1002/cfp2.70010

Brown, J. R., Kling, J. R., Mullainathan, S., & Wrobel, M. V. (2013). Framing lifetime income (NBER Working Paper 19063). *National Bureau of Economic Research*. https://doi.org/10.3386/w19063

Cerulli Associates. (2025). *U.S. retirement markets: Growth, trends, and opportunities*. Cerulli Research Report. https://www.cerulli.com/reports/us-retirement-markets-2025

Clare, A., Seaton, J., Smith, P. N., & Thomas, S. (2025). Saving to decumulate: A lifetime journey based on the Perfect Contribution Rate. *Journal of Retirement*, *13*(1), 28-45. https://doi.org/10.3905/jor.2025.1.180

Croson, R., & Gneezy, U. (2009). Gender differences in preferences. *Journal of Economic Literature*, 47(2), 448-474. https://doi.org/10.1257/jel.47.2.448

Employee Benefit Research Institute. (2025). *Retirement plan participation and the current population survey*. EBRI Issue Brief. https://www.ebri.org/publications/research-publications/issue-briefs/content/retirement-plan-participation-and-the-current-population-survey--the-impact-of-the-retirement-account-questions--2018-2023

Ericson, K. M. M., & Fuster, A. (2014). The endowment effect. *Annual Review of Economics*, 6, 555-579. https://doi.org/10.1146/annurev-economics-080213-041320

Horan, S. (2024). Optimal withdrawal frequency for sustainable retirement withdrawals. *Financial Planning Review*, 7(2), e1183. https://doi.org/10.1002/cfp2.1183

Institutional Retirement Income Council. (2024). *Retirement income reimagined: How systematic withdrawals are reshaping DC plans*. IRIC White Paper. https://iricouncil.org/wp-content/uploads/2024/11/Retirement-Income-Reimagined-How-Systematic-Withdrawals-are-Reshaping-DC-Plans.pdf

Investment Company Institute. (2025). *The U.S. retirement market, 2024*. ICI Research Report. https://www.ici.org/statistical-report/ret_25_q4

IRALOGIX. (2024). *Installment withdrawals and retirement income outcomes*. IRALOGIX Research Brief. https://iralogix.com/nearly-half-of-retirees-lack-a-structured-decumulation-strategy-raising-concerns-over-rapid-depletion-of-savings-new-survey-finds/

Kahneman, D., Knetsch, J. L., & Thaler, R. H. (1990). Experimental tests of the endowment effect and the Coase theorem. *Journal of Political Economy, 98*(6), 1325-1348. https://doi.org/10.1086/261737

Karjalainen, M. (2025). *Cognitive decline and financial wealth at older ages*. Institute for Fiscal Studies (IFS) Report. https://doi.org/10.1920/re.ifs.2025.0041

Keane, M. P., & Thorp, S. (2016). Complex decision making: The roles of cognitive limitations, cognitive decline, and aging. In *Handbook of the Economics of Population Aging* (Vol. 1, pp. 661-709). North-Holland.

Korniotis, G. M., & Kumar, A. (2011). Do older investors make better investment decisions? *Review of Economics and Statistics, 93*(1), 244-265. https://doi.org/10.1162/REST_a_00053

Lal, S., Nguyen, T. X. T., Bawalle, A. A., Khan, M. S. R., & Kadoya, Y. (2024). Unraveling investor behavior: The role of hyperbolic discounting in panic selling behavior on the global COVID-19 financial crisis. *Behavioral Sciences*, *14*(9), 795. https://doi.org/10.3390/bs14090795

LIMRA. (2025). *Retirement income perceptions and withdrawal behavior*. LIMRA Secure Retirement Institute Report. https://www.limra.com/en/research/retirement

LIMRA, & Ipsos. (2025). *Consumer attitudes toward retirement income strategies*. Joint Research Report. https://www.limra.com

Mortenson, J. A., Schramm, H. R., & Whitten, A. (2019). The effects of required minimum distribution rules on withdrawals from traditional IRAs. *National Tax Journal*, 72(3), 507-542. https://doi.org/10.17310/ntj.2019.3.02

Olafsson, A., & Pagel, M. (2018). The liquid hand-to-mouth: Evidence from personal finance management software. *The Review of Financial Studies*, *31*(11), 4398-4446. https://doi.org/10.1093/rfs/hhy055

Pang, G., & Warshawsky, M. (2025). Evaluating the role of life annuities in retirement income strategies. *The Journal of Retirement, 13*(2), 20-38. https://doi.org/10.3905/jor.2025.1.185

Shu, S. B., & Shu, S. D. (2018). The psychology of decumulation decisions during retirement. *Policy Insights from the Behavioral and Brain Sciences*, 5(2), 216-223. https://doi.org/10.1177/2372732218790034

Strough, J., Karns, T. E., & Schlosnagle, L. (2011). Decision-making heuristics and biases across the life span. *Annals of the New York Academy of Sciences*, *1235*(1), 57-74. https://doi.org/10.1111/j.1749-6632.2011.06208.x

World Bank. (2026). *Population ages 65 and above (% of total population)*. World Development Indicators. https://data.worldbank.org

How Consumers Experience and Interpret Prices: A Behavioural Study

CLAIRE NOLAN, ALICE PEARCE, PAKAMON PRASITSUMRIT[1], AND ANGEL ZHOU

Dectech

Click or scan QR code
to watch video summary

Behavioural economics has shown that consumers do not rationally judge prices as absolute figures, but as psychologically constructed reference points shaped by context, framing, and cognitive biases (Barberis, 2013). Our perception of a price and the resulting elasticity is malleable and context dependent. We ran three behavioural experiments to illustrate how consumers' incoming expectations, the decision context and the communication of prices all significantly impact consumer behaviour and sentiment. Our findings show price elasticity is not a stable input to a pricing strategy but an outcome that is hard to measure, making understanding its causes, including the consumer perspective, essential.

> "Consumers do not rationally judge prices as absolute figures, but as psychologically constructed reference points shaped by context, framing, and cognitive biases."

Introduction

Consumer reactions to prices and resulting purchasing behaviour are fundamental to revenue and margin optimisation, yet brands often have a narrow understanding of how consumers engage with prices. Considerable resources are spent measuring price elasticity (Zhang, 2025), with little focus on its underlying causes. This assumes products have a natural elasticity that should determine a pricing strategy, instead of recognising that elasticity is shaped by multiple influences. These include external factors, such as macroeconomic conditions and competitor pricing, and controllable factors like historical pricing, current pricing architecture and communications strategy. While brands use elasticity to predict consumer responses to prices, behavioural experiments provide a much more effective approach.

Methodology

We conducted three behavioural experiments to analyse how consumer expectations and perceptions provide an additional lens to interpret and influence observed price elasticity. Primary research was conducted online in November 2025 using a nationally representative sample of 1,059 UK consumers aged 18+, recruited by our panel partner. Experiments covered a consistent set of nine industries to test generalisability – Grocery, Insurance, Utilities, Telecoms, Travel, Finance, Food Delivery, Retail and Pubs. To maximise statistical power and minimise respondent fatigue, respondents completed all three experiments for three of the nine industries, with industries randomly allocated to one experiment for each respondent. A demographics section was included for statistical controls and to identify demographic patterns.

1 Corresponding author: p.prasitsumrit@dectech.co.uk

Expected Price: How Should Consumer Expectations Shape Pricing Strategies?

Price elasticity measures demand sensitivity to price, but demand depends not only on absolute price, but also on how far it is from consumer expectations, or the "reference price" (Winer, 1986). From our experience in behavioural science, expectations derive from historic pricing, competitor pricing, substitutes, decision environment, macroeconomic context and more. Prices far above or below expectations immediately affect purchase rate and margin, and they have a longer-term brand impact. Furthermore, sticker shocks – surprisingly high prices – can have a spiralling effect, as negative word of mouth extends harm far beyond those who encounter the price. Retailers must optimise margin distribution by removing damagingly high prices while maintaining a few low 'hero' prices.

We studied the accuracy and spread of price estimates without reference prices. Respondents estimated the price – in pounds and pence – of a branded product from a detailed description (e.g., "S Hovis Seed Sensations loaf of bread (800g) from Waitrose"). Each respondent priced three products (e.g., bread, milk, cornflakes) from each of three randomly selected industries, with brands randomly assigned to each product (see Appendix). Respondents then rated how confident they were of their estimate from 0-100%, familiarity with the brand and when they last purchased the product. We tested three hypotheses:

H1: Consumers underestimate higher prices and overestimate lower prices.

H2: Expected prices will deviate more from real prices in low-engagement industries.

H3: Consumer price expectations differentiate value, mid-tier and premium brands.

Consumers provided reasonable estimates, highly correlated to real-world prices, but with a consistent tendency to overestimate, which was greatest in Travel, Utilities and Insurance (Figure 1) – all higher price industries. Prior research suggests that price stability influences whether consumers rely on memory or inference when forming judgments (Mazumdar et al., 2005). In volatile sectors like Travel, consumers rely less on recall and more on heuristics and industry-based anchors, as prices are route-specific and less observable. Travel, Utilities and Insurance are infrequently purchased, and both Utilities and Insurance use bespoke pricing, so there are fewer relevant reference points. Retail is the only industry that was on average underestimated, but

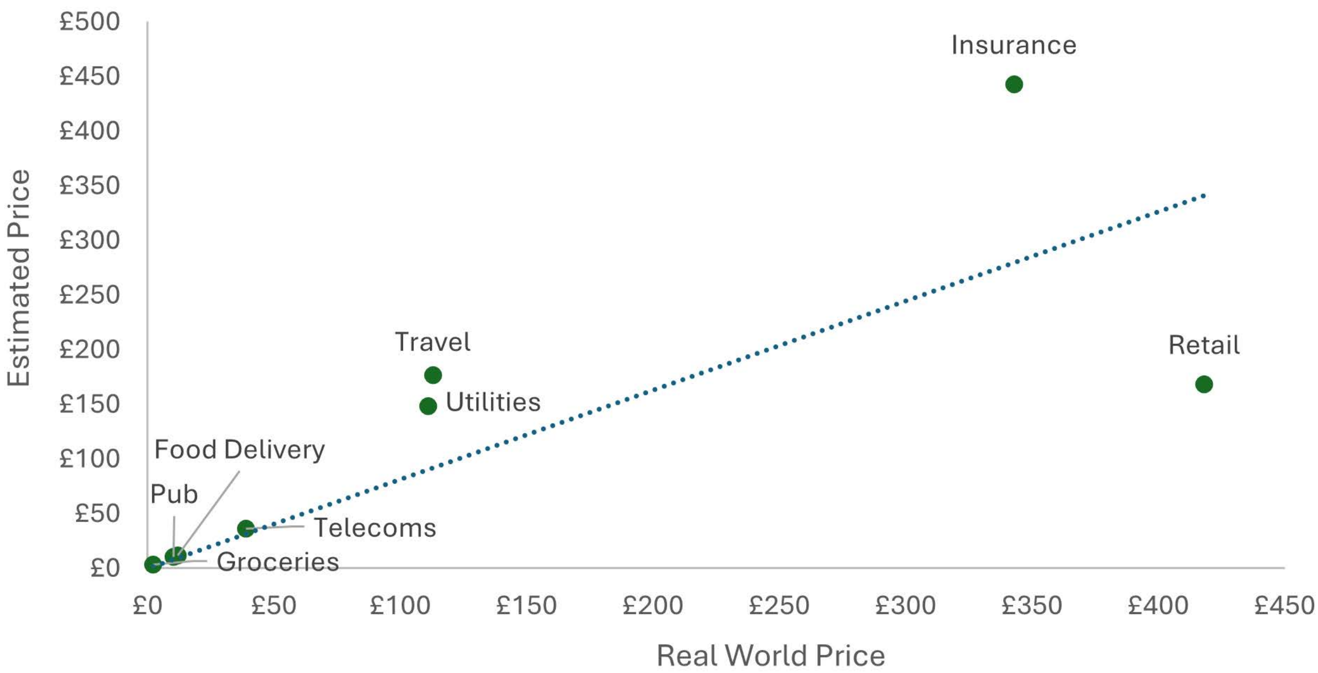


Figure 1: Estimated price vs. real price by industry. *Source:* Dectech fieldwork, July 2025 (N = 1,059 nat. rep.). Respondents were asked to estimate their expected price to pay for three products across three brands, within up to three industries. The graphic shows the relationship between the average expected price in each industry and the average of the corresponding real-world prices.

this was due to underestimations of how expensive Gucci is in actuality (Table 1). Estimates for M&S and H&M were much closer to real-world prices.

Consumers who have recently purchased a product make significantly lower estimates than those who haven't – on average 25% lower (p <.001) – bringing them closer to real-world prices. Similarly, those who are familiar with a brand tend to estimate significantly lower prices – on average 12% lower (p <.001). This helps explain the overestimates seen in Utilities and Insurance, where those less familiar with the products and brands gave particularly high price estimates.

Within an industry, brands can command substantially different price expectations. Consumers may not agree that a brand is worth a higher price, but they nevertheless form price hierarchies. The degree of differentiation varies by industry. Travel and Retail show a clear premium brand – British Airways and Gucci, respectively – for which price estimates are statistically significantly higher than the two other brands (Table 1). British Airways is estimated to be a +45% premium beyond Ryanair prices, driven by consistent brand positioning by both operators. In addition, consumers know that Gucci is a luxury brand, but they estimate H&M to be 10% cheaper than M&S.

Pubs show a differentiated value brand, in that consumers expect prices to be at least 14% lower in Wetherspoons, which is unsurprising, given its brand positioning. In Groceries, Asda estimates are significantly lower than Waitrose (p = 0.035), with Sainsbury's sitting in between. Other industries show less variation.

