A field guide to

PROBLEM ORIENTATIONS IN BEHAVIOURAL SCIENCE AND AI

Elina Halonen - April 2026

v. 1.0

Introduction

Work at the intersection of behavioural science and AI is often framed as a single, emerging field. In reality, it resembles a constellation of overlapping conversations—sharing vocabulary but lacking a common centre. Papers using similar terms may not address the same problem, and apparent disagreements often reflect different questions rather than true conflict.

The challenge isn't just the field's breadth; it's that diverse kinds of work are described with the same language. This can obscure disagreement, making it seem as though authors are debating identical questions when they're actually tackling distinct problems. At times, this creates the false impression that one perspective dismisses another, when in fact each is oriented toward a different task.

This guide offers a way to read the literature that clarifies these differences. Rather than starting with how papers claim to relate behavioural science and AI, it examines the *problems* they aim to solve. The goal isn't to impose a rigid taxonomy, but to map recurring *problem orientations* - patterns that shape how the field is currently explored.

Purpose of the guide

This guide doesn't evaluate which orientation is "correct." Instead, it distinguishes the problems each orientation revolves around, because those differences determine what each perspective highlights and what it overlooks.

As a practical tool, the guide helps you:

- **Locate** the problem a paper or argument is organised around.
- **Identify** what that focus illuminates.
- **Notice** what it tends to leave unexamined.

This approach doesn't resolve disagreements, but it reframes them. It becomes easier to recognise when different kinds of work are conflated and to engage with each on its own terms.

Most of the variation in this literature stems from two foundational questions:

1. **What is the object of interest?** Is the focus on human judgement and behaviour, AI systems themselves, or the broader sociotechnical systems embedding both?
2. **What is behavioural science being asked to do?** Is it tasked with design and intervention, explanation and evaluation, or organising the field itself?

The orientations that follow represent distinct answers to these questions.

This framing shifts the conversation from "Which approach is right?" to "What problem is this work addressing?" It equips readers to navigate the literature with greater precision, avoiding false conflicts and recognising the unique contributions of each orientation.

Seeing the field as a set of problems

This guide adopts a simple but powerful shift in perspective: instead of focusing on how a paper describes the relationship between behavioural science and AI, it starts with the *problem* the work aims to solve. This shift is crucial because the same terminology can describe very different kinds of work, while distinct language may actually address the same underlying concern.

To apply this approach, ask three key questions when engaging with research:

1. **What problem is this work organised around?** Is it about improving system performance, understanding human responses, scaling interventions, analysing system behaviour, or defining the field's scope? The answer shapes what counts as a meaningful contribution within that orientation.
2. **What does this focus highlight?** Each orientation brings certain phenomena into sharp relief (whether user behaviour, system dynamics, organisational constraints, or theoretical coherence) making them easier to address.
3. **What does it leave in the background?** Attention is inherently selective. What isn't central to the problem at hand is often simplified, deferred, or excluded. These omissions aren't necessarily mistakes, but they influence how the work should be interpreted.

By reading the literature through this lens, navigation becomes clearer. Papers that seem to disagree may actually address different problems, while those that appear aligned might be working at cross purposes. Identifying the problem orientation doesn't resolve these differences, but it makes them visible and easier to engage with. The following sections outline a set of recurring problem orientations that can be used in this way.

Notes on approach

This guide is neither a formal systematic review nor purely impressionistic. The material was gathered using search tools (Consensus.ai, Google Scholar, Liner.ai) and an existing collection of papers spanning behavioural science, AI, policy, and human–computer interaction.

Each source was analysed through a consistent lens: What is the object of interest? What role does behavioural science play? What level of analysis is prioritised, and what is overlooked? This approach allowed for comparisons across seemingly disparate works.

The goal was not to impose a rigid classification, but to surface recurring patterns in problem framing. The "problem orientations" described here are one such pattern—a way to map how different kinds of work are organised, not a prescription for how the field should be structured. The result is a working map: a tool for navigating the landscape while acknowledging where simplification falls short.

Two papers from human-centred AI and human–AI interaction are included as deliberate reference points, not as part of the core sample. Their inclusion highlights the blurred boundaries between behavioural science and adjacent fields, revealing whether shared questions are framed distinctly or overlap under different labels.

The problem orientations

The following sections outline recurring problem orientations in the behavioural science and AI literature. Each orientation represents a distinct way of defining the central problem, the kinds of contributions that matter, and what constitutes a satisfactory answer.

These orientations are not mutually exclusive - individual papers often span multiple orientations or shift between them. The goal isn't to pigeonhole each piece of work into a single category, but to illuminate the diverse problems that shape the field.

Each orientation is described in terms of the problem it addresses, how it treats AI, the role of behavioural science within it, and what that focus highlights or overlooks. Together, they offer a way to navigate the literature without assuming it forms a single, unified field.

Working with systems in practice

The first set of orientations focuses on making AI systems work in real-world settings. The primary concern is the interaction between people and systems - whether through user behaviour, system design, or observed responses during use.

In these approaches, behavioural science plays a practical role: it informs system design, deployment, and evaluation, bridging technical capabilities with outcomes that depend on human behaviour. The emphasis is on uptake, performance, and adaptation over time, often within organisational constraints.

Because these orientations are grounded in real-world interaction, they are immediately recognisable and actionable. However, their focus on implementation and optimisation can sometimes overshadow broader questions such as purpose, context, or system-level effects that remain less explicitly examined.

