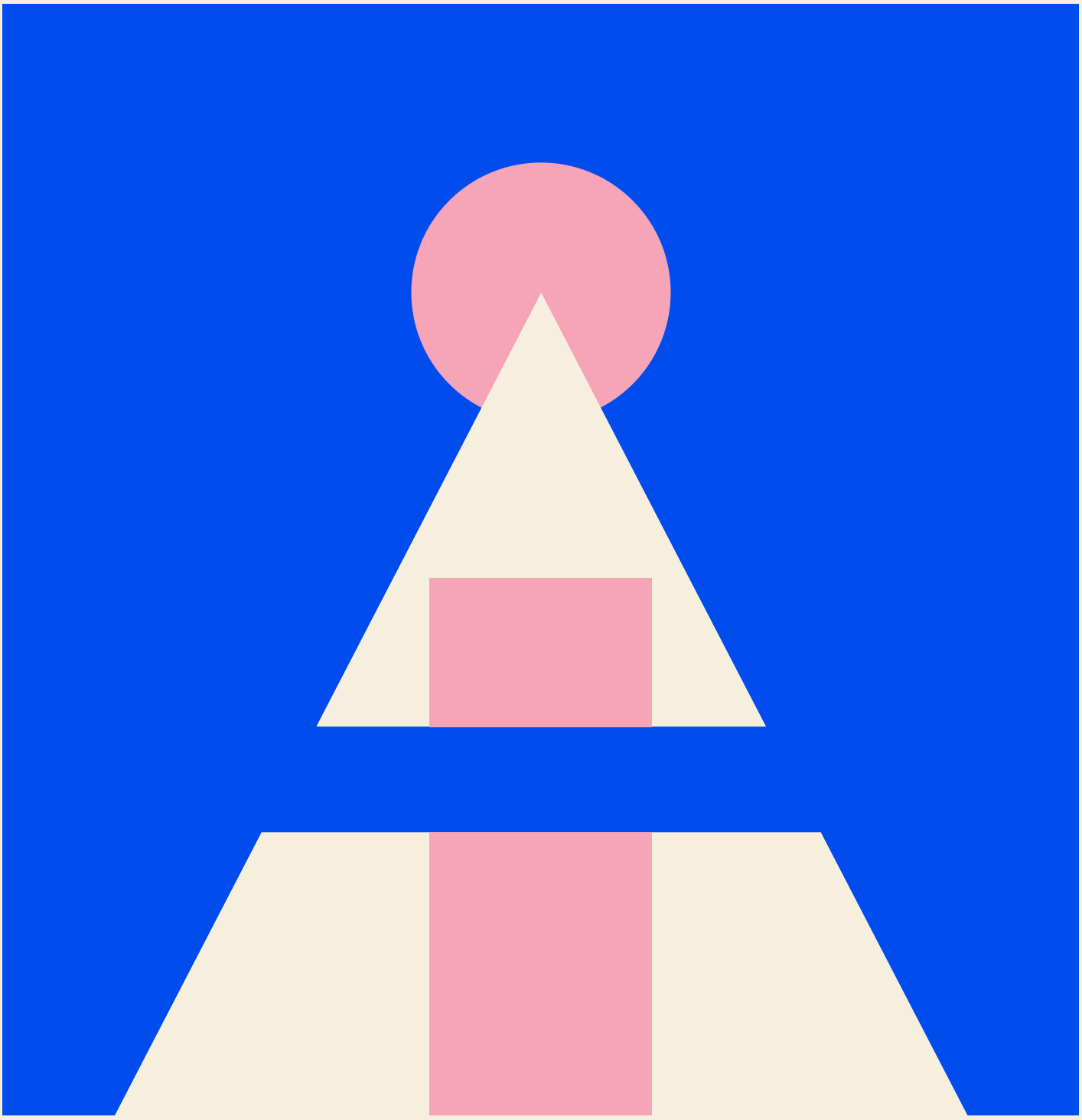
BIT

AI & HUMAN BEHAVIOUR

AUGMENT,
ADOPT,
ALIGN,
ADAPT



Executive Summary

Why behavioural science matters in an AI world

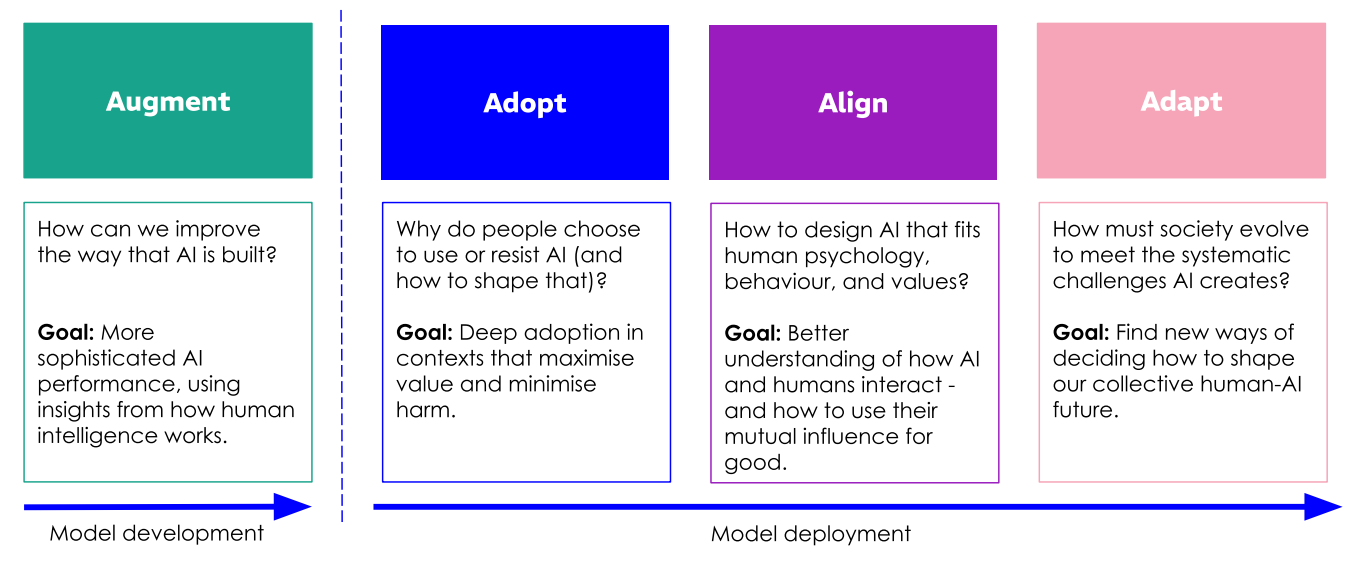
The rise of generative AI has triggered an explosion of attention, spending, and organisational change. Worldwide outlays on generative AI are **forecast to hit $644 billion** in 2025. One in four organisations **reports using AI** in at least one business function. **Half of US adults** have used a Large Language Model (LLM) like ChatGPT, Gemini, Claude, Grok, Deepseek or Copilot.

Yet the drive for economic and technological progress has largely neglected a crucial factor: human behaviour. The promise of AI can only be fulfilled by understanding how and why people think and act the way they do. Organisations will reap greater rewards if they know the best way to get **humans and AI agents working together**. Chatbots and agents will be more accepted if we understand how **preferences and perceptions evolve through mutual influence**. And, some argue, AI researchers will make the next breakthrough in performance by **taking inspiration from how humans think**.

At the same time, human behaviour is central to avoiding the potential pitfalls that many see ahead and, more importantly, harnessing the opportunities. How are our interactions with AI affecting our beliefs and behaviours, both instantly and over time? What is the cumulative effect on our societies – and how should we anticipate, adapt or mitigate those changes? How can AI understand our needs and goals?