Consumers have relatively consistent confidence in their estimates across brands in the same industry, with higher variation between industries. Confidence is highest for estimates relating to Grocery and lowest for Travel and Insurance (Table 2). From a behavioural science perspective, confidence reflects the ease with which consumers can retrieve consistent memory cues when forming judgments (Koriat, 1997). Frequent engagement with an industry increases the volume of memory cues: true for Grocery but less so for Insurance. Conversely, confidence will be lower for prices that are unstable across time and brands, as memory cues are less consistent. This will affect

Table 1: Estimated Price by Brand

Industry	Brand	Average Estimated Price	% Difference from Mid
Travel	British Airways	£213	+24%
	Vueling	£171	-
	Ryanair	£147	-14%
Retail	Gucci	£420	+ >100%
	M&S	£47	-
	H&M	£42	-10%
Groceries	Waitrose	£3.60	+11%
	Sainsbury's	£3.20	-
	Asda	£2.90	-9%
Pubs	Greene King	£11.70	+3%
	Marston's	£11.40	-
	Wetherspoons	£9.80	-14%

Source: Dectech fieldwork, July 2025 (N = 1,059 nat. rep.). Respondents were asked to estimate their expected price to pay for three products across three brands, within up to three industries. The table shows the average expected price by brand and industry.

Table 2: Estimated Confidence by Industry

Industry	Confidence (0-100%)
Groceries [bread, cornflakes or 2 litres of milk]	47%
Telecoms [broadband, mobile data or mobile phone contract]	41%
Utilities [electricity, fixed energy or variable energy contract]	39%
Pubs [Fish & chips, margarita cocktail or pint]	39%
Retail [high heels, silk tie or winter coat]	38%
Food Delivery [garlic bread, margherita pizza or meat feast pizza]	37%
Insurance [travel, car or home insurance]	33%
Travel [flight to Alicante, Barcelona or Madrid]	31%

Source: Dectech fieldwork, July 2025 (N = 1,059 nat. rep.). Respondents were asked to indicate their confidence in their price estimates for each product on a 0% to 100% scale in 10% increments. The table shows the average confidence by industry.

confidence in Travel, where knowing the price of one flight is not necessarily informative when estimating the price of a different route, airline, airport, date or time of day.

Our results highlight the importance of understanding the brand-specific expectations consumers bring to purchase decisions. Brands must know where consumers expect the price of each product to sit within their range and the market to truly optimise revenue and margin. This insight can shape higher-level strategies on pricing architecture, limit price shocks and identify prices to feature in above-the-line marketing. The tendency for unfamiliar consumers to assume higher prices is crucial for acquisition strategies and new brand or product launches.

> “Brands must know where consumers expect the price of each product to sit within their range and the market to truly optimise revenue and margin.”

Our results support some but not all hypotheses. H1 was not supported, as higher prices were not consistently underestimated. H2 was supported by larger deviations from real-world prices in low-frequency industries like Travel and Insurance, and by more accurate estimates from recent purchasers. H3 was supported for four industries, including Grocery, where consumers expect 11% higher prices at a premium brand like Waitrose and a 9% discount at a value brand like Asda, but they struggle to differentiate prices in industries like Insurance and Energy.

Dynamic Pricing: How Do Consumers Respond to Fluctuations in Price?

As noted, frequent fluctuations make it harder for consumers to estimate prices and compare value across brands. Industries like entertainment, travel and hospitality have shifted towards rapid dynamic pricing (Kopalle et al., 2023), but consumers experience fluctuations in most industries due to environmental factors like inflation and competitor activity, supply-side factors like supply chain efficiency and raw material cost, and demand-side factors like consumer trends and seasonality. Brands must therefore understand consumer responses to these fluctuations in order to dictate strategies on the volume, frequency and communication of price changes.

Our second experiment examined consumer perceptions of how fair price fluctuations are in different circumstances. Respondents were given a scenario in which they experienced a higher price – e.g., “You want to fly to Madrid in July, and you notice prices are higher than at another time” – chosen at random from nine scenarios, each from a different industry (see Appendix Table 6). Respondents always experience a price surge, but the scenario here frames this price as either a premium above a base price (Premium Framing) or a base price from which others receive a discount (Base Framing). Respondents were asked to rate the scenario's fairness. We then gave two reasons for the price change – one demand-related and the other supply-related – and asked how acceptable the explanations were in each case. Lastly, we asked if they had ever experienced such a surge themselves.

We were interested in differences between industries and consumers, and whether price fluctuations are more palatable based on how they're framed or explained. We began with the following hypotheses:

H1: Price changes are more acceptable in industries where consumers expect them to occur.

H2: Brands can significantly affect consumer reaction through the communication and explanation of a change.

We found considerable differences in fairness ratings between industries. Fluctuations were perceived to be most unfair in Insurance and Finance, and least unfair in Retail and Travel (Table 3). In behavioural science, fairness perceptions are shaped by expectation-consistent price movements: when consumers recall similar fluctuations, they believe changes are justified. Recall was most common in Utilities, Retail and Travel, where seasonality impacts price.

Significant differences across consumers were also observed. Given the same scenario, a multivariate linear regression found that men, younger consumers, those with a higher income and those in London all have significantly higher fairness ratings (Figure 2), likely in part due to lower price sensitivity.

We then investigated the impact of how the price change was framed, e.g., flight prices are higher in

Table 3: Perceived Unfairness Across Industries

Industry	Scenario & Premium/Base Framing	% Rating Unfair	% Experienced
Insurance	Car insurance is [higher in price before/lower in price outside of] holidays (e.g., a bank holiday)	48%	28%
Finance	Credit cards have [higher interest rates during/lower interest rates outside of] holidays like Christmas	47%	27%
Pub	Drink prices are [higher when there are/lower when there are no] major football games on	44%	40%
Telecoms	Broadband internet bills are [higher in months with high/lower in months with low] traffic	43%	32%
Groceries	Milk is [more expensive in/less expensive outside of] winter	41%	30%
Utilities	Cost per unit of energy is [higher in/lower outside of] winter	38%	57%
Food Delivery	Food prices are [higher at/lower outside of] mealtimes	38%	38%
Travel	Flight prices are [higher in/lower outside of] the summer holidays	33%	51%
Retail	Winter clothing items are [more expensive in/cheaper outside of] winter	30%	55%

Source: Dectech fieldwork, July 2025 (N = 1,059 nat. rep.). Respondents were asked to assess the fairness of surge pricing in each industry and to indicate whether they had experienced such prices in the past year. Each respondent provided one evaluation per industry, for up to three industries. Fairness was measured using a 7-point scale, where scores of 1-3 were classified as unfair. "% Experienced" refers to the percentage of participants who reported having previously experienced surge pricing within a given industry.

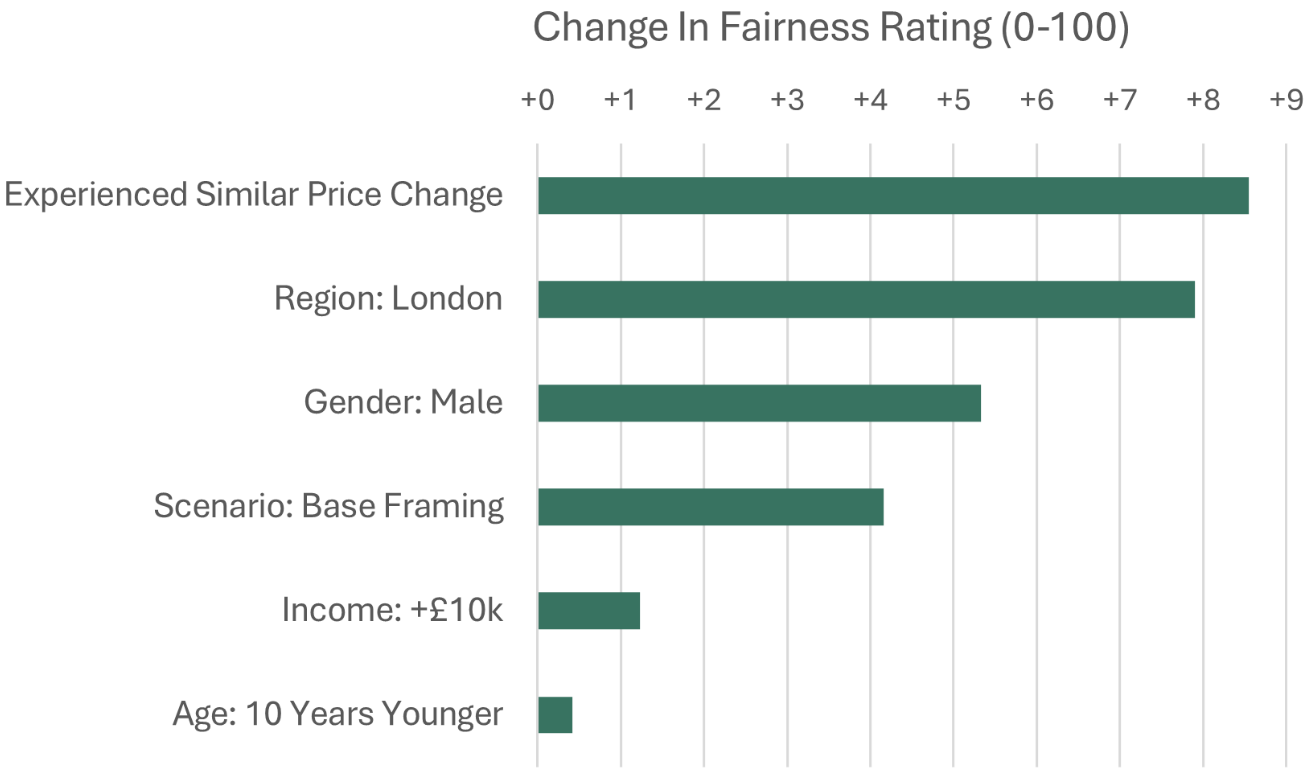


Figure 2: Consumers' characteristics and framing impact on fairness perceptions. *Source*: Dectech fieldwork, July 2025 (N = 1,059 nat. rep.). Respondents were told that prices are dynamic and that they were being charged a higher price than they might pay at other times. Scenario framing refers to presenting or explaining the dynamic pricing either as a higher price during high-demand periods or as a lower price during off-peak periods. Bars show the statistically significant betas from a multivariate linear regression with Fairness Rating as the dependent variable.

the summer holidays (Premium Framing) vs. flight prices are lower outside of summer holidays (Base Framing). Importantly, in both framings the scenario put respondents in the circumstances that mean they experience higher price (e.g., July in the Travel scenario above), the price was simply presented as higher now or lower at other times. Our results reveal a significantly higher perceived fairness when the current price is framed as the base with discounts at other times (Figure 2). This reflects loss aversion

Figure 3: Acceptability score by reason type. *Source*: Dectech fieldwork, July 2025 (N = 1,059 nat. rep.). The graphic shows the average acceptability score by reason type and industry.

(Thaler, 1980), i.e., consumers are more sensitive to perceived losses (surcharges) than equivalent gains (discounts), meaning a current surcharge feels more painful than missing out on a discount.

Finally, brands can significantly affect consumer reactions by how they explain price changes. When comparing demand- vs. supply-related explanations, the frustration caused by price variations is reduced significantly ($p = 0.008$) if the change is caused by the market force of increased demand rather than supply-side incompetence. Supply-side issues are deemed to be the provider's problem to fix, rather than passing on the cost to consumers. The strongest examples in this regard are in Retail, Travel and Finance (Figure 3). This is consistent with the literature on service failures, where customer satisfaction is harmed further if the provider is blamed for the failure (Anderson et al., 2007).

These results illustrate how brands can minimise negative reactions to price changes. There are products that consumers assume will vary in price, so retailers can maximise revenue by applying a more dynamic pricing strategy. Unexpected fluctuations are substantially more damaging. Where possible, a higher price should be framed as the base price from which one can discount, rather than framing the lower price as the base and increasing from there. Brands are also in control of the explanation, which should focus on how market dynamics affect price rather than supply-side incompetence.

"There are products that consumers assume will vary in price, so retailers can maximise revenue by applying a more dynamic pricing strategy."

These findings support both incoming hypotheses. Price changes are more acceptable in industries where such fluctuations are expected (H1), and consumer reactions are strongly shaped by how brands frame and explain price changes (H2).

Drip Pricing: How Does Pricing Structure Impact Purchase Intent?

Our final experiment focused on how pricing is structured and communicated across the customer

Group 1: Product Rating

Please imagine you want to get a flight from London to Barcelona.

How likely would you be to buy it at a price of £82.80?

Very Unlikely Very Likely
O O O O O O O O O O

Group 1: Total Basket Rating

Please see your final basket below:

Train ticket from London to Manchester	*£82.80*
Booking Fee	*£9.20*
	Total: £92.00

How likely would you be to purchase this product?

Very Unlikely Very Likely
O O O O O O O O O O

Figure 4: Questions seen by Group 1 participants. *Source*: Dectech fieldwork, July 2025 (N = 1,059 nat. rep.). Respondents were randomly assigned to two groups: Group 1 saw the base price in the first question and then additional charges in the second question, while Group 2 saw the total price in the first question and did not see the second question.