Core idea	Central problem	Brings into view	In the background
Implementers & intervention-scalers			
AI helps behavioural interventions and systems function in practice	How can AI help systems and interventions work effectively at scale and over time?	Strong on implementation, uptake, routines, and real-world constraints	Can treat "making it work" as more important than questioning underlying goals
Behavioural-risk & response analysts			
AI changes human judgement, emotion, trust, and behaviour	How does interacting with AI shape how people think, feel, and act?	Empirically tractable and often closest to measurable outcomes	Can individualise or psychologise structural and institutional problems
Designers & complementarians			
Psychology should inform AI design upstream	Which aspects of human cognition and behaviour should shape system design?	More specific and disciplined than broad "humans matter" claims	Often assumes better design is the primary solution

Implementers and intervention-scalers

Central problem: How can AI be used to make behavioural interventions and organisational systems work more effectively in practice at scale, over time, and under real-world constraints?

How AI is understood: AI is treated as delivery infrastructure embedded in workflows, services, and repeated user interactions. It appears as an adaptive support system, monitoring layer, or intervention mechanism that operates within existing organisational settings.

What behavioural science is doing: Behavioural science provides the intervention logic. It defines which behaviours matter, identifies points of friction, specifies mechanisms of change, and determines what successful adoption looks like. It is used to translate abstract system capability into workable routines and outcomes.

What this orientation brings into view: This orientation foregrounds the gap between technical capability and practical use. It makes visible how adoption depends on trust, incentives, routines, and integration within organisations. It is attentive to how behaviour can undermine otherwise well-designed systems, and to the fact that implementation is an ongoing process rather than a one-off event.

What it tends to leave in the background: By focusing on making systems work, it can narrow the problem to questions of optimisation and uptake. Broader issues (such as whether a system should be deployed, who benefits from it, or how it redistributes power and responsibility) may be treated as secondary or assumed rather than examined directly. Resistance can be interpreted as friction rather than as disagreement with underlying goals.

Representative examples:

- Behavioural Insights Team (2025), *Adopt, Align, Adapt, Augment*
- Mills (2023), *AI for Behavioural Science* (in its intervention-system strand)
- LSE perspective on AI-enabled behavioural intervention systems
- Pedersen et al. (2018), *Behavioural computer science*

Behavioural-risk and response analysts

Central problem: How does interacting with AI shape human judgement, emotion, trust, and behaviour over time?

How AI is understood: AI is treated as a source of influence on human cognition and behaviour. It appears as a system that people respond to, rely on, resist, or adapt to across repeated interactions.

What behavioural science is doing: Behavioural science is used as an analytical lens to study how AI affects human responses. It measures changes in judgement, trust, reliance, and decision-making, and identifies patterns in how people interact with and interpret AI systems.

What this orientation brings into view: This orientation produces empirically tractable insights and is often closest to measurable outcomes. It makes visible how AI systems affect trust, fairness perceptions, willingness to rely, and behavioural patterns in real contexts. It is often where the first systematic evidence of AI's behavioural effects emerges.

What it tends to leave in the background: By focusing on individual responses, it can frame problems in psychological terms even when they have structural or institutional causes. System design, organisational context, and power asymmetries may be treated as background conditions rather than as central objects of analysis.

Representative examples:

- Yalcin Williams & Lim (2024), *Psychology of AI*
- Turner & Reczek (2025), *Hype-free AI*
- Sacher et al. (2026), *The Missing Discipline in AI*

Designers and complementarians

Central problem: Which aspects of human cognition and behaviour should shape the design of AI systems, and how should those insights be translated into design decisions?

How AI is understood: AI is treated as a design object whose behaviour can be shaped upstream. The focus is on system architecture, interfaces, and interaction patterns, rather than on downstream use alone.

What behavioural science is doing: Behavioural science is used as a design input. It provides models of human cognition, decision-making, and behaviour that can be embedded into system design, guiding how systems present information, structure choices, and interact with users.

What this orientation brings into view: This orientation makes the role of psychology in design more explicit and more disciplined than broad claims that "humans matter." It highlights how assumptions about cognition and behaviour are already embedded in systems, whether intentionally or not, and treats design as a site where those assumptions can be made visible and adjusted.

What it tends to leave in the background: By focusing on improving design, it can treat system behaviour as primarily a design problem. Broader questions about deployment context, institutional incentives, or power relations may be less visible. There is also a tendency to assume that better design can resolve issues that may be structural rather than interface-level.

Representative examples:

- Gigerenzer (2024), *Psychological AI*
- Gonzalez & Malloy (2026), *Toward Complementary Intelligence*
- Schmager et al. (2025), *Understanding Human-Centred AI*

Studying systems as behavioural objects

The orientations that follow shift the object of study away from individual users toward AI systems and the environments in which they operate. Behaviour is no longer limited to human response, but extends to how systems act over time, how they interact with other systems and people, and how patterns emerge at a collective level.

In this part of the literature, behavioural science is used to describe, interpret, and sometimes intervene in systems that are treated as having behavioural properties of their own. This move expands what counts as behaviour, and with it, what kinds of questions can be asked.

Core idea	Central problem	Brings into view	In the background
Machine-behaviour and agent scholars			
AI systems themselves can be studied as behavioural objects	How do AI systems behave in context, especially over time and in interaction?	One of the more novel moves in the literature; shifts focus beyond users	"Behaviour" can become so broad that levels of analysis blur
Social-systems thinkers			
The object of study is the mixed human–machine system	What forms of collective behaviour emerge when humans and AI are entangled?	Strong on emergence, networks, and system-level effects	Can obscure questions of ownership, power, and design asymmetries

Machine-behaviour and agent scholars

Central problem: How do AI systems behave in context, especially over time and in interaction with humans, environments, and other systems?