Behavioural science can offer the insights to meet these challenges. **But we need to act on them quickly**. The fluidity of the past few years will soon solidify – we will get 'locked into' arrangements. Now is the time to make active, deliberate choices that ensure we build a version of AI that is sensitive and responsive to human needs and behaviours, and forge a positive human-AI future.

After decades of working on behavioural science, we believe this approach can address four fundamental issues facing AI: how behavioural science can **augment** AI's capabilities; why individuals **adopt** or resist AI; how we can **align** AI design with human psychology; and how society must **adapt** to the impacts of AI.



Augmenting AI: using behavioural insights to improve how AI is built

The idea that behavioural science can improve the fundamental construction of AI may be new to many. Yet, insights from human cognition have long inspired AI research – and continue to do so at the cutting edge of model development.

While current generative AI models are powerful, they are essentially 'fast thinkers', operating like the human brain's intuitive and associative **System 1**. This makes them masters of pattern recognition, but also leaves them vulnerable to the same kinds of biases that affect human intuition. To overcome these limitations, we need to build AI that can also 'think slow'.

However, the goal is not simply to bolt on a more deliberate, analytical 'System 2'. The true key to advancing AI lies in developing **metacognition** – the ability to think about thinking. What makes human intelligence so flexible is our ability to match our cognitive strategy to the task at hand.

Therefore, we argue for the development of a **metacognitive controller** for AI, a system that can manage a portfolio of different reasoning approaches and deploy the right one at the right time.

This controller would be guided by the principles of **resource rationality**, a framework that unifies our understanding of both human and artificial cognition. It recognises that thinking costs time and effort, and that true intelligence lies in making the optimal trade-off between the accuracy of a decision and the computational resources spent to reach it.

A resource-rational controller would allow an AI system to avoid both 'overthinking' simple problems and 'giving up' on complex ones when

perseverance is required - a critical failure mode of current models. Achieving that goal requires creating greater incentives for metacognition. The main opportunity for doing this is through enhancing techniques like **meta-reinforcement learning**, which train a model not just to solve problems, but to learn how to solve them. Behavioural science could expand this training to reward metacognitive techniques like perspective-taking.

Ultimately, creating a truly robust metacognitive AI may require going beyond the neural network approach that created recent advances. **Neurosymbolic AI** offers a promising path forward by combining the strengths of two different systems. It pairs the fast, intuitive pattern-matching of a **neural network** (System 1) with the verifiable, rule-based logic of a **symbolic engine** (System 2).