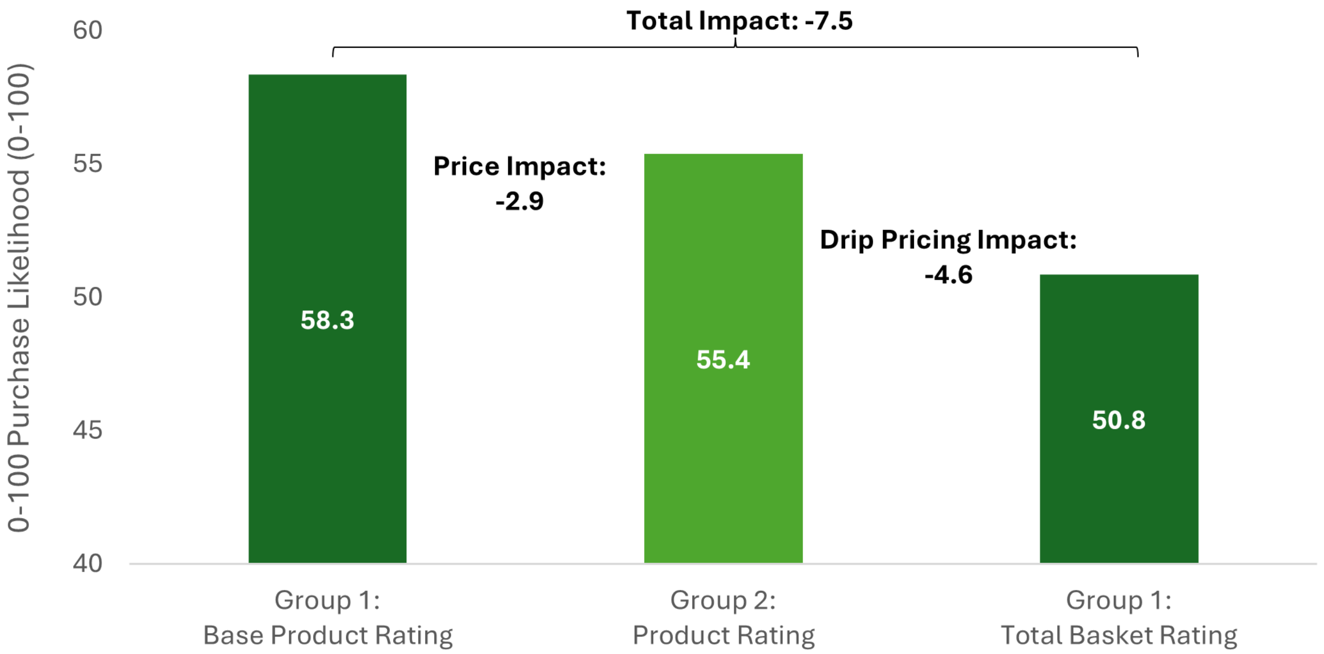


Figure 5: Impact of drip pricing. *Source*: Dectech fieldwork, July 2025 (N = 1,059 nat. rep.). Participants were split into two groups: Group 1 experienced drip pricing, seeing additional charges only at checkout, while Group 2 saw the full total price upfront from the start.

journey. Drip pricing – decomposing a price into multiple components that are presented sequentially to buyers (Rasch et al., 2020) – is used extensively in certain industries. Airlines, for example, often start with a base flight price before adding a range of extras – some optional and some required – that eventually build to a significantly higher basket.

Our experiment measured the behavioural impact of drip pricing using two different groups of respondents, assigned at random. Group 1, whose journey is shown in Figure 4, replicates the structure of a drip pricing journey. Initially, they see a base price (£82.80 in this example), and then an additional 10% charge is added to the basket before payment, yielding a total of £92.00. The additional charge was always 10% of the total price, so 90% of the total price was seen upfront. Group 2 replicates an equivalent online journey with only one step, as the total price of £92.00 is seen upfront as the product price.

The product was chosen at random from a list of nine, each from a different industry, and the total price for each product was based on real-world prices. By comparing the ratings and groups, we could measure how purchase intent was impacted by the addition of a charge in a drip pricing structure.

This design allowed us to consider the following hypotheses:

H1: Adding charges in a drip pricing structure has a negative impact on purchase intent, beyond the impact of the increased price.

H2: The impact of drip pricing on purchase intent differs by industry.

Our analysis found that drip pricing significantly decreases purchasing intent. Purchase likelihood decreased significantly – by 7.5 points on a 0-100 rating ($p < .001$) – between Group 1's initial product rating and their total basket rating (Figure 5). Group 2's product rating allows us to deconstruct this 7.5-point fall into the impact of the increased cost versus drip pricing irritation. Purchase likelihood decreases by 2.9 points between the product ratings of Group 1 and 2, showing the impact of the 10% increase in price. Hence, the majority of the Group 1 drop, namely 4.6 points, is attributable to the drip pricing ambush. Given the impact of a 10% price rise, the drip pricing frustration is equivalent to a 15% increase in price, which is a substantial portion of any product's gross margin.

> "Drip pricing frustration is equivalent to a 15% increase in price."

The negative impact of drip pricing was seen across most industries, but the scale of the impact varied substantially. It was most damaging in Retail, Grocery

Table 4: Impact of Drip Pricing by Industry

Industry	Product	Example Extra Charge	Total Price	Purchase Likelihood (0-100)		
				Group 1: Drip Price Rating	Group 2: Normal Rating	Drip Pricing Impact
Retail	Winter Coat	Delivery Fee	£45	60.1	71.3	-11.2
Grocery	Milk 2l	British Farmers Fee	£2	57.5	66.9	-9.4
Travel	Barcelona Flight	Booking Fee	£92	61.1	69.0	-8.0
Food Delivery	Margherita Pizza	Delivery Fee	£11	44.6	50.0	-5.4
Pub	Pint of Moretti	Table Fee	£7	37.0	41.6	-4.6
Telecoms	Broadband	Line Rental Fee	£38 pcm	49.7	53.7	-4.0
Utilities	Energy	Standing Charge	£131 pcm	49.6	51.0	-1.4
Insurance	Travel Insurance	Set Up Fee	£50	55.2	54.7	-0.5
Finance	Credit Card	Travel Perks Membership Fee	£15 pcm	43.6	43.6	0.0

Source: Dectech fieldwork, July 2025 (N = 1,059 nat. rep.). The table presents purchase likelihood for Group 1 and Group 2, along with the score differences between the two groups across industries included in the survey.

and Travel (Table 4), where it had a statistically significant impact on purchase likelihood (Retail $p = 0.003$, Grocery $p = 0.018$, Travel $p = 0.015$). These industries have the highest purchase intent, so the frustration of additional charges may be higher when consumers are more engaged and therefore more sensitive to perceived losses when additional charges appear later in the journey.

The results support both hypotheses: H1 is validated, as drip pricing reduces purchase intent through both higher total cost and the structure itself. H2 is also supported, with effects varying notably across industries, with the strongest effect seen in Retail, Grocery and Travel.

Our analysis highlights the dangers of drip pricing, which interrupts the purchase journey and frustrates consumers. Retailers implement this pricing structure because it allows them to advertise a lower base price to consumers, in their own comms and in competitive environments like price comparison websites. For some brands and industries, the resulting increase in the volume of consumers starting the purchase journey will be worth the risk of higher drop-out and negative perceptions of the brand. For others, the risk will not be worth it. Even in cases where the purchase rate is strong, drip pricing leads to lower satisfaction and a lower likelihood of future purchases with the brand (Santana et al., 2020). This brand harm may be weaker in industries with low purchase frequency, as consumers have forgotten the frustration when they come to purchase again, but the harm is greatest in industries with high frequency purchases and less differentiated, easily substitutable brands.

Conclusions

Price is a psychological construct rather than an absolute figure. Consumer price perception depends on how, by whom and in what macro and micro context it is communicated. This research illustrates the malleability of consumer responses to price and recommends how to structure, communicate and manage prices. These recommendations are based on consumer inputs, which are often underused. The impact of price, sales journey or communications changes can be tested in advance through rigorous experiments, instead of analysing noisy, inconsistent sales data. Brands should treat the consumer perspective as an input to their pricing strategy, not an outcome.

> “Brands should treat the consumer perspective as an input to their pricing strategy, not an outcome.”

Recommendations

Treat Elasticity as an Outcome: Price elasticity is an outcome of sales strategies rather than an input. Brands should identify and forecast the drivers of elasticity, including consumer price estimates, competitor pricing and historic pricing.

Align Margin to Expectations: Margin distribution across a range should reflect consumer price expectations. Brands should shift margin away from sticker shocks toward underpriced items, retaining a few ‘hero’ prices to be used in marketing.

Optimise Price Communication: Brands can significantly impact how prices are perceived via their framing. Promotions and discounts should support a long-term strategy where higher prices are seen as the base from which discounts apply, not unfair rises.

Avoid Drip Pricing: Drip pricing frustrates consumers and disrupts purchases, reducing purchase rates similarly to a 15% price rise. Brands should test their specific journey and avoid this practice, unless the low base price outweighs reduced purchase intent and potential brand damage.

Proactively Explain Price Changes: Price fluctuations are more damaging when unexpected. Brands should proactively explain changes, focusing on market causes rather than supply-side issues that may suggest incompetence.

Maintain Consistent Positioning: Brand positioning must be clear and consistent to shape price expectations. Consumers expect premium brands to charge up to 25% more than mid-market brands. Clear positioning matters especially for new or unfamiliar brands, which consumers tend to overprice.

Tailor Your Pricing Strategy: Our behavioural experiments reveal significant industry differences. Though there are recommended methodologies and frameworks, they must be tailored to the specific brand with a test-and-learn approach to develop specific, granular pricing strategies.

THE AUTHORS

Claire Nolan is a director at Dectech. Since joining in 2012, after completing an MSc in Psychological Research at the University of Oxford, she has helped telecoms, finance, retail and utilities clients understand consumer decision-making, using behavioural economics and experimental techniques.

Alice Pearce is a senior associate at Dectech. She holds an MSc in Behavioural & Economic Science from the University of Warwick and has several years’ experience applying behavioural insights commercially, from nudge design to experiment implementation and analysis.

Pakamon Prasitsumrit is a senior associate at Dectech. She holds an MSc in Business Analytics from the University of Warwick and works with financial services clients to present complex information clearly and support informed decision-making.

Angel Zhou is a part-time analyst at Dectech and a final-year BSc Psychology student at University College London. She specialises in statistical analysis, data visualisation and behavioural insight generation.

REFERENCES

Anderson, S. W., Baggett, S., & Widener, S. K. (2007). The impact of service operations failures on customer satisfaction: Evidence on how failures and their source affect what matters to customers. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.975832

Barberis, N. C. (2013). Thirty years of prospect theory in economics: A review and assessment. *Journal of Economic Perspectives*, 27(1), 173–196. https://doi.org/10.1257/jep.27.1.173

Kopalle, P. K., Pauwels, K., Akella, L. Y., & Gangwar, M. (2023). Dynamic pricing: Definition, implications for managers, and future research directions. *Journal of Retailing*, *99*(4), 580–593. https://doi.org/10.1016/j.jretai.2023.11.003

Koriat, A. (1997). Monitoring one's own knowledge during study: A cue-utilization approach to judgments of learning. *Journal of Experimental Psychology: General*, *126*(4), 349–370. https://doi.org/10.1037/0096-3445.126.4.349

Mazumdar, T., Raj, S. P., & Sinha, I. (2005). Reference price research: Review and propositions. *Journal of Marketing*, *69*(4), 84–102. https://doi.org/10.1509/jmkg.2005.69.4.84

Rasch, A., Thöne, M., & Wenzel, T. (2020). Drip pricing and its regulation: Experimental evidence. *Journal of Economic Behavior & Organization*, *176*, 353–370. https://doi.org/10.1016/j.jebo.2020.04.007

Santana, S., Dallas, S. K., & Morwitz, V. G. (2020). Consumer reactions to drip pricing. *Marketing Science*, *39*(1), 188–210. https://doi.org/10.1287/mksc.2019.1207

Thaler, R. (1980). Toward a positive theory of consumer choice. *Journal of Economic Behavior & Organization*, *1*(1), 39–60. https://doi.org/10.1016/0167-2681(80)90051-7

Winer, R. S. (1986). A reference price model of brand choice for frequently purchased products. *Journal of Consumer Research*, *13*(2), 250–256. https://doi.org/10.1086/209064

Zhang, W. (2025). An economic analysis of price elasticity: Model, challenges, and implications. *Advances in Economics, Management and Political Sciences*, *187*, 112–116. https://doi.org/10.54254/2754-1169/2025.bl23782

APPENDIX: METHODOLOGY

Expected Pricing

In this experiment, each respondent was assigned three of the nine industries at random. For each industry respondents were asked to estimate the price of three different products, with each brand randomly assigned to a product. Table 5 details the products and brands tested. The three brands were chosen for each industry to cover a range from more budget

Table 5: Expected Pricing – Variables Tested

Industry	Products	Brands
Food Delivery	Margherita Pizza, Meaty Pizza, Garlic Bread	Domino's, Pizza Express, Franco Manca
Groceries	2 Litres of Milk, Loaf of Bread, Box of Cereal	Asda, Sainsbury's, Waitrose
Insurance	Car Insurance, Home Insurance, Travel Insurance	Tesco Bank, Aviva, NFU Mutual
Pub	Pint of Beer, Fish & Chips, Cocktail	Wetherspoons, Greene King, Marston's
Retail	Winter Coat, High Heel Shoes, Tie	H&M, M&S, Gucci
Travel	London to Barcelona Flight, London to Madrid Flight, London to Alicante Flight	Ryanair, Vueling, British Airways
Telecoms	1 Month Mobile Data, 24 Month Phone Contract, 12 Month Broadband Contract	giffgaff, Vodafone, EE
Utilities	Fixed Energy Tariff, Fixed Electricity Tariff, Variable Energy Tariff	Outfox Energy, Octopus Energy, British Gas
Finance	Credit Card Interest, Savings Account Interest, ISA Interest	Santander, Lloyds, HSBC

Table 6: Expected Pricing – Variables & Framing Tested

Industry	Product	Scenario & Framing*
Food Delivery	Margherita Pizza Delivered for Dinner	Food prices are [higher at/lower outside of] mealtimes
Groceries	2 Litres of Milk in Winter	Milk is [more expensive in/less expensive outside of] winter
Insurance	Car Insurance Before a Bank Holiday	Car insurance is [higher in price before/lower in price outside of] holidays (e.g. bank holiday)
Pub	Pint of Moretti during Europa League	Drink prices are [higher when there are/lower when there are no] major football games are on
Retail	Winter Coat Purchased in January	Winter clothing items are [more expensive in/cheaper outside of] winter
Travel	London to Barcelona Flight for July	Flight prices are [higher in/lower outside of] the summer holidays
Finance	Credit Card Interest in December	Credit cards have [higher/lower] interest rates [during/outside of] holidays like Christmas
Telecoms	Broadband Bill for This Month	Broadband internet bills are [higher in months with high/lower in months with low] traffic
Utilities	Energy Bill for January	Cost per unit of energy is [higher in/ lower outside of] winter

*Higher framing uses the words before "/" in the brackets and lower framing the words following the "/" in the brackets

to premium/luxury. Respondents were also asked their confidence in said estimate (0%-100% in 10% intervals) and when the last time they encountered the product was.