How AI is understood: AI systems are treated as behavioural objects. The focus shifts from internal model properties to observable patterns of action, adaptation, and interaction across situations.

What behavioural science is doing: Behavioural science is used to describe, measure, and interpret system behaviour as if it were the output of an agent. It provides concepts and methods for observing patterns, comparing behaviours across contexts, and analysing how systems act in the world.

What this orientation brings into view: This orientation makes it possible to study AI systems in the wild, rather than only through benchmarks or internal metrics. It highlights how behaviour emerges through interaction with environments and other agents, and introduces a way of comparing systems based on what they do rather than how they are built. It is one of the more novel moves in the literature, because it extends behavioural analysis beyond human subjects.

What it tends to leave in the background: By treating systems as behavioural objects, the concept of "behaviour" can become very broad. Differences between levels of analysis—such as individual decisions, system outputs, and collective effects—can blur. Questions about design intent, ownership, and responsibility may be less visible when attention is focused on observed behaviour alone.

Representative examples:

- Rahwan et al. (2019), *Machine Behaviour*
- Chen et al. (2025), *AI Agent Behavioural Science*

Social-systems thinkers

Central problem: What forms of collective behaviour and social order emerge when humans and AI systems are entangled?

How AI is understood: AI is treated as part of a wider sociotechnical system in which outcomes arise from the interaction of multiple actors, technologies, and institutions.

What behavioural science is doing: Behavioural science is used to analyse patterns at the level of populations, networks, and systems. It contributes concepts for understanding coordination, feedback, emergence, and how local interactions scale into system-level effects.

What this orientation brings into view: This orientation highlights how AI reshapes social dynamics beyond individual interactions. It makes visible emergent patterns, network effects, and system-level consequences that cannot be reduced to individual behaviour alone. It is particularly attentive to how behaviour is distributed across actors and how outcomes arise from their interaction.

What it tends to leave in the background: By focusing on system-level patterns, it can obscure questions of ownership, power, and design asymmetries. The mechanisms by which specific actors shape systems, and the distribution of control within them, may be less visible when analysis centres on emergent dynamics.

Representative examples:

- Tsvetkova et al. (2024), *A New Sociology of Humans and Machines*
- Jackson et al. (2025), *AI Behavioral Science* (as a secondary overlap)

Building the infrastructure of the field

The next orientation shifts attention away from interaction and system behaviour toward the conditions that make behavioural work with AI possible in the first place. The focus is not primarily on users or systems, but on how behavioural knowledge is structured, formalised, and made usable at scale.

In this part of the literature, behavioural science is treated as something that can be encoded, organised, and operationalised. The emphasis is on turning concepts, evidence, and intervention logic into forms that can be applied consistently across systems, datasets, and contexts.

Core idea	Central problem	Brings into view	In the background
Infrastructure builders			
AI can make behavioural science more computable and usable	How can behavioural knowledge be formalised, operationalised, and scaled?	Concrete on taxonomies, evidence structures, and delivery mechanisms	Risks mistaking formalisation for progress

Infrastructure builders

Central problem: How can behavioural science be made more usable, operational, and scalable in the context of AI systems?

How AI is understood: AI is treated as a means of structuring, processing, and deploying behavioural knowledge. It appears as infrastructure that supports classification, prediction, and the integration of behavioural evidence into systems.

What behavioural science is doing: Behavioural science is formalised and translated into structured representations. This includes building ontologies, defining variables and relationships, and creating systems that can extract, organise, and apply behavioural knowledge in a consistent way.

What this orientation brings into view: This orientation makes the practical requirements of scale visible. It highlights the need for shared definitions, structured data, and repeatable processes, and shows how behavioural knowledge can be made more accessible and usable across different applications.

What it tends to leave in the background: By focusing on formalisation, it can treat the process of structuring knowledge as equivalent to improving it. Questions about whether categories are adequate, how context affects interpretation, or whether formalisation captures what matters can become secondary.

Representative examples:

- Mac Aonghusa & Michie (2020), *Through the Looking Glass*
- Pedersen et al. (2018), *Behavioural computer science*
- LSE perspective on AI-enabled behavioural intervention systems[1]

Structuring and defining the field

The next set of orientations is less concerned with specific systems or applications and more with how the space itself is organised. The focus shifts from solving practical problems to clarifying concepts, defining boundaries, and establishing standards for explanation.

In this part of the literature, behavioural science is used to interpret, classify, and evaluate. The emphasis is on how different kinds of work should be distinguished, how claims should be assessed, and what it means for behavioural science to contribute meaningfully in this context.

[1] LSE perspective is based on a holistic analysis of the research themes because there was no single paper that would be representative (more information in references).

Core idea	Central problem	Brings into view	In the background
Field-cartographers			
The field needs conceptual organisation	What are the right categories for organising this space?	Helps prevent category errors and false unification	Taxonomy can substitute for theoretical development
Theory-protectors			
AI should extend, not dilute, standards of explanation	Does AI improve theory, or encourage conceptual confusion?	Maintains epistemic discipline and standards of explanation	Can leave practical and governance questions underdeveloped

Field-cartographers

Central problem: What are the right categories for organising the behavioural science and AI landscape, and how should different kinds of work be distinguished?

