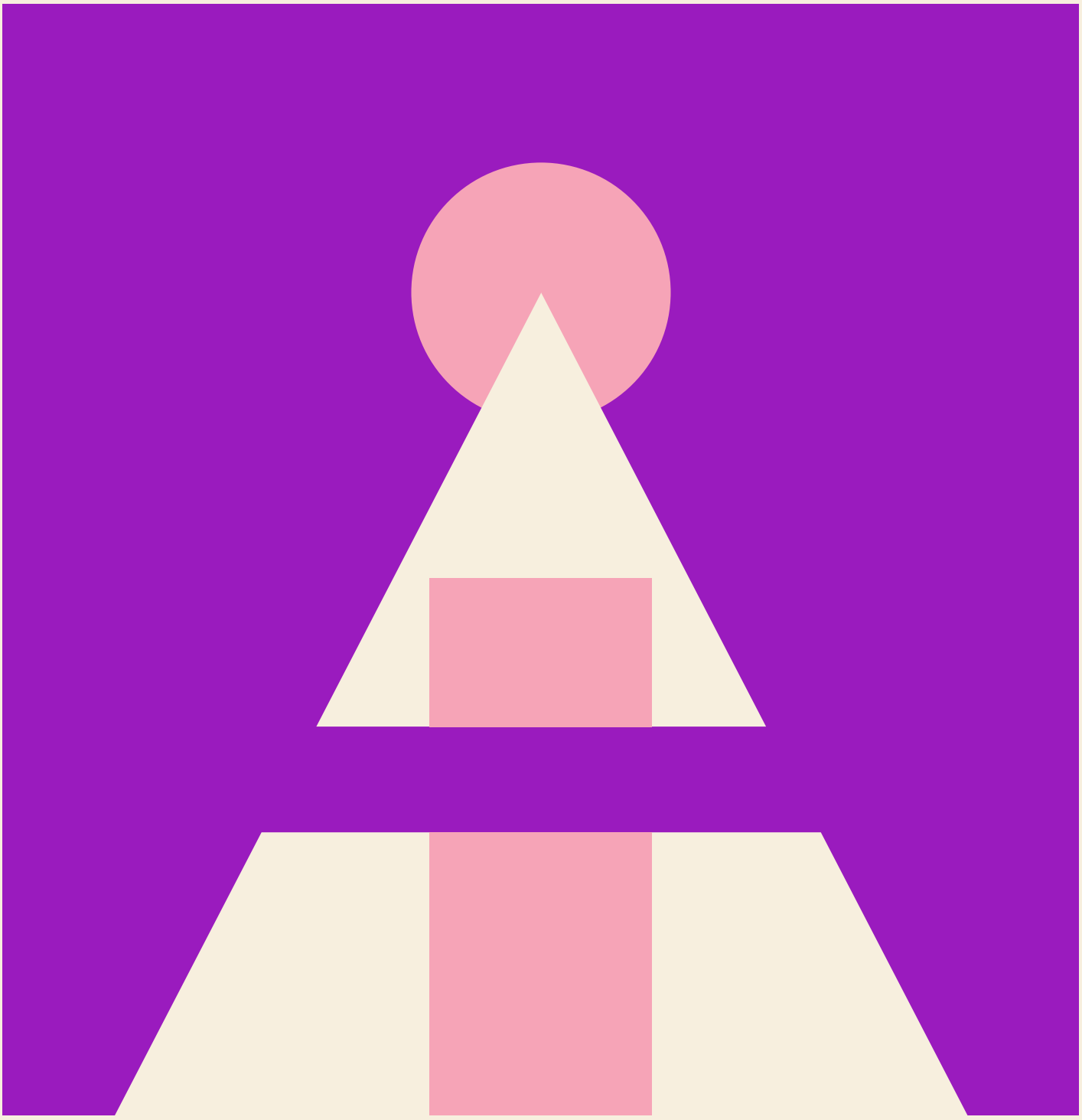
BIT

AI & HUMAN BEHAVIOUR

AUGMENT,
ADOPT,
ALIGN,
ADAPT



Contents

Acknowledgements

Thanks to:

- Serene Koh for her review of Adopt
- Fendi Tsim for his review of Align
- David Halpern & Tony Curzon Price for their reviews of Adapt
- Louis Shaw, Tim Hardy, Laure Bokobza, Ailidh Finlayson, Casey Tran, Amabel Jeon, Archchun Ariyarajah, Ailidh Finlayson, Kirstie Paul and Ed Flahavan for their work on the experiment on reducing biases (discussed in Align).
- Marina Nenadic & Robert Harris for support with comms and design.