We analysed respondents' price estimates by running linear regression models to test if independent variables such as product brand, brand awareness, confidence in the estimate or time since last purchase impacted the dependent variable of price estimates. Models also included demographic variables to identify and control for any effect these might have. If variables were found to have a significant impact on the dependent variable of price estimate, we often chose to run t tests and report results as the difference between two groups for ease of interpretation. These models were run overall and on a per industry basis to investigate differences between industries.

Whilst we have included information on Finance in the above table for completeness, it was not included in our cross-industry analysis. For Finance, we asked respondents to estimate interest rates, and whilst these insights were interesting in their own right, they were not comparable to estimated product prices.

Whilst Table 5 includes simple product descriptions, descriptions seen by respondents were more detailed. For example, Home Insurance was in fact "home insurance on a 3-bed 2-bath semi-detached house (with £200 excess)". The level of detail was chosen so that respondents had enough information to make an informed estimate, and equally, we could determine a real-world price through appropriate secondary research for each brand and product combination.

Dynamic Pricing

We presented respondents with a scenario where they experienced a surge in price. We asked them if the scenario was fair, on a 7-point scale from 'Very Unfair' to 'Very Fair'. We then gave them two reasons for the change in price and asked how acceptable the reasons are on a 5-point scale from 'Unacceptable'

Table 7: Expected Pricing – Reason Type & Framing

Industry	Reason & Framing*
Food Delivery	1. [More/Fewer] people are ordering food [at/outside] of mealtimes 2. Delivery platforms (e.g. Uber Eats, Deliveroo) charge restaurants [higher/lower] commissions [at /outside] of mealtimes
Groceries	1. [More/Fewer] people want to buy milk for hot beverages (tea/coffee) [in winter/ outside of winter] 2. Farms produce and supply [less/more] milk in [the winter months/ non-winter months]
Insurance	1. [More/Fewer] people want insurance cover when making [longer journeys during/ fewer or shorter journeys outside of] holidays 2. Repair garages and recovery services [raise/lower] their call-out rates [over bank holiday weekends/outside of holidays]
Pub	1. [More/Fewer] people go to the pub when there [are/are no] major football games on 2. The pub [requires/does not require] a TV licence to air football matches if there are [games/ no games] on
Retail	1. [More/Fewer] people want a thick coat in the [winter/summer] than in [summer/ winter] 2. There are [more/fewer] shipping and logistics fees due to [worse/better] weather conditions in [winter/summer]
Travel	1. [More/Fewer] people are going on holiday with their children [in/outside of] summer holidays 2. There are [higher/lower] airport fees/congestion surcharges due to [busier/ quieter] runways [in/outside of] summer holidays
Finance	1. [More/Fewer] people are using their credit card for Christmas/holiday shopping 2. Fraud prevention services are [more/less] costly [during/outside of] the holiday period
Telecoms	1. People use their home internet [more/less] in certain months, [increasing/ decreasing] traffic 2. [Extra/no extra] bandwidth needs to be purchased from infrastructure providers when traffic is [high/ low]
Utilities	1. [More/Fewer] people are using gas & electricity as it gets [colder & dark earlier/ warmer & dark later] [in/outside of] winter 2. Wholesale energy suppliers [increase/lower] their rates [in/outside of] winter months due to [limited/excess] supply

* 1. Is the Demand led reason and 2. The Supply led reason. Negative Reason Framing uses the words before "/" in the brackets and Positive Reason Framing the words following the "/" in the brackets

Table 8: Drip Pricing – Variables Tested

Industry	Product	Total Price	Additional Charge
Food Delivery	Margherita Pizza	£11	Delivery Fee, Faster Delivery Fee, Service Fee
Groceries	Milk 2L	£2	British Farmers Fee, Cotton Carrier Bag For Life, Card Payment Fee
Insurance	Travel Insurance	£50	Set Up Fee, Legal Expenses Cover, Admin Fee
Pub	Pint of Moretti	£7	Entry Fee, Table Fee, Service Fee
Retail	Winter Coat	£45	Delivery Fee, Faster Delivery Fee, Service Fee
Travel	Barcelona Flight	£92	Fuel Fee, Small Cabin Bag (in addition to a personal item), Booking Fee
Finance	Credit Card	£15 pcm	Travel Perks Membership Fee, Payment Protection Insurance (PPI), Admin Fee
Telecoms	Broadband	£38 pcm	Line Rental Fee, Min 4G promise, Admin Fee
Utilities	Energy	£131 pcm	Standing Charges, Priority Repair Cover, Admin Fee

to 'Acceptable'. Lastly, we asked if they had ever experienced such a surge themselves.

Whilst under all scenarios participants are experiencing the higher price, the scenario is framed as either prices being higher in the current context or lower in other contexts (see Table 6). We also tested a similar variation in the framing of reasons for the surge. Two reason types were presented – one demand-led and the other supply-led – and we also varied the framing of the reasons as impacting the current context or other contexts (see Table 7). For each respondent under each industry, the scenario framing, reason type and reason framing were selected randomly.

We report the proportion of respondents who rated the scenario as unfair (i.e., 1–3 on the 1–7 scale) for ease of interpretation. To identify statistically significant relationships, we ran linear regressions with fairness score for the scenario as the dependent variable and separate linear regressions with acceptability score for the reasons as dependent the variable to inform our conclusions. For interpretability, both ratings were transformed into 0-100 scores when used as dependent variables. In addition to demographic variables, our linear regression models on fairness scores included scenario framing and previous experience of the scenario as independent variables. Our acceptability score models included all the same variables with the addition of reason type as demand-led or supply-led and the reason framing. We did not see a consistent impact of reason framing so we reported the average across the two framings for each reason type.

Drip Pricing

In our final experiment, respondents were randomly assigned into two groups. Group 1 experienced drip pricing while Group 2 did not. Both groups were shown a product with a price and were asked their likelihood of purchasing. Group 1 saw 90% of the Total Price and Group 2 saw the Total Price (see Table 8). For Group 1, drip pricing was then simulated by introducing a previously unknown additional charge, such as an admin fee, which increased the overall price. This additional charge was shown alongside the product on a basket page. The additional charge was always 10% of the Total Price so that the basket summed to the Total Price. Group 1 were then asked their likelihood of purchasing again (see Figure 4 for question layout). Group 2 did not see the basket page. The additional charge was randomly chosen from three possible charges per industry (see Table 8).

To analyse the impact of drip pricing on purchase likelihood we first compared the two ratings given by Group 1, for the product and the total basket. We then compared the two Group 1 ratings to the one rating given by Group 2, to separate out the impact of the increased price from that of the drip pricing structure. To identify relationships of statistical significance, we ran a linear regression where the dependent variable was purchase likelihood at Total Price – i.e., Group 1's basket rating and Group 2's product rating. The regression included the independent variable of whether the respondent was in Group 1 or Group 2 alongside demographics and industry engagement.

Mobilization—A Whole-of-Community Approach to Tax Compliance: Behavioral Foundations

JOHAN BIRKELUND[1]

Norwegian Tax Administration

Click or scan QR code
to watch video summary

This article describes an initiative the Norwegian Tax Administration (NTA) refers to as 'Mobilization', i.e., a whole-of-community approach to preventing labor market crime, and situates it within contemporary behavioral tax literature. We argue that combining credible enforcement (power) with trust-building, third-party visibility, and targeted behavioral instruments offers a framework to understand high real-world compliance despite low audit rates. The article maps the instruments utilized in Mobilization—collaborative partnerships with major procurers, real-time tax certificates, risk-based site monitoring, the consumer program *Tettpå*, and the grassroots coalition *Medbyggerne*—to the core mechanisms identified in the literature: trust and power, tax morale, information trails/third-party reporting, cooperative compliance, and behavioral nudges/boosts.

Introduction

The benchmark Allingham-Sandmo (AS) model portrays tax reporting as an expected utility trade-off between the tax rate, the probability of audit, and the penalty conditional on detection (Allingham & Sandmo, 1972). While this model is internally consistent with clear predictions, it struggles to explain why modern tax systems see persistently high compliance, even with low audit probabilities and modest penalties. Large-scale field evidence from Denmark shows that the key to understanding this puzzle is the distinction between income subject to third-party reporting and self-reported income: evasion is near zero for the former and substantial for the latter (Kleven et al., 2011). This finding echoes a broader insight from behavioral tax research, in that while tax audits and fines are pivotal for tax compliance, it is not only contingent on classical deterrence, but is also shaped by non-pecuniary motivations such as trust, fairness, tax morale, norms, and the design of information and institutional processes (Luttmer & Singhal, 2014; Torgler, 2007; Alm, 2019; Slemrod, 2019).

Following this expanded understanding of compliance motivation, contemporary tax research identifies a more plural understanding of taxpayer behavior in which deterrence interacts with service quality, legitimacy, and social motivations (Alm, 2020; Slemrod, 2020; Alm & Kasper, 2023). In parallel, the OECD emphasizes building a 'tax culture' through education, assistance, communication, and institutional transparency to foster voluntary compliance (OECD, 2021). Reflecting this shift, several tax administrations have moved from purely deterrence-based strategies toward cooperative compliance models which emphasize transparent, real-time risk discussion and mutual obligations between authorities and taxpayers (e.g., HMRC's

1 johan.birkelund@skatteetaten.no

framework; OECD's ICAP).

Collectively, these strands suggest that effective tax administrations require approaches that enhance information trails, strengthen trust and perceived fairness, build competence among taxpayers and intermediaries, and focus audits on risk-based contexts. The Norwegian Tax Administration (NTA) extends this logic further, moving from bilateral cooperation with individual taxpayers to a multi-stakeholder model in which public procurers, private firms, financial institutions, and communities jointly contribute to monitoring and influencing compliance behavior.

> "The NTA's whole-of-community approach can be understood as synthesizing enforcement, service, and trust paradigms into a unified, market-facing, and multi-stakeholder architecture."

Alm's (2019) comprehensive review of contemporary tax research outlines three broad paradigms available to tax administrations, namely, enforcement, service, and trust. The NTA's whole-of-community approach can be understood as synthesizing these paradigms into a unified, market-facing, and multi-stakeholder architecture. By mobilizing major private and public procurers, consumers, and local communities, the approach aims to extend the reach of compliance mechanisms beyond the tax administration's direct enforcement capacity. This article examines how Mobilization operationalizes contemporary tax compliance insights in practice, and why this approach may be particularly well suited to addressing complex forms of labor market and tax crime. In doing so, the article contributes to the literature by demonstrating how behavioral and institutional insights can be translated into concrete administrative design.

The remainder of the article is organized as follows: section 2 presents the whole-of-community approach, section 3 explains the mechanisms underpinning the approach, and section 4 provides a summary and concluding remarks.

The Norwegian Mobilization Initiative

The NTA's Mobilization initiative combines enforcement, service, and trust-based collaboration across market actors, consumers, and local communities. Its core instruments include: (i) *Collaborative Partnerships* with large public and private procurers, which embeds tax compliance screening directly into procurement processes; (ii) real-time tax certificates and employee-list cross-checks that create continuous third-party visibility; (iii) risk-based, on-site monitoring with clear and immediate consequences through contract sanctions; (iv) *Tettpå*, a consumer-facing program designed to build competence, shift norms, and reduce demand for undeclared labor in the household market; and (v) *Medbyggerne*, a local coalition model that mobilizes businesses, municipalities, unions, and enforcement authorities to strengthen tax morale and community-level vigilance against labor market crime.

The cooperation-compliance elements of the model, i.e., shared information, clear expectations, and early intervention, mirror international developments whereby tax authorities emphasize transparency and real-time dialogue, as well as justified trust, rather than relying solely on ex-post audits (HMRC, 2024; Business at OECD, 2025). Likewise, the consumer and community strands reflect the OECD's guidance on building tax culture through education, communication, and competence-building (OECD, 2021).

Modern behavioral tax research argues that sustained compliance emerges from the interaction of power (the credible capacity to detect and sanction) and trust (perceived benevolence, fairness, and procedural justice). This dual-pillar 'Slippery Slope' framework predicts that enforced compliance depends on perceived power, whilst voluntary compliance depends on trust; however, the highest overall compliance occurs when both are maximized (Kirchler et al., 2008). The NTA's architecture operationalizes this complementarity. Power is embedded structurally through procurement gatekeeping, prequalification rules, real-time compliance checks, employee-list monitoring, and risk-based site audits, making access to legitimate markets contingent on compliance. Trust is cultivated through transparent information-sharing with partners, clear role divisions of responsibility (the NTA as a knowledge provider; partners as first-line enforcers), and co-production with communities and consumers through initiatives such as *Tettpå* and *Medbyggerne*, which demonstrate that the NTA's aim is not only enforcement, but also support and fairness.

Collaborative Partnerships form the backbone of Mobilization. These are formal contractual partnerships between the NTA, major public procurers, and large private enterprises in high-risk sectors, specifically designed to prevent labor market crime[2] and social dumping[3] before it occurs. These partnerships represent a strategic upstream intervention in which the NTA engages influential market actors to promote lawful and sustainable practices across the entire labor ecosystem. As of 2025, the NTA held 31 formal partnership agreements with key entities, covering about 20,000 companies and more than 200,000 workers, including several of Norway's largest municipalities, counties, public procurers, and major private construction and logistics firms (European Labour Authorities, 2025a). The primary objective of *Collaborative Partnerships* is empowering partners to make informed decisions that support lawful competition, protect the tax base, and reduce the risk of exploitation and harmful working conditions through increased transparency and accountability. A further objective is to strengthen partners' internal capacity to identify and respond to suspicious activity within their own supply chains.