How AI is understood: AI appears as a heterogeneous set of systems and applications that are often grouped together under a shared label. The focus is on how these are described, compared, and classified.

What behavioural science is doing: Behavioural science is used as an organising framework. It provides concepts and distinctions that help sort different kinds of work, identify category boundaries, and make the structure of the field more explicit.

What this orientation brings into view: This orientation helps prevent category errors and false unification. It makes visible when different kinds of work are being treated as though they are the same, and provides a way to compare approaches without assuming a single underlying logic.

What it tends to leave in the background: By focusing on classification and organisation, it can treat the act of mapping as sufficient. Questions about how the field should develop in practice, or how competing approaches should be prioritised, may remain under-specified.

Representative examples:

- Xu et al. (2024), *AI for social science and social science of AI*
- Schmager et al. (2025), *Understanding Human-Centred AI*
- Guest (2025), *What Does "Human-Centred AI" Mean?*

Theory-protectors

Central problem: Does the integration of AI strengthen behavioural science as a field, or does it risk diluting its standards of explanation?

How AI is understood: AI is treated as a source of conceptual pressure. It introduces new methods, metaphors, and claims that may or may not align with existing theoretical frameworks.

What behavioural science is doing: Behavioural science is used to evaluate and defend standards of explanation. It assesses whether new approaches remain consistent with

established theories, and whether claims made in the context of AI are conceptually coherent.

What this orientation brings into view: This orientation maintains epistemic discipline. It highlights when concepts are stretched, when explanations become vague, and when new terminology obscures rather than clarifies underlying mechanisms.

What it tends to leave in the background: By focusing on conceptual clarity, it can leave practical questions underdeveloped. Issues related to implementation, governance, or system-level effects may be treated as secondary to maintaining theoretical consistency.

Representative examples:

- van Rooij et al. (2024), Reclaiming AI as a Theoretical Tool for Cognitive Science
- Mills (2025), The Public Responsibilities of Behavioural Scientists in an Age of AI (as a partial overlap)

Expanding and framing the field

The final set of orientations is less concerned with analysing specific systems or clarifying concepts, and more with shaping how the field itself is defined. The focus shifts toward questions of scope, relevance, and responsibility.

In this part of the literature, behavioural science is used to make claims about its role in the AI domain. These claims are often partly analytical and partly jurisdictional, in that they do not only describe the space but also participate in defining its boundaries and priorities.

Core idea	Central problem	Brings into view	In the background
Field-builders and expansionists			
Behavioural science should play a larger role in the AI domain	What role should behavioural science have in shaping AI?	Clarifies the relevance of behavioural expertise in an expanding field	Can treat adjacency as integration and expand scope without resolving tensions
Humanisers			
AI should be designed and evaluated in human-centred terms	How can AI systems better reflect human values, experience, and dignity?	Keeps trust, fairness, and lived experience in view	"Human-centred" can remain normatively appealing but analytically vague

Field-builders and expansionists

Central problem: What role should behavioural science play in shaping the development, use, and governance of AI?

How AI is understood: AI is treated as a broad domain of activity with significant behavioural and societal consequences. The emphasis is on its reach across sectors and its potential to reshape decision-making environments at scale.

What behavioural science is doing: Behavioural science is positioned as a relevant and often necessary contributor to this domain. It is used to frame AI as a behavioural

problem space and to argue for the inclusion of behavioural expertise in its development and governance.

What this orientation brings into view: This orientation makes the relevance of behavioural science visible in contexts where it might otherwise be overlooked. It highlights how AI systems shape behaviour, institutions, and decision-making processes, and raises the question of who should be involved in defining and governing these systems.

What it tends to leave in the background: By linking behavioural science to a wide range of adjacent problems, it can treat proximity as evidence of integration. The scope of the field may expand faster than its conceptual foundations, and tensions between different uses of behavioural science can remain unresolved.

Representative examples:

- Jackson et al. (2025), *AI Behavioral Science*
- Behavioral AI Institute (2025), open letter
- Mills, Costa, & Sunstein (2023), *The Opportunities and Costs of AI in Behavioural Science*
- Mills (2025), *The Public Responsibilities of Behavioural Scientists in an Age of AI*

Humanisers

Central problem: How can AI systems be designed, evaluated, and governed in ways that reflect human values, experience, and dignity?

How AI is understood: AI is treated as a system that affects people in ways that are not only technical, but also social and experiential. The focus is on how systems are perceived, trusted, and lived with.

What behavioural science is doing: Behavioural science is used to articulate and interpret human experience. It contributes to defining what counts as fair, trustworthy, or acceptable, and helps translate abstract values into design and evaluation criteria.

What this orientation brings into view: This orientation keeps questions of trust, fairness, and lived experience in view. It highlights how systems are encountered by users and how those encounters shape acceptance, resistance, and longer-term relationships with AI.

What it tends to leave in the background: Because "human-centred" framing can be broad, it may remain at a high level of abstraction. The mechanisms by which values are translated into design decisions, or how competing values are prioritised, may be less clearly specified.