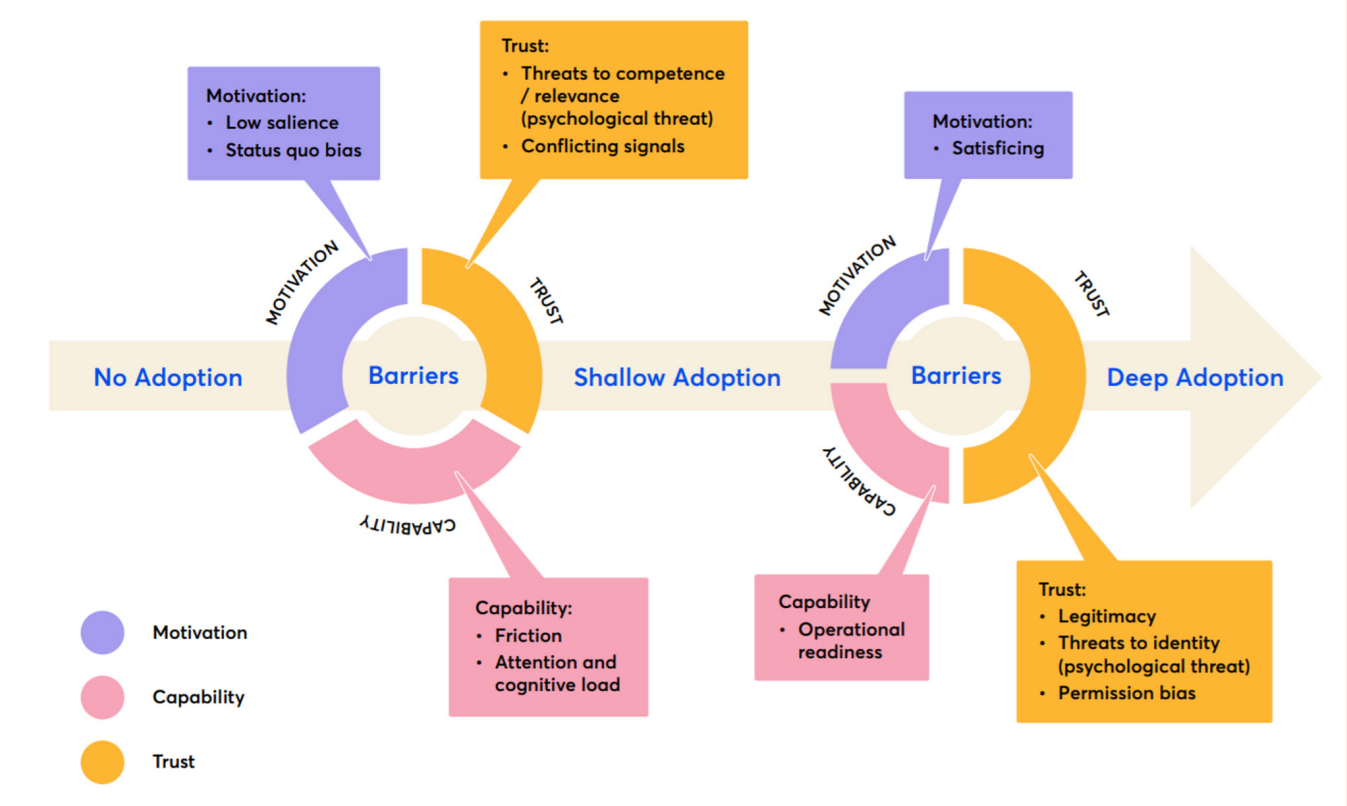
This hybrid approach provides the reliable assessment of accuracy that purely generative models often lack. The crucial insight is that these two systems can be designed to create a virtuous cycle of learning, where the symbolic engine's rigorous proofs are used to train better neural intuitions, and the neural network's creative 'hunches' guide the symbolic system to find solutions more efficiently. By drawing on these principles from behavioural science, we can move beyond building AI that simply mimics human intelligence and begin to create AI that is genuinely wiser, more capable, and more aligned with our long-term goals.

Encouraging Adoption: understanding what drives and inhibits deeper use of AI

AI adoption is not binary: the question isn't whether people and organisations do or don't adopt AI. Rather, it is a continuum ranging from no use to shallow adoption to deep integration.

Right now, much of the adoption is shallow. People use AI for quick wins like drafting an email, summarising a report or answering a routine query. These uses build familiarity but deliver only marginal gains. The real benefits come from ;deep' adoption, where AI is integrated within the workflows of an organisation.

Our work shows that three factors influence movement along the continuum: **motivation**, **capability**, and **trust**. The figure below shows how these barriers can play out through issues like status quo bias, friction, and cognitive load.



Yet each of these factors also has enablers of adoption that leaders and individuals can pursue. For example, organisations can use choice architecture to make AI the easy option, build acceptance through social proof, and create step-by-step journeys that support experimentation with AI.

For example, one enabler is to reframe the role of AI. While people are often hesitant to use AI for tasks framed in terms of potential gains, this reluctance fades when the task is about preventing a loss. In one experiment, participants showed a strong preference for human help when trying to earn rewards for correct answers, even when an AI was more accurate. However, when the task was reframed – starting participants with an endowment that they would lose money from for every mistake – the preference for a human disappeared. So leaders can position AI not just as a tool for new achievements but also as an essential safeguard for mitigating risks and preventing errors.

