Thinking clearly about behavioural science and AI: A guide for the perplexed

Scaling interventions / Improving AI-systems / Personalized messaging

Summary

A three-dimensional matrix to provide a structured approach to identify the intersection of AI and behavioral science, categorizing the relationship based on Purpose (is AI a tool or a subject of study?), Scale (is the focus on individuals or entire systems?), and Interaction (is the engagement transactional or relational?).

Elina Halonen
Founder and Strategist at Prismatic Strategy

Artificial intelligence is reshaping how decisions are made, communicated, and acted upon. For organisations, this creates opportunities but it also creates uncertainty about where and how behavioural science can add value. The term “AI + behavioural science” can refer to very different activities:

- Using AI tools to accelerate research, review evidence, or scale interventions.
- Analysing how algorithms influence behaviour through personalisation, choice architectures or hypernudges.
- Applying behavioural insights to shape the design, alignment, or governance of AI systems themselves.

Each of these is valid, but without clarity, the label risks becoming so broad that it is hard for clients to act on, and hard for consultants to explain where they add value. To clarify these differences and make this emerging space easier to navigate, I developed a matrix. It maps the roles behavioural science can play with AI, the modes of engagement, and the kinds of systems involved.

For clients, the matrix can help:

- Scope and focus projects by identifying where behavioural science can add the most value
- Select the right expertise by ensuring the right skills are matched to the right kind of challenge
- Reduce the risk of blind spots by accounting for the full behavioural context of AI systems

The Matrix

To make sense of the many ways behavioural science intersects with AI, the matrix maps three core dimensions. Each one captures a meaningful difference in how we engage with AI systems, what kind of behavioural expertise is needed, and where the work is focused.
Before we get to the dimensions, there's one important distinction to keep in mind: the term AI can mean many things. In the past few years, AI has become synonymous with Generative AI, even though it is only a part of a bigger field. For simplicity, I will group everything else into "non-generative AI" that operates through classification, prediction, or personalisation. These includes e.g. recommender algorithms, credit scoring tools, or dynamic choice architectures – all of which have been around for much longer.

The matrix combines three intersecting axes:

- **Purpose:** Is AI being used as a tool to support behavioural science work, or is behavioural science being used as a lens to interrogate or guide AI systems?
- **Scale:** Is the focus on individual-level decisions and experiences, or on the broader population- or system-level dynamics that AI creates or reinforces?
- **Interaction:** Are humans engaging with AI in a one-off, transactional way, or in an ongoing, adaptive, relational dynamic?

Each axis gives a different view of the relationship between behavioural science and AI. Used together, they create a multidimensional map of where contributions can be made and where different kinds of expertise might be needed. In reality, projects often span multiple areas so these dimensions are best understood as continuums.

1. Purpose: AI as a tool vs. behavioural science as a lens

This axis helps clarify whether AI is the instrument being used, or the subject being examined: is AI supporting behavioural science work, or is behavioural science being used to understand or shape AI systems? At one end, AI is used as a tool to extend or enhance what behavioural scientists already do. This includes work where AI enables faster analysis, broader reach, or greater precision.

AI as a tool for behavioural science

Here, AI is used to support behavioural science workflows or goals, such as:

- Reviewing literature at scale
- Analysing behavioural data
- Designing or delivering personalised interventions
- Running simulations or virtual experiments

At the other end, behavioural science acts as a lens for interrogating how AI systems influence human behaviour—whether by shaping decisions, altering incentives, or introducing new sources of bias and uncertainty.

Behavioural science as a lens on AI

Here, behavioural science is used to examine, critique, or guide the development of AI systems:

- Identifying behavioural risks, biases, or blind spots
- Understanding how systems shape cognition, trust, or agency
- Aligning AI design with human needs, norms, and decision environments

Most real-world projects fall somewhere between these poles. For example, a team using an LLM to streamline intervention design might also need to assess how the model's outputs reflect its training data and embedded assumptions.

2. Scale: From individual experiences to system-level effects

This axis helps clarify the **level at which behavioural dynamics are being addressed**: is the work focused on **individual experiences and decisions**, or on **system-wide effects and patterns**?

At one end, behavioural science is applied to micro-level interactions, where AI influences how individuals perceive, decide, and act. This is where BeSci traditionally operates: modelling cognitive processes, shaping decision environments, or designing friction and feedback loops.

Micro-level focus

Here, behavioural insights are used to shape or evaluate how AI systems interact with individual users:

- Personalising recommendations or messages
- Reducing friction in decision pathways
- Modelling user attention, motivation, or trust
- Designing choice architectures in AI-enabled interfaces

At the other end, behavioural science is used to understand macro-level impacts—how AI systems affect populations, institutions, and social structures over time. This includes evaluating not just what an AI system does to one user, but how it scales, who it benefits or excludes, and how it reshapes norms and expectations.