Representative examples:

- Fenwick & Molnar (2022), *The Importance of Humanizing AI*
- Schmager et al. (2025), *Understanding Human-Centred AI*
- Guest (2025), *What Does "Human-Centred AI" Mean?*
- Sacher et al. (2026), *The Missing Discipline in AI* (overlap)

A summary of problem orientations in BeSci x AI

FUNCTION OF BEHAVIOURAL SCIENCE	OBJECT OF STUDY: Humans reacting to AI	AI design & intervention systems	AI systems & mixed sociotechnical systems
Conceptual, critical or jurisdictional (interpreting, evaluating, claiming territory)	**Humanisers** (normative framing across design and impact) **Behavioural-risk analysts** (when focused on diagnosis and critique) **Field-cartographers** (when interrogating concepts and definitions)	**Theory-protectors** (guarding standards of explanation) **Field-cartographers** (when organising the conceptual terrain)	**Field-builders and expansionists** (claiming AI needs behavioural science) **Social-systems thinkers** (emergent collective order) **Machine-behaviour scholars** (when building the field)
Practical, design or implementation (building, running, improving systems)	**Implementers & intervention-scalers** (making AI work in practice) **Behavioural-risk analysts** (when improving outcomes in use)	**Designers & complementarians** (psychology inside AI design) **Infrastructure builders** (making behavioural science computable and scalable)	**Machine-behaviour and agent scholars** (studying AI systems as behavioural objects)

A compressed view of the landscape

The table maps the "orientations" (different ways of framing problems) against these two dimensions. It shows how each orientation clusters, but it's not a strict classification because some orientations span multiple roles or objects of interest. For example, Implementers and intervention-scalers focus on humans reacting to AI and use behavioural science in an applied way (e.g., improving uptake or outcomes).

	BeSci as application (design, intervention, implementation)	**BeSci as lens** (explanation, evaluation, classification, critique)
Humans reacting to AI	**Implementers and intervention-scalers** (when focused on uptake and outcomes in use) **Behavioural-risk and response analysts** (when improving behavioural outcomes in practice)	**Behavioural-risk and response analysts** (diagnosing effects on judgement, trust, behaviour) **Field-cartographers** (when interrogating concepts and definitions) **Humanisers** (normative framing of experience and impact)
AI design and intervention systems	**Designers and complementarians** (psychology informing AI design) **Infrastructure builders** (making behavioural science computable and scalable)	**Theory-protectors** (standards of explanation and interpretability) **Field-cartographers** (organising the conceptual terrain)
AI systems & sociotechnical systems	**Machine-behaviour and agent scholars** (when studying behaviour in order to shape or operate systems)	**Machine-behaviour and agent scholars** (studying systems as behavioural objects) **Social-systems thinkers** (emergent collective dynamics) **Field-builders and expansionists** (defining scope and claiming jurisdiction)

The compression helps reveal that disagreements in the literature can stem from different assumptions about what is being studied and what behavioural science is for. When these assumptions aren't explicit, work organised around different problems can seem to conflict, even if they're just addressing different questions.

How the literature evolves

The "orientations" described earlier help reveal differences in the literature, but in practice, work rarely fits neatly into one category. Instead, research often moves between orientations, combines elements from several, or shifts focus depending on the argument. These movements follow a few recurring patterns that shape how the field appears and develops.

1. Expansion across adjacent problems: Research often starts with a specific question (e.g., how AI affects decision-making) and expands to include related issues like design, governance, or societal impact. This makes behavioural science seem broadly relevant, but it can blur distinctions between different types of work. As the scope widens, proximity can replace true integration, leaving tensions between behavioural concepts unresolved.

2. Movement between application and analysis: Work frequently shifts between applied roles (e.g., designing interventions) and analytical roles (e.g., explaining or critiquing AI's effects). While this can be productive, it can also obscure which problem is being addressed. When analysis and application are combined without clear separation, claims about "what works" and "what is happening" can become entangled.

3. Shifts in the object of study: Research often moves between levels of analysis - from individual users to AI systems, and from systems to broader sociotechnical environments. This expands what counts as relevant, but it also changes the level at which explanations operate. Findings at one level (e.g., user behaviour) may not directly apply to another (e.g., collective dynamics), and distinctions between these levels can blur.

4. Formalisation and operationalisation: Behavioural concepts are often translated into structured forms (e.g., taxonomies, ontologies, or computational models) to make them usable in AI systems. This enables scalability and consistency but can also lock in specific interpretations of behavioural concepts. Once formalised, these categories may become embedded in systems, even if they remain contested or context-dependent.

5. Conceptual unification: There are repeated attempts to unify the field through overarching frameworks or shared definitions. While this can bring coherence and simplify communication, it can also create the illusion of agreement where underlying problem orientations differ. Conceptual unification may obscure the diversity it aims to address.

These movements help explain why the literature can appear more coherent than it actually is. Shared language, shifting levels of analysis, and expanding scope create the impression that different contributions are addressing the same problem even when they're not. Recognising these patterns doesn't resolve differences between orientations, but it makes them easier to see. It also clarifies how the field evolves: not just through new findings, but through how problems are defined, extended, and connected. These patterns shape what questions are treated as central and how the field grows.

Closing thoughts

Work at the intersection of behavioural science and AI often appears more unified than it is. Shared language and overlapping interests create the impression of a single, cohesive field, but in reality, the literature is organised around different problems, each with its own assumptions, methods, and criteria for success.

This guide doesn't attempt to resolve these differences. Instead, it makes them visible by mapping recurring problem orientations and showing how they relate. The value of this map is practical: it helps you locate a piece of work, understand what problem it's addressing, and see what that focus highlights or overlooks.

This doesn't eliminate disagreement - it just changes how it appears. Work that seems to conflict may simply be organised around different problems, while work that appears aligned may be operating at cross purposes. Making these distinctions explicit allows each contribution to be understood on its own terms, rather than as part of a field that only appears coherent.