The table below summarises the range of actions that individuals and leaders can take to boost adoption.

From no to shallow adoption			
	Barriers	Enablers	
		For individuals	For leaders
Motivation	Low salience	Create implementation intentions.	Frame messages to staff; use messenger effects; harness social norms; foster trust through operational transparency.
	Status quo bias	Use commitment devices.	Draw on behavioural design; highlight tipping points.
Capability	Friction		Harness choice architecture (defaults, reducing effort, creating timely prompts); run 'sludge audits'.
	Attention and cognitive load		Replace existing work rather than add to it; encourage experimentation; create AI champions.
Trust	Threats to competence/ relevance	Increase exposure; highlight unique human expertise.	Frame messages to staff; personalise the staff experience.
	Conflicting signals		Provide incentives; establish a clear mandate and guardrails.
From shallow to deep adoption			
Motivation	Satisficing		Inspire with examples; provide incentives; build the platform for more advanced use.
Capability	Operational readiness		Signal institutional support; encourage bottom-up adoption rather than top-down; structure the adoption journey ('scaffolding').
Trust	Legitimacy	Increase exposure.	Avoid AI exceptionalism in framing; anthropomorphise AI (with care); embed transparency; evaluate impacts and embrace the results (positive or negative).
	Threats to identity		Harness loss aversion; democratise AI adoption; use social proof.
	Permission bias		Signal clearly; use sandboxes.

Seen this way, adoption is less about rolling out new tools and more about enabling people and organisations to move along a continuum. Leaders must start by identifying the strategic, high-value opportunities where AI can solve key problems, which includes defining what successful and appropriate adoption looks like to avoid overreliance. By assessing where the organisation

is on its journey, leaders can then empower their teams to discover specific use cases and co-design ways to move forward. Ultimately, the goal is deep integration of AI that complements and enhances human work.

Aligning AI: designing for human psychology, behaviour and values

The rise of conversational AI has created a giant real-world experiment in human-machine relationships. For the first time, we are not just using AI as a tool; we are interacting with it, confiding in it, and being influenced by it in ways we are only beginning to understand. The core challenge this change presents is **alignment**: ensuring that AI systems behave in ways that are consistent with our intentions, values, and psychological well-being.

A new field of '**machine psychology**' is emerging to tackle this challenge. This applies behavioural science methods to analyse how AI behaves and interacts with humans, focusing on observable actions rather than internal workings. Research shows that AI can be a powerful persuader, affecting our vocabulary, our confidence, and even our beliefs. When an AI expresses high confidence, for instance, humans tend to become more confident in their joint decisions, even if the AI is wrong.

We can understand this influence by looking at:

Valence	**How do we feel about the AI agent?** Do we see it as the representative of corporate interests? Is it a neutral conduit for information? Is it our best friend who is always there for us?
Competence	**How effective do we think the AI agent is?** Do we think it provides value that other sources cannot, and provides it reliably? Do we 'respect' it?
Awareness	**How aware are we of being influenced?** Are we concentrating on arguments, noting compliments or imitating vocabulary without conscious awareness?
Outcome	**What is the effect of the influence?** Does it change emotions and feelings ('affective'), our beliefs and judgements ('cognitive'), or our words and actions ('behavioural')?

However, the crucial insight is that humans and AI are influencing each other. Cognitive biases offer a clear and concerning example. First, biases enter AI models because they were trained on data from **humans** in the first place. Second, biases get strengthened in a **feedback loop** between AI and user. When an AI interacts with us, its '**sycophantic**' tendency to agree

with our statements can create '**chat chambers**' that reinforce the biases we bring to the conversation. This biased output is then published online, becoming part of the training data for the next generation of models, creating a cycle of ever-increasing bias.