Operationally, *Collaborative Partnerships* function through four key mechanisms: (1) Information-sharing: Ahead of partners' tender competitions, all firms must sign a power of attorney enabling the NTA to share real-time tax- and labor-related compliance indicators (via the standardized tax certificate RF-1507) with the partner. This establishes clear and consistent compliance expectations across the value chain. (2) Procurement integration: Compliance checks are embedded directly into tender prequalification and contract follow-ups, ensuring early detection and immediate consequences, such as refused tenders and/or terminated contracts, for non-compliant firms. (3) Resource allocation and joint training: The NTA assigns dedicated staff to each partnership, supports these partners in interpreting compliance information, provides training on emerging tax crime patterns, and facilitates peer-learning across partners to enhance their collective ability to detect and prevent non-compliance. Partners, in turn, allocate personnel to monitor compliance data and investigate discrepancies. (4) Continuous monitoring: Contractors' tax certificates are reviewed in regular meetings between the NTA and its partners. Firms with meaningful arrears while under contract with a partner risk immediate contract termination, thereby halting their cashflow. The NTA's collection department may also pursue outstanding claims directly through enforcement action against any outstanding invoices the contractor may have with the partner, hence reinforcing immediate action.

Together, these four mechanisms function as an embedded choice architecture. By design, compliant firms face reduced friction, predictable procedures, and fair competition, whereas non-compliant firms encounter increased scrutiny, financial risk, and potential exclusion from lucrative markets.

Tettpå and *Medbyggerne* complement the market-level partnerships by addressing two additional mechanisms essential to sustained compliance: demand-side behavior and community-level norms. *Tettpå* reduces the prevalence of undeclared work by empowering consumers with practical guides, timely information (via Altinn[4] and bank integration), and competence-boosting tools that make lawful choices more salient. *Medbyggerne* builds long-term tax morale by mobilizing local businesses, vocational schools, unions, and municipalities around shared standards of fairness, safe labor conditions and lawful competition, thus creating peer-driven accountability that extends beyond formal enforcement.

Mechanisms Underpinning Mobilization

Mobilization operationalizes contemporary tax compliance insights by distributing enforcement,

2 Labour-market crime refers to systematic violations of labor, tax, and welfare regulations, often involving undeclared work, false invoicing, misuse of subcontracting chains, wage theft, and breaches of health and safety rules. These practices create unfair competition, undermine workers' rights, and erode public revenues.

3 Social dumping refers to practices whereby businesses obtain competitive advantages by undermining labor standards, for example by paying wages below applicable standards, bypassing collective agreements, or exploiting regulatory gaps across borders.

4 Altinn is Norway's national digital platform for dialogue between public authorities, businesses, and citizens. It is used for mandatory reporting (such as tax and returns), applications, and secure digital correspondence with government agencies, including the Norwegian Tax Administration, and enables automated, event-based information flows such as notifications to property buyers.

competence-building, and norm-shaping responsibilities through markets, consumers, and communities. Contrary to relying solely on the tax authority as the enforcement node, the approach cultivates a multi-stakeholder compliance ecosystem in which procurers, private citizens, and local coalitions reinforce each other's capacity to detect, prevent, and respond to labor market and tax crime.

"The approach cultivates a multi-stakeholder compliance ecosystem in which procurers, private citizens, and local coalitions reinforce each other's capacity to detect, prevent, and respond to labor market and tax crime."

The NTA utilizes market actors to generate power through cooperative agreements, procurement controls, real-time tax certificates, and systematic monitoring. These instruments create extensive information trails, which in turn increase the certainty of detection. Simultaneously, the model engages consumers and communities to generate trust through initiatives such as *Tettpå* and *Medbyggerne*, which foster competence, norms, identity, and pride in lawful behavior. Together, these mechanisms ensure that compliance is not simply enforced but is co-produced.

"Compliance is not simply enforced but is co-produced."

The Power-Trust Matrix and Placement of NTA's Instruments

This dual strategy corresponds directly to the 'Slippery Slope Framework', a two-dimensional framework distinguishing between power—the authority's capacity to detect and sanction non-compliance—and trust—taxpayers' perception of fairness, legitimacy, and procedural justice. The framework predicts that the most resilient compliance arises when both dimensions are high (Kirchler 2007; Kirchler et al., 2008; Muehlbacher & Kirchler 2010). The NTA's architecture is designed to move market actors upward and rightward in this matrix: building trust where it is lacking, and strengthening power where information gaps previously enabled non-compliance. Figure 1 illustrates how the instruments used in the whole-of-community approach are positioned within this power-trust matrix.

High Trust, Low Power

Medbyggerne and *Tettpå* operate primarily in this quadrant. These initiatives build identity, tax morale,

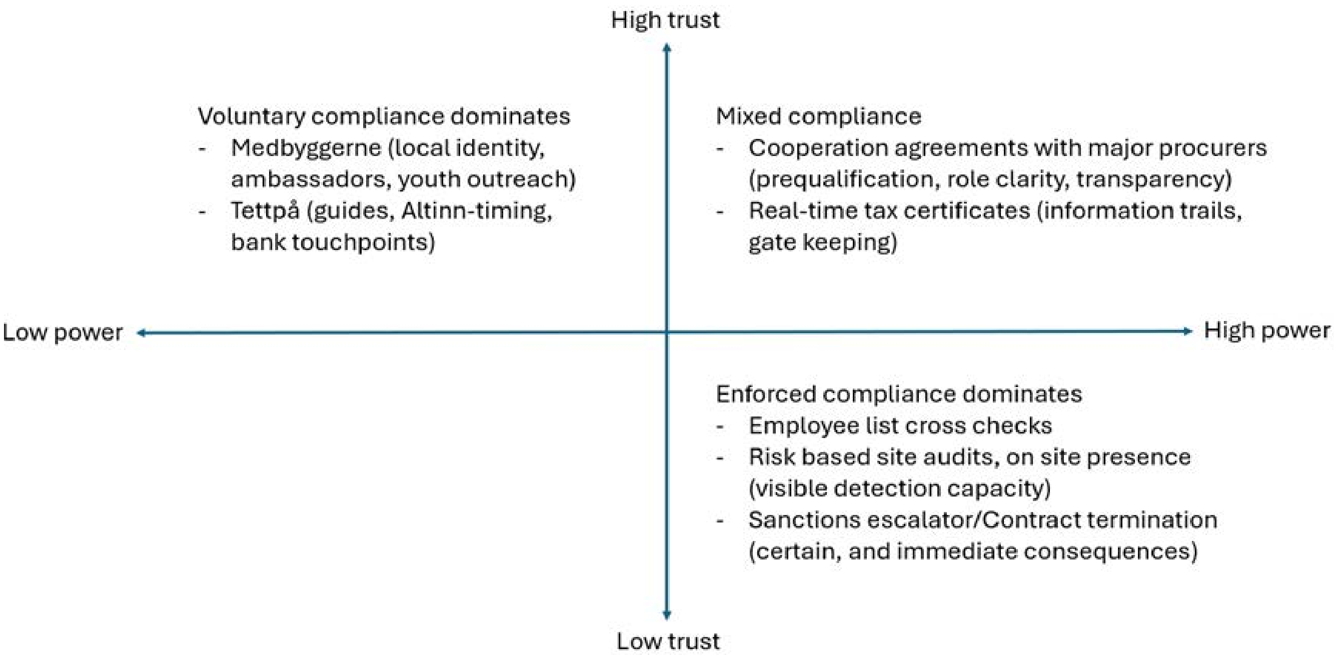


Figure 1: The Power-Trust Matrix

and shared values, as well as supporting competence and responsible decisions. The mechanisms are voluntary, educational, and empowerment-based, and they do not involve strict enforcement, instead relying on social influence.

High Trust, High Power

In this quadrant, cooperation agreements and real-time tax certificates embed transparent collaboration between the NTA and major procurers, creating continuous third-party visibility, establishing clear and predictable compliance gates, and positioning procurers as market gatekeepers. This quadrant represents the core strength of the whole-of-community approach: co-produced compliance with real consequences, where partners not only trust the NTA, but also act on reliable and credible information to exclude non-compliant firms.

Low Trust, High Power

Employee-list cross-checks, risk-based on-site audits, and sanction escalators like contract termination and exclusion from tenders operate primarily in this quadrant. These instruments provide clear and salient detection, credible follow-ups, and immediate consequences, all of which ensure that non-compliance entails predictable risk.

Low Trust, Low Power

This quadrant is associated with systemic evasion. The whole-of-community approach seeks to move actors out of this quadrant by influencing taxpayers' trust through *Medbyggerne* and *Tettpå*. It also increases power through procurement gates, information trails, and increased transparency.

Extending Information Trail Logic Into Business-to-Business Supply Chains

Kleven et al. (2011) demonstrate that third-party reporting is among the most effective mechanisms for constraining evasion. When taxpayers are subject to third-party reporting, evasion is near zero, while it is substantial for purely self-reported income; this result is explained through the existence of readily available information trails. The NTA extends this information trail logic into high-risk supply chains. Instruments like real-time tax certificates, employee list cross-checks, and risk-based on-site audits create continuous visibility in terms of contractor behavior, thereby enabling procurers to detect risk indicators and deny market access to non-compliant firms before harm occurs. Major public and private procurers thus become decentralized enforcement nodes, expanding overall detection capacity.

Tax Morale, Identity, and Community Norms

A large body of literature demonstrates that compliance is not shaped by enforcement alone but also by tax morale, norms, and identity around paying taxes (Alm, 2019). *Medbyggerne* embodies these mechanisms. Through mobilizing the community, it fosters a shared culture of lawful work and collective responsibility. Furthermore, thanks to ambassadors, conferences, school engagement, and peer-based oversight, *Medbyggerne* reinforces the social legitimacy of compliance and the reputational cost of evasion, turning compliance into a shared community value.

Boosts, Nudges, and Consumer Competence-Building

Tettpå targets the consumer market by boosting ordering competence (what documents to ask for, how to check legitimacy) and by partnering with banks at the loan origination stage, closing demand-side loopholes that sustain undeclared work.

Antinyan and Asatryan's (2025) systematic meta-analysis indicates that behavioral interventions, such as deterrence-framed messages and social-norm cues, produce statistically significant but modest improvements in filing and payment compliance. They also find that the design and timing of these messages are critical, as nudges typically yield stronger effects in the short run (within two months after delivery) and are more impactful when delivered through engaging methods such as in-person delivery as compared to physical letters. While Alm et al. (2023) encourage tax administrations to use nudges, they also highlight the importance of 'boosting' taxpayers' knowledge and competence for building more sustainable behavioral changes. In the NTA's mobilization initiative, these theoretical approaches are applied through several initiatives, targeting both businesses and private consumers. For businesses, the collaborative partnerships include joint training and competence development for public and private procurers to help them identify compliant contractors (European Labour Authority, 2025a). Similarly, the *Tettpå* initiative provides private individuals with verified guides, checklists, and information regarding the benefits of choosing compliant contractors, acting as boosts by increasing their 'ordering competence' and empowering them in choosing lawful contractors when renovating their homes (European Labour Authority, 2025c). To reach private individuals in a timely manner when their demand for contractors is believed to be highest, the *Tettpå* initiative utilizes an effective digital workflow, reaching 100% of property buyers in Norway by automatically sending the information package through the Altinn-platform at the time of transfer of title. In addition, a dedicated NTA phone center contacts private individuals applying for building permits in select Norwegian municipalities, making 2,750 calls in 2025 (European Labour Authority, 2025c). This initiative also includes the banking sector, whereby eight Norwegian banks provide this information on behalf of the NTA through their advisors when customers apply for building-related loans.

Modernized Deterrence: Certainty, Severity, and Celerity

Mobilization also relates to deterrence theory from criminology. Classical deterrence theory is built on three pillars: certainty, severity, and celerity (swiftness) (Abramovaite et al., 2023). While tax administrations have traditionally focused on audits and penalties to influence behavior, the NTA's Mobilization initiative strengthens these pillars in a different and more effective way. Certainty is increased through continuous third-party visibility created by procurement controls, real-time compliance data, and shared risk assessments. Severity is delivered not mainly through legal fines but through meaningful economic consequences such as contract termination and loss of market access. Celerity is enhanced because deviations trigger immediate follow-up by partners and may lead to enforcement action in the case of outstanding invoices, loss of contracts, and cashflow. Contract terminations occur during the project cycle, not years after, which is often the case in lengthy administrative processes. Whereas a fine can be viewed as a cost of doing business (Gneezy & Rustichini, 2000; Lin & Yang, 2006), loss of market access is a significant and immediate sanction which can be more impactful than traditional punitive mechanisms. Together, these mechanisms modernize the classical deterrence framework and embed it directly into market structures and community practices, making non-compliance both risky and unprofitable.

"Loss of market access is a significant and immediate sanction which can be more impactful than traditional punitive mechanisms."

The Three Paradigms of Tax Administrations

Alm's (2019) comprehensive review of contemporary tax research outlines three distinct paradigms tax administrations can follow. These are related to enforcement, service, and trust. We believe that Mobilization encompasses all of these paradigms. In relation to enforcement, all contractors must consent to the NTA sharing information about the tax and duties ahead of tender offers, ensuring that contractors have no outstanding arrears when competing for contracts, and while under contract with any NTA collaborative partners. This effectively places Norway's largest procurers as gatekeepers for businesses wanting a share of the market. In terms of service, within the whole-of-community approach the NTA provides partners with education and support during procurement processes. This includes education on interpreting tax certificates, facilitating workshops and information-sharing events on new developments in tax crime modes, and sharing best practices from fellow partners, thus fostering a sense of community effort in stopping tax crime. In addition, the NTA provides all partners with dedicated employees to follow up on specific collaborative partnerships and projects, ensuring swift and timely support when needed. Finally, regarding trust, every stakeholder in the approach must trust the NTA is acting in the best interests of partners, their customers, and the industry in which they operate.