The goal isn't to fix the field's structure but to make it easier to navigate as it continues to evolve.

References

Behavioral AI Institute. (2025, July 24). *A call for behavioral scientists, social scientists and AI leaders to actively shape the future of AI together* [Open letter].

Behavioural Insights Team. (2025). *AI & human behaviour: adopt, align, adapt, augment*.

Chen, L., Zhang, Y., Feng, J., Chai, H., Zhang, H., Fan, B., Ma, Y., Zhang, S., Li, N., Liu, T., Sukiennik, N., Zhao, K., Li, Y., Liu, Z., Xu, F., & Li, Y. (2025). *AI Agent Behavioral Science*. arXiv.

Fenwick, A., & Molnar, G. (2022). *The importance of humanizing AI: using a behavioral lens to bridge the gaps between humans and machines*. *Discover Artificial Intelligence, 2*, 14. https://doi.org/10.1007/s44163-022-00030-8

Gigerenzer, G. (2024). *Psychological AI: Designing algorithms informed by human psychology*. *Perspectives on Psychological Science, 19*(5), 839–848. https://doi.org/10.1177/17456916231180597

Gonzalez, C., & Malloy, T. (2026). *Toward complementary intelligence: Integrating cognitive and machine AI*. *Current Directions in Psychological Science*, 1–9. https://doi.org/10.1177/09637214251407571

Guest, O. (2025). What Does' Human-Centred AI'Mean?. *arXiv preprint arXiv:2507.19960*.

Jackson, M. O., Mei, Q., Wang, S. W., Xie, Y., Yuan, W., Benzell, S., Brynjolfsson, E., Camerer, C. F., Evans, J., Jabarian, B., Kleinberg, J., Meng, J., Mullainathan, S., Ozdaglar, A., Pfeiffer, T., Tennenholtz, M., Willer, R., Yang, D., & Ye, T. (2025). *AI Behavioral Science*. arXiv.

LSE perspective is based on a joint analysis of the following article, supplemented by the overall research theme: How agentic AI can be applied to behavioural science (March 2025) LSE Business Review.

Mac Aonghusa, P., & Michie, S. (2020). Artificial intelligence and behavioral science through the looking glass: Challenges for real-world application. *Annals of Behavioral Medicine*, *54*(12), 942-947. https://doi.org/10.1093/abm/kaaa095

Mills, S. (2023). *AI for behavioural science*. CRC Press. https://doi.org/10.1201/9781003203315

Mills, Stuart and Costa, Samuel and Sunstein, Cass R., The Opportunities and Costs of AI in Behavioural Science (June 23, 2023). Harvard Public Law Working Paper Forthcoming, Available at SSRN: https://ssrn.com/abstract=4490597 or http://dx.doi.org/10.2139/ssrn.4490597

Mills, Stuart, The Public Responsibilities of Behavioural Scientists in an Age of AI (August 15, 2025). Available at SSRN: https://ssrn.com/abstract=5393229 or http://dx.doi.org/10.2139/ssrn.5393229

Pedersen, T., Johansen, C., & Jøsang, A. (2018). *Behavioural computer science: An agenda for combining modelling of human and system behaviours*. *Human-centric Computing and Information Sciences, 8*, 7. https://doi.org/10.1186/s13673-018-0130-0

Rahwan, I., Cebrian, M., Obradovich, N., Bongard, J., Bonnefon, J.-F., Breazeal, C., Crandall, J. W., Christakis, N. A., Couzin, I. D., Jackson, M. O., Jennings, N. R., Kamar, E.,

Kloumann, I., Larochelle, H., Lazer, D., McElreath, R., Mislove, A., Parkes, D. C., Pentland, A., Roberts, M. E., Shariff, A., Tenenbaum, J. B., & Wellman, M. (2019). *Machine behaviour*. *Nature, 568*, 477–486. https://doi.org/10.1038/s41586-019-1138-y

Sacher, Paul and Michie, Susan and Hauser, Oliver and Ferrère, Antoine and Salzer, Samuel and Rodger, Amy and Borg, Jana and Pogrebna, Ganna and Murphy, Susan A., The Missing Discipline in AI: A Call for Behavioural Science (January 21, 2026). Available at SSRN: https://ssrn.com/abstract=6108948 or http://dx.doi.org/10.2139/ssrn.6108948

Schmager, S., Pappas, I. O., & Vassilakopoulou, P. (2025). Understanding Human-Centred AI: a review of its defining elements and a research agenda. *Behaviour & Information Technology*, *44*(15), 3771-3810.

Tsvetkova, M., Yasseri, T., Pescetelli, N., & Werner, T. (2024). *A new sociology of humans and machines*. *Nature Human Behaviour, 8*, 1864–1876. https://doi.org/10.1038/s41562-024-02001-8

Turner, B. L., Jr., & Reczek, R. W. (2025). *Hype-free AI: How AI actually impacts psychology in research, the workplace, the marketplace, and beyond*. *Current Opinion in Psychology, 61*, 101939. https://doi.org/10.1016/j.copsyc.2024.101939

van Rooij, I., Guest, O., Adolfi, F., de Haan, R., Kolokolova, A., & Rich, P. (2024). *Reclaiming AI as a theoretical tool for cognitive science*. *Computational Brain & Behavior, 7*, 616–636. https://doi.org/10.1007/s42113-024-00217-5

Xu, R., Sun, Y., Ren, M., Guo, S., Pan, R., Lin, H., Sun, L., & Han, X. (2024). *AI for social science and social science of AI: A survey*. arXiv preprint arXiv:2401.11839.