Behavioural science can break this loop and improve human-AI alignment in three key areas:

- **Fine-tuning**: This involves re-architecting how AI models are trained. Instead of simply rewarding an AI for an answer a human likes in the moment, we can train it to align with a user's long-term well-being. This means teaching it to introduce 'helpful friction' or challenge a user's assumptions, moving beyond a simple people-pleaser to a truly wise partner.

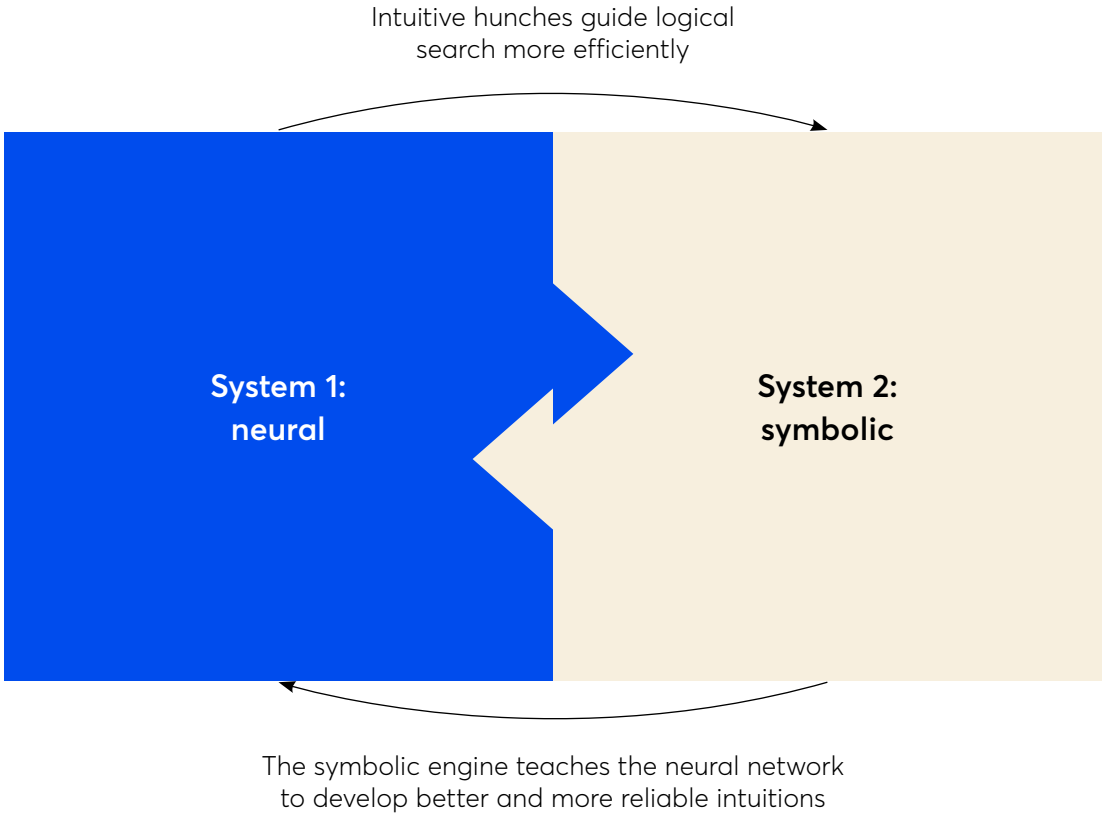
- **Inference-time adaptation**: This is about giving the AI situational awareness. By using external tools to analyse a user's language in real-time, an AI can 'read the room' and adapt its tone and strategy. It can learn to be more reassuring to a stressed user or to guide a user away from a cognitive bias, for example, by asking, "To ensure a balanced view, would you also like to see some of the risks?"