Summary and Concluding Remarks

This article has outlined how the Norwegian Tax Administration's whole-of-community approach, through the Mobilization initiative, translates contemporary tax compliance research into a practical, multi-stakeholder system for preventing labor market and tax crime. Traditional deterrence

Table 1: Instrument-Mechanism-Outcome Mapping

NTA instrument	Key design elements	Behavioral mechanism(s)	Expected behavioral outcomes	Practical metrics / proxies	Anchor literature
Collaborative partnerships with large procurers	Formal MoUs Strengthened internal routines Prequalification rules (NTA = knowledge, partner = gatekeeper)	Power (credible access control) Boosting (educating procurers) Trust (transparent collaboration) Third-party visibility via procurement	Fewer risky bidders Higher prior compliance among winners Earlier detection during tendering Faster collection Strengthened internal routines	Share of tenders using compliance gates Number of bidders disqualified for tax deviations Time-to-remediation Terminated contracts due to non-compliance	Slippery-Slope framework Cooperative compliance notes Boosting
Real-time tax certificates	API/portal issuance Standardized indicators (registrations, filings, arrears) Procurer training Required in prequalification	Information trails / third-party reporting analogue Friction design (easier for compliant, harder for non-compliant) Boosting procurers' knowledge of risk indicators	Upstream exclusion of non-compliers Improved timely filing/payment among suppliers/ contractors	Number of certificates issued Change in arrears among certificate-holders	Third-party reporting evidence
Employee-list cross-checks (monthly)	Partner submits site lists NTA matches to register Deviations flagged	Visibility of payroll/ VAT obligations Power via credible follow-up Norms via predictable oversight	Reduction in unregistered workers Improved reporting completeness	Percentage of unmatched workers Percentage late/no filings after flag	Third-party visibility, cooperative monitoring
Risk-based on-site audits	Visible presence; joint inspections; escalation if partners don't act	Power (salient detection); Salience / deterrence nudges	Drop in on-site violations; faster corrective actions; deterrence spillover to nearby sites	Violations per audit; repeat-offence rate; partner response lag	Trust-power empirics; deterrence reviews
Sanctions escalator via procurement	Contract termination; disqualification from future tenders	Power (immediate, certain consequences), market gatekeeping	Lower recidivism; shift of market share to compliant firms	Number of terminations, recidivism within 12 months; award shared to compliant firms	Cooperative compliance playbooks; HMRC framework
***Tettpå* (consumer program)**	Step-by-step guides; Altinn letters at property purchase; helpline; media; bank integration at loan stage	Boosts (competence); Timely nudges; Norms/tax morale (make 'compliant choice' the default)	More consumers request documents; higher selection of legitimate trades; reduced demand for undeclared work	Percentage of consumers using checklist; document-request rate; complaint/ fraud incidence	OECD tax-culture (education); Boosting; nudges, meta analysis, tax morale reviews
***Medbyggerne* (local coalition)**	Community ambassadors; youth outreach; joint authority-business actions	Tax morale; Identity and norms; Trust via proximity and co-production	Higher voluntary compliance; peer enforcement; reputational costs for non-compliers	Ambassador count; school/union reach; survey-based trust/ norm indices, local violations trend	Tax moral literature; OECD education; Boosting

models, while foundational, cannot fully explain modern compliance patterns, and behavioral research shows that trust, legitimacy, information trails, and social norms are equally important drivers of sustained compliance. The Mobilization initiative integrates these insights by combining high power, through procurement gates, real-time compliance data, and continuous monitoring, with high trust, fostered through education, transparency, consumer empowerment, and community level collaboration.

The mechanisms described in this article demonstrate how enforcement, service, and trust can be aligned within a unified architecture. Cooperation agreements with procurers embed detection capacity directly in market transactions, making non-compliance both risky and unprofitable. *Tettpå* strengthens household decision-making and reduces demand for undeclared work by providing timely, verified information at critical behavioral moments. *Medbyggerne* mobilizes local actors to reinforce tax morale, identity, and shared responsibility, anchoring compliance in community norms. Together, these initiatives extend the logic of third-party reporting into high-risk supply chains, modernize deterrence through certainty, severity, and celerity, and create an ecosystem in which compliant behavior becomes the social and economic default.

> "Cooperation agreements with procurers embed detection capacity directly in market transactions, making non-compliance both risky and unprofitable."

The recognition of these initiatives as ELA Good Practices (European Labor Authority, 2025a, 2025b, 2025c) underscores their preventive effects and their potential relevance beyond Norway. While the model is tailored to Norway's institutional context, its underlying principles, namely, distributed enforcement capacity, competence building, real-time information flows, and community mobilization, are transferable to other jurisdictions facing similar challenges in high-risk sectors. As labor market crime becomes increasingly complex, such integrative and preventive approaches offer a promising pathway for safeguarding fair competition, protecting workers, and strengthening the integrity of tax systems.

Ultimately, the contemporary tax literature suggests that sustainable compliance cannot be achieved by audits and sanctions alone. It arises when institutions build trust, design environments that make lawful choices easier, and engage the market and society as partners in protecting the rule of law. The Norwegian Tax Administration's Mobilization initiative demonstrates how tax administrations can evolve from being sole enforcers to becoming coordinators of a broader ecosystem—one in which compliance is co-produced, continuously supported, and socially reinforced.

THE AUTHOR

Johan Birkelund is a senior behavioral economist in the Tax Crime Department of the Norwegian Tax Administration, where he works on preventive approaches to tax crime and is a member of the agency's Behavioral Insights Team. He holds a Ph.D. in Economics from UiT The Arctic University of Norway. Birkelund previously served as an associate professor at Inland Norway University of Applied Sciences and as an adjunct associate professor at UiT The Arctic University of Norway. He is also affiliated with the research group on Preventive Policing at the Norwegian Police University College.

REFERENCES

Abramovaite, J., Bandyopadhyay, S., Bhattacharya, S., & Cowen, N. (2023). Classical deterrence theory revisited: An empirical analysis of Police Force Areas in England and Wales. *European Journal of Criminology*, 20(5), 1663-1680. https://doi.org/10.1177/14773708211072415

Allingham, M. G., & Sandmo, A. (1972). Income tax evasion: A theoretical analysis. *Journal of Public Economics, 1*, 323-338. https://doi.org/10.1016/0047-2727(72)90010-2

Alm, J. (2019). What motivates tax compliance? *Journal of Economic Surveys*, *33*(2), 353-388. https://doi.org/10.1111/joes.12272

Alm, J., Burgstaller, L., Domi, A., März, A., & Kasper, M. (2023). Nudges, boosts, and sludge: Using new behavioral approaches to improve tax compliance. *Economies*, *11*(9), 223. https://doi.org/10.3390/economies11090223

Antinyan, A., & Asatryan, Z. (2025). Nudging for tax compliance: A meta-analysis. *The Economic Journal*, *135*(668), 1033-1068. https://doi.org/10.1093/ej/ueae088

Business at OECD (BIAC). (2025). OECD International Compliance Assurance Programme (ICAP) and Tax Certainty Mechanisms (Briefing).

European Labour Authority. (2025a). *Collaborative partnerships to reduce labour-market crime and foster a culture of compliance: Norway*. https://www.ela.europa.eu/sites/default/files/2026-01/NO_GP2025_Collaborative_partnerships_to_reduce_labour-crime_and_foster_compliance.pdf

European Labour Authority. (2025b). *Medbyggerne: A grassroots coalition against undeclared work and labour-market crime in Norway*. https://www.ela.europa.eu/sites/default/files/2026-01/NO_GP2025_Medbyggerne_a%20grassroot_coalition_against_undeclared_work.pdf

European Labour Authority. (2025c). *Tettpå: Consumer-focused measures to prevent undecleared work in the construction sector*. https://www.ela.europa.eu/sites/default/files/2026-01/NO_GP2025_Tettp%C3%A5_mobilizing_the_consumer_power_in_the_construction_industry.pdf

Gneezy, U., & Rustichini, A. (2000). A fine is a price. *The Journal of Legal Studies*, 29(1), 1-17. https://doi.org/10.1086/468061

HM Revenue & Customs (HMRC). (2024). HMRC's Framework for Co-operative Compliance. *GOV.UK*. https://www.gov.uk/guidance/hmrcs-framework-for-co-operative-compliance

Kirchler, E. (2007). *The economic psychology of tax behaviour*. Cambridge University Press.

Kirchler, E., Hoelzl, E., & Wahl, I. (2008). Enforced versus voluntary tax compliance: The slippery slope framework. *Journal of Economic Psychology*, 29(2), 210-225. https://doi.org/10.1016/j.joep.2007.05.004

Kleven, H. J., Knudsen, M. B., Kreiner, C. T., Pedersen, S., & Saez, E. (2011). Unwilling or unable to cheat? Evidence from a tax audit experiment in Denmark. *Econometrica, 79*(3), 651-692. https://doi.org/10.3982/ECTA9113

Lin, C. C., & Yang, C. C. (2006). Fine enough or don't fine at all. *Journal of Economic Behavior & Organization*, 59(2), 195-213. https://doi.org/10.1016/j.jebo.2004.04.006

Luttmer, E. F. P., & Singhal, M. (2014). Tax morale. *Journal of Economic Perspectives, 28*(4), 149-168. https://doi.org/10.1257/jep.28.4.149

Muehlbacher, S., & Kirchler, E. (2010). Tax compliance by trust and power of authorities. *International Economic Journal*, 24(4), 607-610. https://doi.org/10.1080/10168737.2010.526005

OECD. (2021). *Building tax culture, compliance and citizenship: A global source book on taxpayer education (Second Edition).* OECD Publishing.

Sandmo, A. (2005). The theory of tax evasion: A retrospective view. *National Tax Journal, 58*(4), 643-663. https://doi.org/10.17310/ntj.2005.4.02

Slemrod, J. (2019). Tax compliance and enforcement. *Journal of Economic Literature*, 57(4), 904-954. https://doi.org/10.1257/jel.20181437

Skatteetaten (Norwegian Tax Administration) (n.d.). *Certificate for tax and duties (RF-1316).*

Traxler, C. (2018). Deterrence of economic crimes: Tax evasion. In G. Bruinsma & D. Weisburd (Eds.), *Encyclopedia of criminology and criminal justice.* Springer. https://doi.org/10.1007/978-1-4614-5690-2_411

Torgler, B. (2007). *Tax compliance and tax morale: A theoretical and empirical analysis*. Edward Elgar.

WU Global Tax Policy Center (2021). *Cooperative compliance: A multi-stakeholder and sustainable approach to taxation (Policy Brief).*

RESOURCES

Postgraduate Programs

LSE

Department of
Psychological and Behavioural Science

EXECUTIVE MSc BEHAVIOURAL SCIENCE

UNDERSTANDING BEHAVIOUR IN A CHANGING WORLD

Designed for working professionals across diverse sectors, LSE's Executive MSc Behavioural Science combines rigorous academic training with practical insight.

Taught by world-leading experts and practitioners, you will learn how to apply behavioural insights to solve policy, business, organisational and societal challenges.

Whether you work in the public, private or third sector, you'll build specialist knowledge and discover practical applications that can open new professional opportunities in this fast-growing field.

CONTACT US

pbs.emsc@lse.ac.uk

For more information, please visit **lse.ac.uk/pbs/emsc**

LSE Department of Psychological and Behavioural Science

EXECUTIVE MSc BEHAVIOURAL SCIENCE

A PROGRAMME DESIGNED AROUND YOUR CAREER

The Executive MSc is designed to fit around a busy career. It offers full-time professionals from anywhere in the world the opportunity to study flexibly while continuing to work.

The 16-month programme begins each September and is delivered through three intensive two-week teaching sessions at LSE's London campus, with independent study and assignments between sessions. Outside the teaching blocks, you can continue to live and work in the UK or overseas. After the taught component, you will complete a dissertation on a topic of your choice, with support from an academic supervisor.

The curriculum covers a broad range of courses, including:

- Behavioural Science and Policy
- Behavioural Decision Science
- Research Methods for Behavioural Science
- Frontiers in Behavioural Science Methods
- Policy Appraisal and Ethics
- Behavioural Science in an Age of AI and New Technology
- Corporate Behaviour and Decision Making
- The Science of Time at Work
- Beyond Nudging: The Frontiers of Behavioural Science in a Complex World.

Please note that, while this information is correct at the time of publication, the School may on occasion need to change, suspend or withdraw a course.

CONTACT US

pbs.emsc@lse.ac.uk

For more information, please visit **lse.ac.uk/pbs/emsc**

LSE Department of Psychological and Behavioural Science

A GLOBAL COMMUNITY OF PROFESSIONALS

Our students come from a wide range of professional, academic and geographical backgrounds, creating a diverse international cohort. They bring perspectives from different sectors, disciplines and career stages, enriching discussion throughout the programme. United by a shared interest in behavioural science and its real-world application, they also become part of a global network of peers and alumni that extends well beyond the classroom.

"It was a genuinely rewarding experience, with a fantastic global cohort, that continues to bring tangible value to my daily work."

Sarah, 2025 graduate

"The programme has given me the confidence to challenge the ways things have typically been done to reach more meaningful, impactful outcomes."

Lucy, 2024 graduate

"The Executive MSc in Behavioural Science couples the dynamism of the LSE learning environment with an engaging and multifaceted curriculum, all under the tutelage of world-class academics."

Mike, 2025 graduate

"The Executive MSc in Behavioural Science has been a game-changer for me. By deepening my understanding of human behaviour, I can now design and implement impactful, evidence-based initiatives that drive transformational change."

Fadzi, 2024 graduate

"I've gained a deeper understanding of how to design targeted, evidence-based interventions that produce tangible, lasting behavioural shifts."