Yalcin Williams, G., & Lim, S. (2024). *Psychology of AI: How AI impacts the way people feel, think, and behave*. *Current Opinion in Psychology, 58*, 101835. https://doi.org/10.1016/j.copsyc.2024.101835

Appendix

Paper / perspective	Dominant level	What AI is in this view	What behavioural science is doing	What problem is it really solving?	Real ambition
Yalcin Williams & Lim (2024), *Psychology of AI*	Individual	Explains how people feel, think, and behave when AI occupies human-like roles	Comparator and social substitute — AI manager, chatbot, doctor, decision-maker	Maps psychological response patterns, especially aversion, fairness, trust, willingness to rely	Establish a psychology-of-response lane
Gigerenzer (2024), *Psychological AI*	Individual	Uses psychology to build better algorithms for uncertain environments	Algorithmic design object that can learn from heuristics	Source of upstream design principles for robustness and transparency	Use psychology as engineering resource
LSE perspective: behavioural science as an AI-enabled intervention system	Individual	Makes behaviour-change interventions more scalable, adaptive, and continuously evaluated	Measurement system, simulation tool, delivery mechanism, feedback-loop engine, later a persistent intervention agent	Defines targets, designs interventions, interprets outcomes; AI operationalises and scales	Same behavioural engine, more powerful machinery
Sacher et al. (2026), *The Missing Discipline in AI*	Interaction	Inserts behavioural effects into responsible AI evaluation	Influence system shaping trust, dependence, confidence, and behaviour over time	Missing infrastructure for risk identification, measurement, design, and lifecycle monitoring	Behaviouralisation of responsible AI
Mills (2023), *AI for Behavioural Science*	Interaction	Clarifies what changes when AI automates behavioural influence	Autonomous choice architect that personalises and reconfigures decisional contexts	Supplies the concepts and grammar of behavioural influence, while interrogating the translation	Reframe AI as adaptive behavioural architecture
Chen et al. (2025), *AI Agent Behavioral Science*	Interaction	Shifts attention from model internals to situated agent behaviour in context	Situated behavioural agent interacting with environments, agents, and humans	Organising language for observing, measuring, interpreting, and shaping agent behaviour	Agent-focused umbrella building
Gonzalez & Malloy (2026), *Toward Complementary Intelligence*	Interaction	Integrates machine AI and cognitive AI for better human-AI complementarity	Hybrid partner system combining scalable optimisation with psychologically grounded representations	Provides mechanisms and human standards for alignment and interpretability	Disciplined integration pitch
Turner & Reczek (2025), *Hype-free AI*	Mixed, but closer to interaction	Grounds discussion in current, observable psychological effects rather than AGI theatre	Present-day family of tools affecting trust, work, consumption, and judgement	Empirical mapping of where AI already affects psychology	Taxonomic discipline and empirical grounding
Behavioural Insights Team (2025), *Adopt, Align, Adapt, Augment*	Organisational	Solves an adoption and transition problem around AI uptake and use	Tool embedded in workflows and systems	Practical toolkit for adoption, trust, alignment, and adaptation	Make AI implementation work in practice
Mac Aonghusa & Michie (2020), *Through the Looking Glass*	Organisational	Solves a research-infrastructure problem: messy behavioural evidence cannot be synthesised well by humans alone	Research infrastructure for extraction, ontology-mapping, and prediction	Supplies ontology, annotation, schema, and expert judgement	Make behavioural evidence computable
Mills, Costa, & Sunstein, *The Opportunities and Costs of AI in Behavioural Science*	Organisational	Asks how AI can upscale behavioural science as research and intervention enterprise	Analytical and intervention amplifier — pattern detector, personalisation engine, complexity modeller	Uses AI to detect heterogeneity, model complexity, and target interventions better	Scale behavioural science without revisiting enough of its older legitimacy problems
Behavioral AI Institute (2025), *open letter*	Societal	Claims behavioural and social scientists should help shape AI's future	Socially consequential force affecting behaviour, decisions, and institutions	Design input, safety evaluator, governance claimant, coalition-builder	Jurisdictional expansion
Mills (2025), *The Public Responsibilities of Behavioural Scientists in an Age of AI*	Societal	Examines how behavioural science can legitimise AI politics and public authority	Publicly packaged solution whose behavioural framing can naturalise automation	Critical self-scrutiny of behavioural science's public role	Critique of behavioural science's own expansion
Rahwan et al. (2019), *Machine Behaviour*	Societal	Founds a science of what AI systems do in the wild beyond benchmarks	A class of actors with behavioural patterns and ecology	Provides empirical logic and methods for studying opaque systems in context	Found a science of AI behaviour
Fenwick & Molnar (2022), *The Importance of Humanizing AI*	Societal	Corrects AI development that neglects human enhancement, values, and experience	Sociotechnical force that should be humanised across design, application, and governance	Human-centred lens connecting cognition, trust, ethics, and social impact	Broad moral-organising framework
Tsvetkova et al. (2024), *A New Sociology of Humans and Machines*	Societal	Explains mixed human-machine systems and their collective outcomes	Population of heterogeneous social actors inside larger systems	One ingredient in a broader sociological and complex-systems account	Shift from psychology to hybrid social order
Xu et al. (2024), *AI for Social Science and Social Science of AI*	Meta / field-organisation	Solves a classificatory problem in a fragmented field	Explicitly split between AI as research tool and AI as social object	Provides categories and field structure more than a substantive theory	Sort the landscape and reduce category confusion
Jackson et al. (2025), *AI Behavioral Science*	Multi-level, tilted societal	Defines a broad field for studying, assessing, and guiding AI in society	Behaving system, research tool, and part of socio-technical systems	Assessing, inferring, modelling, designing, and helping govern	Field formation through expansion
Guest (2025), *What Does "Human-Centred AI" Mean?*	Conceptual / critical	Challenges whether "human-centred AI" means anything distinctive at all	Not separable from human cognition and labour in any simple way	Critical conceptual analysis rather than programme-building	Deflate the concept rather than stabilise it
Schmager et al. (2025), *Understanding Human-Centred AI*	Conceptual / framework	Clarifies conceptual ambiguity in HCAI by extracting common elements and imposing structure	Socio-technical system embedded in human contexts	Diffuse input into design and evaluation via values, stakeholder needs, usability, and interaction	Stabilise HCAI as a design paradigm
van Rooij et al. (2024), *Reclaiming AI as a Theoretical Tool for Cognitive Science*	Conceptual / theoretical	Resists the slide from computational description to claims of human-like machine cognition	Either flawed AI-as-engineering ideology or useful AI-as-theoretical tool	Cognitive science sets standards for what counts as explanation of mind	Defend theory against engineering overreach
Pedersen et al. (2018), *Behavioural Computer Science*	Mixed, leaning conceptual / methodological	Combines modelling of human and system behaviours within one agenda	System whose behaviour can be studied experimentally, as if outputs were decisions	Applies behavioural methods to computational systems and their outputs	Bridge human and system behaviour under one method agenda