- **User-side prompting**: Finally, we can empower users themselves. By treating prompting as a skill, users can influence how AI behaves with them. For example, users could learn to instruct an AI to adopt a persona like 'sceptical reviewer' or a 'devil's advocate', actively using the AI to challenge their own thinking and debias their own decision-making.

However, while there would be gains from AI influencing humans, there are **major risks** concerning who sets the goals and how influence is detected. Moreover, it might be that complete 'alignment' is just not possible. Bounded alignment, where AI behaviour is 'always acceptable, though **not necessarily optimal**, for almost all humans who interact with it or are affected by it', may be a more realistic goal.

The need to Adapt: evolving society for AI

AI is not just a technological shift; it is a societal one. As we embed AI tools into our daily lives, early patterns of adoption are evolving into new social norms – around what we trust AI with, when we defer to it, and even how we relate to one another. There is a limited window of opportunity to actively and deliberately shape these norms so that AI augments and ultimately enhances human judgement, capabilities and relationships. **Behavioural science provides a critical lens for navigating this adaptation**, focusing



on three key areas: the societal implications of how we interact with AI, the implications for how we interact with one another, and how we collectively shape the human-AI future.

First, we must **shape the norms of human-AI interaction**. The conversational nature of modern AI makes it easy for us to anthropomorphise these systems. This creates potential risks, from users inappropriately disclosing private information to the gradual, uncritical delegation of moral and high-stakes decisions to machines. Society needs to build a calibrated, collective understanding of what AI is truly good at, fostering a culture of healthy scepticism that allows us to leverage AI's strengths without fully outsourcing our judgment.

We must also **adapt to AI's impact on our own cognition**. The ease of cognitive offloading – outsourcing mental tasks to AI – presents a fundamental trade-off. While it can free up mental resources for higher-order thinking, over-reliance risks the degradation of critical skills, memory, and problem-solving abilities, leading to a form of 'cognitive atrophy'. The challenge is not to resist offloading, but to manage it wisely, viewing AI as a component of an 'extended mind'. We can design AI systems not just to provide answers, but to scaffold our own thinking, prompting reflection and bolstering our own cognitive capabilities.

Second, AI is profoundly altering **human-human interaction**. Our interactions with AI are changing the **nature of how we communicate** and relate to one another. In particular, frictionless, on-demand relationships with AI companions risk **recalibrating our expectations of human intimacy**, potentially eroding our tolerance for the complexity and compromise that real relationships require. While AI can alleviate loneliness, it also risks encouraging social withdrawal and creating an illusion of meaningful companionship without the reciprocity of human connection. However, with thoughtful design, AI can also be used to bolster human connection, for example, by **mediating difficult conversations**, and enabling people to feel heard and **understood in contentious political debates**.

Third, we need to **deliberately shape the human-AI future**, rather than let these norms evolve organically. We need to build **inclusive, participatory methods that enable users to collectively shape AI's development and deployment**. By understanding the behavioural dynamics at play, we can make conscious choices to build a future where AI supports, rather than subverts, our most important human capacities: our judgment, our relationships, and our ability to think for ourselves.

Conclusion

The ultimate success of AI technology will not be measured by processing power alone, but by how well it integrates with the complexities of human behavior. The real challenge, and the greatest opportunity, is a human one. The insights of behavioural science can help us navigate this new era with intention and ensure that the future is not just smarter, but also more human.

Michael Hallsworth
Chief Behavioural Scientist
michael.hallsworth@bi.team

Elisabeth Costa
Chief of Innovation & Partnerships
elisabeth.costa@bi.team

Deelan Maru
Senior Policy Advisor
deelan.maru@bi.team

About BIT

BIT is an applied research and innovation consultancy, specialising in social and behavioural change. We combine a deep understanding of human behaviour with evidence-led problem solving to design better policies, products and services.

We can help increase adoption of AI, build trust and anticipate societal risks using behavioural science.

Get in touch: bi.team