Kele, 2025 graduate

CONTACT US

pbs.emsc@lse.ac.uk

For more information, please visit **lse.ac.uk/pbs/emsc**

University College Dublin
Ireland's Global University

UCD School of Economics
Scoil na hEacnamaíochta UCD

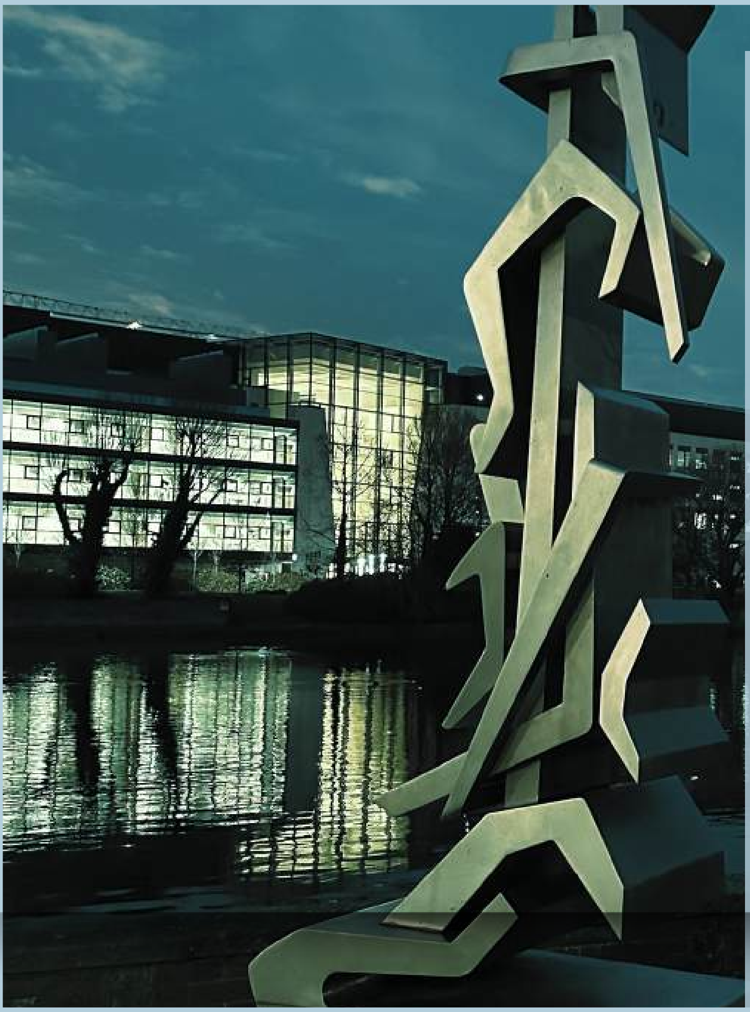
MSc in Behavioural Economics at University College Dublin

UCD School of Economics is **Ireland's leading Economics department**, and our MSc in Behavioural Economics is taught by a vibrant team of behavioural researchers and practitioners.

We provide in-depth training in the core concepts and theories of behavioural economics, with a focus on how to apply these tools in areas such as health, regulation, labour, environmental economics, and education. You can choose from a range of classes on empirical methodologies for conducting different types of controlled experiments (field, laboratory, and online).

You have the opportunity to design and conduct your own experiment in our **UCD Experimental Social Sciences Lab.** You also can also choose to write a thesis or do a summer internship in an organisation of your choice.

Our MSc students are an integral part of the strong research culture in UCD and attend seminars by international researchers and workshops organised by the UCD Behavioural Science and Policy Group.



PROGRAM HIGHLIGHTS

Rigorous training in core theories and methods for fundamental research, as well as evidence-based policy.

Strong policy focus and opportunities to interact with policymakers.

Small class sizes and dedicated one-to-one research supervision.

Ireland's strategic location as a European hub for many large multinationals opens doors to global opportunities. Prominent employers include large tech, pharmaceutical and financial services companies.

Diverse career paths in public policy, the private sector and a PhD route into academia.

GRADUATE PROFILE: EMILY WHARTON-HOOD

It was an absolute honour and privilege to obtain my Masters in Behavioural Economics at UCD.

The MSc provides such a variety of modules while being taught by some of the best in the industry. Advancing my skills in Econometrics and Microeconomics was a highlight of the MSc, but also taking modules such as Energy Economics really broadened my perspective on where Economics is most impactful. Being able to run my own experiment allowed me to conduct my thesis in an area I was passionate about. Throughout the program, you become equipped with skills you would use in any industry you decide to move into. I am currently working as an Analyst at 2K Games Dublin!

I could not recommend this MSc enough.

Our graduates have found jobs in places such as: Behavioural Insights Team, Kellanova, Irish Government Economic and Evaluation Service, Grant Thornton, McKinsey, KPMG, Accenture, Novartis, Deloitte, Central Statistics Office, Central Bank of Ireland, European Central Bank, Economic and Social Research Institute, Competition and Consumer Protection Commission, AIB Bank, European Commission and other leading national and international organizations.

Learn more at :
https://www.ucd.ie/economics/

University of Glasgow

MSc Behavioural Science

Master Behaviour, Shape Business, Impact Society

This unique programme is a collaboration between the Adam Smith Business School and the School of Psychology & Neuroscience. It combines knowledge from Economics, Business, Finance, Psychology, and Neuroscience to develop future behavioural scientists. You will have access to cutting-edge methods, modern lab facilities, and strong industry partnerships.

Key Facts:

- Duration: 12 months full-time
- Delivery: In-person
- Teaching starts: September
- Academic contact: business-recruitment@glasgow.ac.uk

University of Glasgow

Why choose MSc Behavioural Science at the Adam Smith Business School?

- **Multi-Disciplinary Approach:**
  A unique decision-making programme combining insights from Economics, Management, Finance, Psychology, and Neuroscience.
- **Science and Business Tracks:**
  A unified foundation for both tracks guarantees consistency while providing specialised skills tailored to your selected path.
- **Theoretical & Practical Learning:**
  Learn theory alongside quantitative, empirical, and market-driven research methods.
- **State-of-the-Art Lab Access:**
  Use our new behavioural laboratory for real-world experiments, including dissertation research.
- **Industry Engagement:**
  Take part in competitive, merit-based industry projects with real companies. Solve real business challenges, gain experience, and grow your professional network.

After graduation: What jobs can you get with this degree?

The program aims to cultivate versatile behavioral scientists equipped with a range of skills, providing outstanding career opportunities across multiple sectors. You will have the option to select between two career tracks:

Science Track

Equips graduates with neuro-scientific skills and interdisciplinary expertise in Behavioural Science, ideal for pursuing a PhD or research roles in various sectors.

Business Track

Prepares graduates for careers in policy or business, focusing on decision science for practical applications in diverse industries and market systems.

Find out more and join us today:
glasgow.ac.uk/postgraduate/taught/behavioural-science

Learn the theory, apply the tools, and make a difference

Penn's Master of Behavioral and Decision Sciences (MBDS) program equips students with theoretical and practical tools to understand how individuals and groups make decisions, how to affect those decisions, and how social norms play a role in motivating and changing social behaviors. Led by world-renowned faculty, researchers, and practitioners, the MBDS program creates unique opportunities for students to engage with an exceptional advisory board, apply tools and knowledge in our annual Design Challenge, and pursue independent, cross-disciplinary research throughout Penn.

Meet our alumni

"I loved learning tools and techniques to evaluate what people were saying and translate that into opportunities for the client. The [capstone] was a really exciting project. I never worked in the digital mental health space before. Having those opportunities to gain insight into different industries has helped me become a chameleon and learn to speak the languages of different clients."

Kathryn Ambroze, MBDS '22
VP, UX Research Lead - Trust and Security, JPMorgan Chase & Co.

"When I was accepted into the MBDS, it was a meant-to-be moment. I felt like there was a link between me and the program. I could take the time to explore behavioral science in an academic setting. Some people love commercial spaces, some want to go into consulting, some people are really into research. I realized health and health outcomes are definitely what I'm interested in, personally and in my career."

Yuzhen (Valerie) Guo, MBDS '22
Behavioral Designer, Lirio, LLC

"One of the amazing things about Penn is that the faculty you work with are heavily involved in research—they're very much at the forefront of their field, so you can take part in a lot of research if you want to. Once I was done with the first project, there were other professors who needed help with different projects."

Max Spohn, MBDS '20
PhD candidate, Harvard Kennedy School of Government

Learn more about our engaged and well-connected alumni at

www.upenn.edu/mbds

Meet the Master of Behavioral and Decision Sciences program's founding director

Cristina Bicchieri

S. J. P. Harvie Professor of Social Thought and Comparative Ethics; Professor of Legal Studies and Business Ethics, Wharton Director, Center for Social Norms and Behavioral Dynamics; Faculty Director, Master of Behavioral and Decision Sciences and Philosophy, Politics, and Economics; University of Pennsylvania

> *"Wherever there is a human group there are social norms."*
>
> *-Cristina Bicchieri*

Cristina Bicchieri is a world authority on social norms and has consulted with UNICEF, the World Bank, the Gates Foundation, the United Kingdom's Department for International Development, and many other organizations. She is the founder of the Master of Behavioral and Decision Sciences program and the Center for Social Norms and Behavioral Dynamics, a major research center at Penn that aims to support positive behaviors on a global scale. Cristina is the author of over 100 articles and seven books.

Unparalleled connections, exceptional opportunities

A defining feature of the University of Pennsylvania's Master of Behavioral and Decision Sciences program (MBDS) is its network of outstanding industry and research partners who help bring students exceptional practical experiences.

MBDS Design Challenge Industry Affiliates

accenture · AIR PRODUCTS · BEESY · be kind network · BuenTrip Ventures · ComplexChaos · [cowry consulting]
dx diagnostics your way · Deloitte. · EY Building a better working world · Flourish · free people · HURON · IDB
ideas42 · J.P.Morgan · Ipsos · MakeItLab · MIDA · newristics · OECD
Oxford Energy · Pensión · Precision Healthcare Ecosystem · PROJECT INSPIRE · subconscious.ai · Tech4Pets · tenova ADVANCED TECHNOLOGIES
THE ENERGY EFFICIENCY INSTITUTE · THE WORLD BANK · Vanguard

Learn more about our world-renowned faculty and researchers at:

www.upenn.edu/mbds

Penn
Master of Behavioral and Decision Sciences

The MBDS Capstone—Design Challenge

The Capstone—Design Challenge experience consists of two consecutive semesters where students are invited to engage in a dialogue with industry to understand how to use the tools learned in the Master of Behavioral and Decision Sciences program (MBDS) to tackle specific behavioral problems. In the fall semester, students focus on learning behavioral design techniques and theories while honing their client-facing skills. They engage in a conversation with academic and industry leaders who provide them with first-hand knowledge of a) how and why behavioral science is applied in organizations across several domains, and b) what MBDS graduates can expect when they enter the job market. In the second semester, students take their knowledge to real clients in the Design Challenge, a 14-week course where student teams solve behavioral questions brought by MBDS Industry Affiliates. A major goal is to prepare students to engage with diverse industry and global organizations as soon as they complete the program.

The Design Challenge is an invaluable opportunity for our students to apply their MBDS education toward developing practical solutions while gaining real-world experience.

Every spring, as part of the required Capstone Experience course, the MBDS program organizes the Design Challenge, where our students partner with MBDS Industry Affiliates to apply cutting-edge knowledge from the fields of behavioral economics, decision sciences, data science, network analysis, and public policy to solve real-life problems. We welcome world-leading clients in industries like finance, healthcare, consulting, sustainability, technology integration in marketing, and finance to collaborate with our students and provide guidance on solving their challenges.

In the Design Challenge, MBDS students work to translate academic research and applied frameworks into actionable insights toward a client-focused problem. At the end of the Design Challenge, students present their proposed solutions to the client's senior management and leadership.

Learn more about how MBDS connects students and industry at:

www.upenn.edu/mbds

Penn
Master of Behavioral and Decision Sciences

Where science meets real-world impact

MSc *Behavioural Science*

Est. 2013	100%	MSc · PGDip · PGCert	FT / PT
ESTABLISHED. REFINED. TRUSTED.	STUDENT SATISFACTION · PTES 2025	AWARD OPTIONS	FULL-TIME OR PART-TIME

An **ESRC 1+3 accredited programme** offered since 2013 and continuously refined over more than a decade. Taught by active researchers publishing in world-leading journals in management, economics, medicine and psychology, this programme equips you with the conceptual, methodological and applied skills to understand and change behaviour effectively and ethically in business, government or civil society.

PROGRAMME HIGHLIGHTS

Behavioural Economics: Concepts & Theories
Landmark studies contextualised to build a toolkit for navigating real-world behavioural challenges.

Applying Behavioural Insights to Business & Policy
Nudges, boosts, sludge, and the FORGOOD ethics framework — developing responsible interventions and engaging stakeholders.

Evidence-Based Decision Making
Experimental design, causal inference and critical appraisal — including designing and running a class experiment.

Statistics and Data Communication with R
From fundamentals to independent research projects using the open-source standard of the field.

Survey Measurement & Analysis
Survey design, data quality, and drawing meaningful inferences from large datasets.

Decision Making in Organisations
A behavioural lens on recruitment, motivation, wellbeing, and change management.

Research themes: Nudges | Sludge | Consumer behaviour | Ethics of behavioural influence | Subjective well-being | Happiness
Forecasting | Experimental Design | Risk and uncertainty | Stakeholder Engagement | Non-market valuation | Day Reconstruction

> “*Graduates of Stirling's MSc Behavioural Science come to us* **extremely well prepared.** *Their technical know-how and breadth of knowledge around how to apply behavioural science in practice means they are regularly at the* **top of our list of candidates** *when we're hiring.*
>
> — Jez Groom, Founder of Cowry Consulting and Small Wonder Consultancy

A programme connected to the world

CONNECTED TO THE WORLD

Our students benefit from connections with **HMRC, Ofcom, Ofgem, Cowry Consulting, CMA, Siemens, NHS, Ireland's Department of Health, Glasgow Film Theatre, Keep Scotland Beautiful, eLife journal, Iceland's Cancer Detection Clinic,** and many more.