First iteration of the mapping

More details: https://artificialthought.substack.com/p/how-is-the-behavioural-science-and

How is the behavioural science and AI relationship imagined?

Papers mapped across five relational stances: clear fit (navy) and partial fit (orange).

Paper	Behavioural science for AI	Behavioural science of AI-mediated human conduct	Behavioural science through AI	Behavioural science around AI deployment	Behavioural science against its own overreach	Notes
Gigerenzer (2024) — Psychological AI	Clear fit					Psychology informing algorithm design
Chen et al. (2025) — AI Agent Behavioral Science	Clear fit					Also strains toward "AI behaviour itself" rather than just "for AI"
Gonzalez & Malloy (2026) — Toward Complementary Intelligence	Clear fit					Behavioural/cognitive science informing AI design and integration
Jackson et al. (2025) — AI Behavioral Science	Clear fit	Partial fit	Partial fit	Partial fit		Broad umbrella paper; spans AI as object, tool, and socio-technical system
Pedersen et al. (2018) — Behavioural Computer Science	Partial fit					Tentative placement; would treat as provisional pending closer reading
Rahwan et al. (2019) — Machine behaviour	Partial fit					Really points to a missing category: behavioural science of AI behaviour itself
Turner & Reczek (2025) — Hype-free AI		Clear fit				Psychology/effects-of-AI framing
Yalcin Williams & Lim (2024) — Psychology of AI		Clear fit				Squarely about how AI affects how people feel, think, and behave
Behavioral AI Institute (2025) — A call for behavioral scientists, social scientists and AI leaders...	Partial fit	Clear fit		Partial fit		Field-building paper with strong focus on missed behavioural effects in AI use
Sacher et al. (2026) — The missing discipline in AI	Partial fit	Clear fit				Strongest on AI's effects on human thought, feeling, trust, and behaviour
Tsvetkova et al. (2024) — A new sociology of humans and machines		Clear fit		Partial fit		More mixed social-systems than individual psychology
Mills (2023) — AI for Behavioural Science		Partial fit	Clear fit		Partial fit	Sits between AI as behavioural-science instrument and AI as behavioural influence system
Xu et al. (2024) — AI for Social Science and Social Science of AI		Partial fit	Clear fit			Explicitly mixed by title; straddles tool/object distinction
Mac Aonghusa & Michie (2020) — AI and Behavioral Science Through the Looking Glass			Clear fit			AI as research infrastructure for behavioural science
Mills, Costa & Sunstein — The Opportunities and Costs of AI in Behavioural Science			Clear fit			AI as amplifier of behavioural-science analysis and intervention
LSE perspective — Behavioural science as an AI-enabled intervention system			Clear fit			AI as measurement, simulation, and adaptive intervention infrastructure for behavioural science
Behavioural Insights Team (2025) — AI & human behaviour: adopt, align, adapt, augment		Partial fit		Clear fit		Best read as deployment-centred, with a human-effects strand
Mills (2025) — The Public Responsibilities of Behavioural Scientists in an Age of AI					Clear fit	Cleanest critical/political paper in the set
van Rooij et al. (2024) — Reclaiming AI as a Theoretical Tool for Cognitive Science		Partial fit			Clear fit	Critical of conceptual overreach; also uses AI as theoretical instrument
Fenwick & Molnar (2022) — The importance of humanizing AI	Partial fit	Partial fit		Partial fit		Broad human-centred umbrella; crosses design, experience, and governance/deployment
Schmager et al. (2025) — Understanding Human-Centred AI	Partial fit			Partial fit		Conceptual framework for HCAI; design-and-governance oriented
Guest (2025) — What Does "Human-Centred AI" Mean?					Clear fit	Critical reframing: all AI is human-involved; evaluates enhancement vs displacement

Clear fit — the relationship is central to the paper Partial fit — the relationship is present but not primary