Macro-level focus

This involves a systems-level lens on how AI influences behaviour at scale:

- Assessing institutional adoption and policy implications
- Investigating long-term effects on incentives, norms, or public trust
- Analysing disparities in access, influence, or outcomes
- Supporting governance, oversight, and accountability mechanisms

This is a particularly important space for behavioural science to contribute because our perspective helps link individual experiences with potential unanticipated, adverse consequences. For example:

- **Recommender systems:** Optimising for engagement can increase short-term relevance and user satisfaction, but also amplifies filter bubbles, reinforces confirmation bias, and contributes to social polarisation. These effects may go unnoticed unless behavioural science perspectives are applied at the system level.
- **Generative models used in decision support (e.g. LLM copilots):** While they increase speed and confidence in completing tasks, they may also introduce subtle distortions—users can over-trust outputs, anchor on misleading suggestions, or unknowingly internalise the model's assumptions. Behavioural science can help anticipate these risks by modelling how humans interact with AI-generated content over time.

3. Interaction: Transactional vs. relational dynamics

This axis highlights the nature of human–AI interaction. Are people engaging with the system in a one-off, outcome-focused way—or building ongoing, adaptive relationships over time?

Transactional dynamic

At one end, AI systems operate transactionally by generating outputs like credit scores, fraud alerts, or product recommendations in response to predefined inputs. These decisions often feel one-sided: people can't contest, discuss, or adapt them. Behavioural science can help organisations understand not just how people respond in the moment, but how these systems affect long-term perceptions of fairness, control, and trust. That matters because:

- People are more likely to disengage or push back when they feel excluded from decisions
- Opaque outcomes can damage brand trust, even when technically correct
- Small frictions or perceived unfairness can accumulate into reputational risk

Relational dynamic

This dimension brings the human–AI relationship into focus. Some systems do more than just give answers—they behave in ways that feel responsive, even social. As people interact with these tools over time, they start to form expectations, habits, and patterns of trust. How these relationships develop can shape not only what people do, but how they think and feel. Tools based on large language models often behave less like tools and more like collaborators or conversational partners. They respond to tone, adapt to prompts, and influence how users frame their own thinking. Behavioural science can help guide these interactions to support clarity, autonomy, and appropriate trust:

- Analysing how tone, fluency, or confidence shape user perceptions
- Exploring how trust, reliance, or overreliance build over repeated use
- Designing interactions that set realistic expectations about what the system can and can't do

This is another area where behavioural science offers distinctive value. As AI systems become more embedded in everyday interactions, the line between tool and teammate starts to blur. Understanding how users interpret, adapt to, or rely on these systems especially when the system feels responsive or social requires a behavioural lens because these dynamics involve trust, norms, and meaning making, which behavioural scientist are trained to understand. For example:

- **Opaque scores shape behaviour from a distance:** Risk scores and classification tools assign labels like creditworthiness, fraud risk, hiring potential that influence high-stakes decisions. Even when wrong or contestable, these scores feel definitive. Users adapt their behaviour based on opaque outputs, which can entrench disadvantage, erode trust, or drive workarounds.
- **Repeated interactions create relational expectations:** As users engage with AI assistants, copilots, or companions, they develop expectations that go beyond function. Tone, responsiveness, and memory shape how users interpret the system's intent, reliability, and personality. Over time, people may respond to these systems socially, even emotionally which creates new behavioural and ethical dynamics.

Pulling it together

These three dimensions – purpose, scale, and interaction – offer a structured way to think about the role of behavioural science in AI. The framework is intended to support reflection and strategic clarity so that we can notice what kind of dynamics are in play, what questions are being asked, and which ones might be missing.

The dimensions offer a way to map how behavioural science and AI fit together. In practice, most projects land somewhere in this three-dimensional space: for example, a tool built for individual users might end up shifting organisational norms, or a system designed for speed might evolve into something people rely on and relate to.

Every AI system ends up entangled with human expectations, assumptions, and responses. Even when the technical side is dominant, the behavioural consequences tend to show up: what feels like a data pipeline can turn out to be a trust problem, and what looks like a workflow optimisation can start to shape incentives or shift accountability.

As AI becomes more embedded in products, services, and decisions, behavioural perspectives can help teams ask better questions, see unintended effects earlier, and design with people in mind from the start.

Want more? Check these reads out.

AI for Perspective, Not Accuracy: Bringing Behavioural Science to Compliance

We're Leading Behavioural Science + AI for eCommerce Globally