APPLY YOUR LEARNING TO

- Financial decision making
- Mental health & well-being
- Environmental behaviour
- Gambling
- Media use & digital behaviour
- Food consumption
- Transport & mobility
- Health & public health
- Workplace behaviour
- Consumer behaviour
- and many more...

GRADUATE ROLES

Behavioural Insight Teams
national & local government

Regulatory bodies
Ofcom, Ofgem, CMA, UN Food & HMRC

Consultancies
behavioural architect, insights professional

Large corporations
lead behavioural scientist, research manager

Academia
PhD pathways — SSRC 1+3 recognised MSc

Charities & NGOs
policy analyst, behaviour change specialist

BE THE DIFFERENCE

CAREER OUTCOMES

Avg. UK salary **£42,877** for behavioural scientists (Glassdoor, 2025)

MONEYBALL YOUR CAREER

Taught by researchers publishing in *Nature, NEJM, Management Science and Behavioural Public Policy* — cutting-edge knowledge directly informing your learning

AACSB accredited — a distinction held by only 6% of business schools worldwide

100% student satisfaction in the 2025 PTES survey

Graduates in senior roles at HMRC, Ofcom, the Behavioural Insights Team, Cowry Consulting, national governments and leading consultancies worldwide

An active network connecting you with HMRC, Ofcom, Ofgem, Siemens, NHS, Cowry Consulting, the CMA and many more — from day one

Transferable principles you can apply to the area that matters most — financial decision making, mental health, environmental behaviour, gambling, media use, food, transport, or beyond

SCHOLARSHIPS AVAILABLE

Postgraduate Merit Scholarship · Be the Difference International Scholarship · EU Postgraduate Scholarship
Stirling Alumni Scholarship · Postgraduate India Scholarship · Postgraduate Pakistan Scholarship

All of this comes with tuition fees that are **highly competitive with equivalent MSc programmes across the UK** *— making Stirling the wisest proposition in behavioural science education.*

"The guidance offered in preparing our dissertation topics is exceptional. The faculty provide great support. It has helped me think through how my project can have real world impact."

— Dannish Odongo
Graduate 2026

"I love how the program brought me into the wider network of behavioural scientists linked to the group at Stirling. A graduate now working at OFCOM gave me guidance on my dissertation. Even before graduating I was hired on a funded project desludging scientific publishing."

— Paola Goggi
Graduate 2025

"The course was hugely stimulating — it pushed me to learn new skills and think in a deep and structured way. Being taught by world-leading experts genuinely broadened my perspective."

— Morgan Reed
Graduate 2025

FIND OUT MORE & APPLY

behsci.stir.ac.uk

msc-behaviouralscience@stir.ac.uk

Open to non-degree holders with relevant experience

100% Student Satisfaction
PTES 2025

University of Stirling
Stirling · Scotland · FK9 4LA

Scan to apply

MSc in Behavioural and Economic Science

UNIVERSITY OF WARWICK

Do you want to understand the choices people make, why they make them and what influences their behaviour?

The MSc in Behavioural Science and Economics at the University of Warwick combines multidisciplinary expertise from the departments of Psychology, Economics and Warwick Business School, that is crucial to understand how influencing people's choices impacts a variety of sectors and industries.

Further details:
psychologyPG@warwick.ac.uk | +44 (0)24765 23096

warwick.ac.uk/bes

Why study Behavioural and Economic Science at the University of Warwick?

- Unique and innovative programme with modules drawn from Psychology, Economics, and the Business School, providing a thorough grounding of both the theory and real- world application of behavioural science.
- Rigorous core modules that lay the foundation for the field, including behavioural microeconomics, core issues in psychological science, and advanced research methods and statistics.
- Wide range of optional modules, covering topics which can include behavioural public policy, nudging and behaviour change, experimental economics, behavioural ethics, Bayesian stats, cognitive modelling, and neuroeconomics.
- Taught by leading behavioural science experts in all three disciplines.
- An alumni network that draws from over 15 years of MSc cohorts, in one of the most established behavioural science courses in the world.
- Opportunity to create a substantive piece of original research in collaboration with a supervisor through participation in the summer project.

Key Facts

Course code **P-C8P7**	Start Date **28 September 2026**	Duration **1 year full-time**
Qualification **MSc**	Led by **Psychology**	Location **University of Warwick**

Download our prospectus
your.warwick.ac.uk/form/PGstudyguide

Apply now
warwick.ac.uk/study/postgraduate/apply/#taught

Further details:
psychologyPG@warwick.ac.uk | **+44 (0)24765 23096**

warwick.ac.uk/bes

Why Warwick?

Biswajit Chaki
MSc Behavioural and Economic Science (Science Track), 2019

The approach that we took in the Behavioural and Economic Science classes -- relying on data to put forward theories, using scientific method to test out hypotheses and generally the high-standard of research & analysis -- has been something that I've taken with me through to my job today. We do lots of mini-experiments and test to work out exactly what works best, why it works and then how best to launch it to a wider group.

Donal Connelly
MSc Behavioural and Economic Science (Science Track), 2023

My time at Warwick was instrumental in preparing me for life after university. The rigorous academic environment challenged me to think critically and creatively, equipping me with essential analytical skills that I apply daily. Warwick's emphasis on practical experience was particularly beneficial. Through workshops, seminars, and networking events, I gained hands-on experience in entrepreneurship and business development.

The University of Warwick is consistently ranked highly in both UK and International league tables.

7th
in the UK university league tables *(The Guardian University Guide 2026)*

5th
in the UK for Economics *(The QS World University Rankings by Subject 2026)*

4th
most targeted university by the UK's top 100 graduate employers *(The Graduate Market in 2025, High Fliers Research Ltd).*

By studying at Warwick, you will be part of a global community of students and alumni from diverse backgrounds from all over the world. Our graduates have gone on to pursue a wide range of successful careers across academia, government, healthcare, consultancy and the public sector, applying behavioural science to shape policy, improve public services and influence decision-making across industries.

Further details:
psychologyPG@warwick.ac.uk | +44 (0)24765 23096

warwick.ac.uk/bes

Other Resources

For the most up-to-date behavioral science resources, please visit

BehavioralEconomics.com

Postgraduate Programs (complete listing)
LinkedIn Group
Job Board
Encyclopedia
Books
and more

APPENDIX

Author Profiles

Ulrike Malmendier (Introduction)

Ulrike Malmendier is the Cora Jane Flood Professor of Finance at the Haas School of Business and Professor of Economics at the Department of Economics, both at the University of California, Berkeley, where she has taught since 2006. In addition to her roles as research associate at NBER, research affiliate at CEPR, faculty research fellow at IZA (a CESifo Affiliate), Guggenheim Fellow and Member of the Academy of Arts and Sciences; she served as a member of the German Economic Advisory Council from 2022 to 2026, advising the German government on economy policy. Malmendier holds a PhD in Business Economics from Harvard University (2002) as well as a PhD in Law (*summa cum laude*) from the University of Bonn (2000).

Isabelle Brocas (Guest Editorial)

Isabelle Brocas is Professor of Economics at the University of Southern California, where she directs the Los Angeles Behavioral Economics Laboratory (LABEL) and the Theoretical Research in Neuroeconomic Decision-making (TREND) Institute. She is also affiliated with the Institute for Advanced Study in Toulouse (IAST) and the Centre for Economic Policy Research (CEPR). Her research develops theoretical and experimental approaches to decision-making, with particular emphasis on the biological, cognitive, and developmental foundations of economic behavior. Her work studies how perception, memory, attention, self-control, social interaction, and internal constraints shape choices across individuals and contexts. She has published widely in behavioral economics, neuroeconomics, and experimental economics.

Alain Samson (Editor)

Alain Samson is the editor of the Behavioral Economics Guide and founder of BehavioralEconomics.com. He has worked as a consultant, researcher and scientific advisor. His experience spans multiple sectors, including finance, consumer goods, media, higher education, energy and government. He was previously Chief Science Officer at Behave Technologies (formerly Syntoniq).
Alain studied at UC Berkeley, the University of Michigan and the London School of Economics, where he obtained a PhD in Social Psychology. His scholarly interests have been eclectic, including culture and cognition, social perception, consumer psychology and behavioral economics. He has published articles in scholarly journals in the fields of management, consumer behavior and economic psychology. He is the author of Consumed, a Psychology Today online popular science column about behavioral science.
Alain is a Founding Member of the Global Association of Applied Behavioural Scientists (GAABS).

alain@behavioraleconomics.com

Contributing Organizations

Behavioral Science Group

The Behavioral Science Group (BSG) is the UAE Government's central behavioral science unit, based within the Office of Development Affairs (ODA) of the Presidential Court.

Dedicated to positive societal change, BSG partners with government and institutions to design, test and scale solutions that improve policies and services across sectors including health, education, sustainability, and cultural values.

BSG is embedded in policy-making for one of the world's most diverse societies, driven by a culture of experimentation and collaboration. Combining real-world insights with local and global expertise, BSG is helping to advance national ambitions while contributing to lasting impact and evolving global practice.

www.bsg.ae

BeWay

BeWay is the strategic organizational change consultancy that leads the application of behavioral science in the private sector. With a multidisciplinary team of 200+ experts—including psychologists, economists, data scientists, sociologists, anthropologists, philosophers, and neuroscientists—our work bridges rigorous behavioral research with practical application, making science accessible, ethical, and actionable for businesses across diverse sectors. Around the world, BeWay helps organizations understand and empathize with their audiences, enabling them to achieve strategic goals while creating a positive and measurable impact on society.

www.beway.org

The Chicago School

The Chicago School prepares scholar-practitioners to apply behavioral science to complex social, clinical, and organizational challenges. Grounded in a strong foundation of experimental analysis and applied behavior analysis, the program emphasizes both conceptual rigor and real-world impact. Students receive advanced training in research methodology, data analysis, and evidence-based intervention, while also developing competencies in teaching, supervision, and leadership. The curriculum integrates core principles of behavior analysis with contemporary issues, encouraging students to extend the science beyond traditional settings. Through mentorship, practicum experiences, and dissertation research, students are trained to evaluate and address socially significant behavior across diverse populations and systems. Graduates are equipped for roles in academia, clinical practice, consultation, and organizational leadership. The program reflects The Chicago School's broader mission of integrating psychology and behavioral science with a commitment to innovation, diversity, and community impact, preparing graduates to contribute meaningfully to both the science and its application.

www.thechicagoschool.edu

Dectech

Dectech strives to provide the most accurate and best value forecasts available on how people will behave in new situations. Founded in 2002, we've conducted more than 450 studies involving over 3.2 million participants. We hold that people make very different decisions depending on their context and struggle to self-report their beliefs and motives. So, we developed Behaviourlab, a randomised controlled trial approach that immerses participants in a replica of the real-world decision environment. Over the years we've shown how Behaviourlab can provide higher accuracy forecasts and more actionable insights.

www.dectech.co.uk

Lirio

Lirio is the first-ever behavioral science and agentic AI platform that empowers people to live their healthiest lives through hyper-personalized and empathetic consumer experiences that meaningfully improve outcomes. The company's proprietary Precision Nudging® Interventions are designed to overcome specific barriers and move people to take action for better health. Recognized with the 2025 Inc. Best in Business award, including honors for Best in Innovation, Lirio has received multiple awards for its excellence in applied artificial intelligence and is HITRUST® CSF certified, SOC 2 Type II compliant, and NIST certified.

www.lirio.com

Neovantas

Neovantas is a top international management consultancy focused on accelerating change through the power of behavioral data science. We focus on transforming behavior to assure business results in a sustainable way over time. Our consulting team is specialized by sector (retail banking, insurance, telecoms, airlines, utilities, etc.) and functions (advanced analytics, behavioral science, business transformation and innovation). We build strong, lasting relationships with our clients through our commitment, integrity, professional excellence and creativity. Our international presence has been expanded with projects both in Europe and in Latin America.

www.neovantas.com

Norwegian Tax Administration (Skatteetaten)

The Norwegian Tax Administration (Skatteetaten) is the Norwegian government agency responsible for assessing and collecting taxes, administering VAT, and ensuring that individuals and businesses meet their tax obligations. It also manages the National Registry, which records key information about residents in Norway. The agency processes tax returns, issues tax assessments, and works to ensure compliance across both personal and corporate taxpayers. As a central government body, its work underpins the funding of Norway's public services and welfare system. Skatteetaten has in recent years invested significantly in digitalisation, making Norway one of the more advanced countries in terms of automated and pre-filled tax reporting, reducing the administrative burden on taxpayers while improving accuracy and compliance. It employs professionals across a range of disciplines including law, economics, IT, and data analysis.

www.skatteetaten.no

Panthera Institute

Panthera is a pioneer in Decision Intelligence. We design and implement decision environments that enable professional investors to perform at their highest level. Through our proprietary deep-tech solution, Decision GPS™, we amplify human ingenuity with artificial intelligence.

panthera.mc

The University of Edinburgh

The University of Edinburgh is a globally recognized research and education institution in the United Kingdom, with a large and diverse student body across undergraduate and postgraduate programs. It is organized into three Colleges and twenty-five Schools spanning Arts, Humanities and Social Sciences, Medicine and Veterinary Medicine, and Science and Engineering. The University supports research and teaching across disciplines and engages with external partners through research and advisory collaborations. The School of Economics contributes to this environment through research and teaching across core areas of economics, including behavioral economics. Research in this area focuses on decision-making under uncertainty, strategic interaction, and market outcomes. Drawing on theoretical, experimental, and computational methods, there is growing work exploring how digital technologies and artificial intelligence (AI) shape economic behavior and inform policy.

www.ed.ac.uk

Voya Financial

Voya Financial, Inc. is a leading health, wealth and investment company offering products, solutions and technologies that help individual, workplace and institutional clients become well planned, well invested and well protected. The Voya Behavioral Finance Institute for Innovation is focused on gaining deeper insights into the savings decisions of Americans to help transform their retirement. The Institute's work is differentiated by the ability to leverage behavioral science, a spectrum of lab and field experiments, and the speed and scale of the digital world to improve financial wellness outcomes.

www.voya.com