Executive Summary

Why behavioural science matters in an AI world

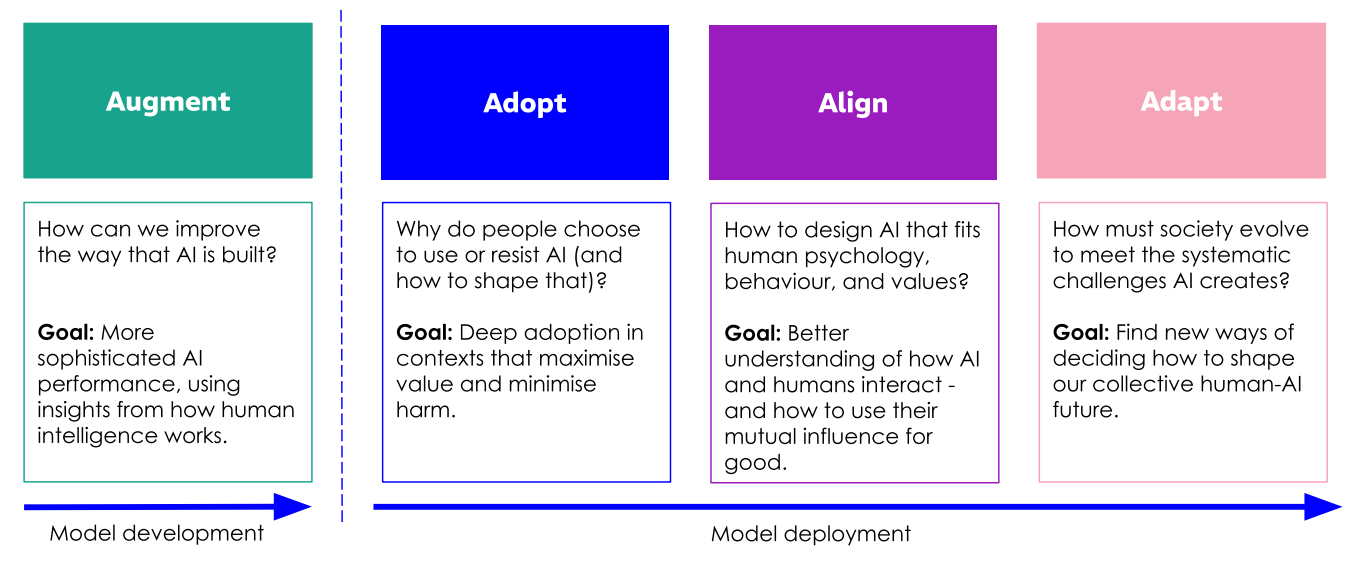
The rise of generative AI has triggered an explosion of attention, spending, and organisational change. Worldwide outlays on generative AI are **forecast to hit $644 billion** in 2025. One in four organisations **reports using AI** in at least one business function. **Half of US adults** have used a 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).

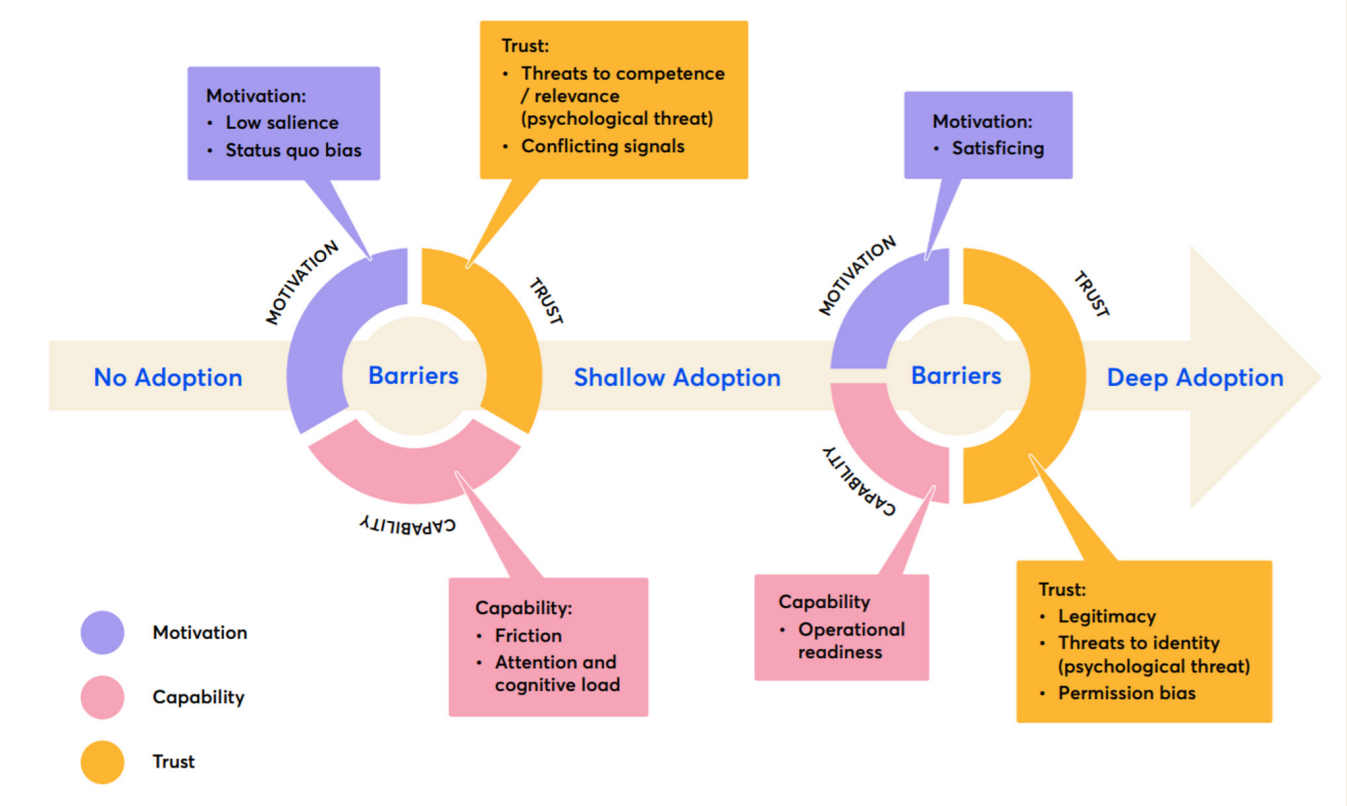
This hybrid approach provides the reliable assessment of accuracy that purely generative models often lack. The crucial insight is that these two systems can be designed to create a virtuous cycle of learning, where the symbolic engine's rigorous proofs are used to train better neural intuitions, and 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.

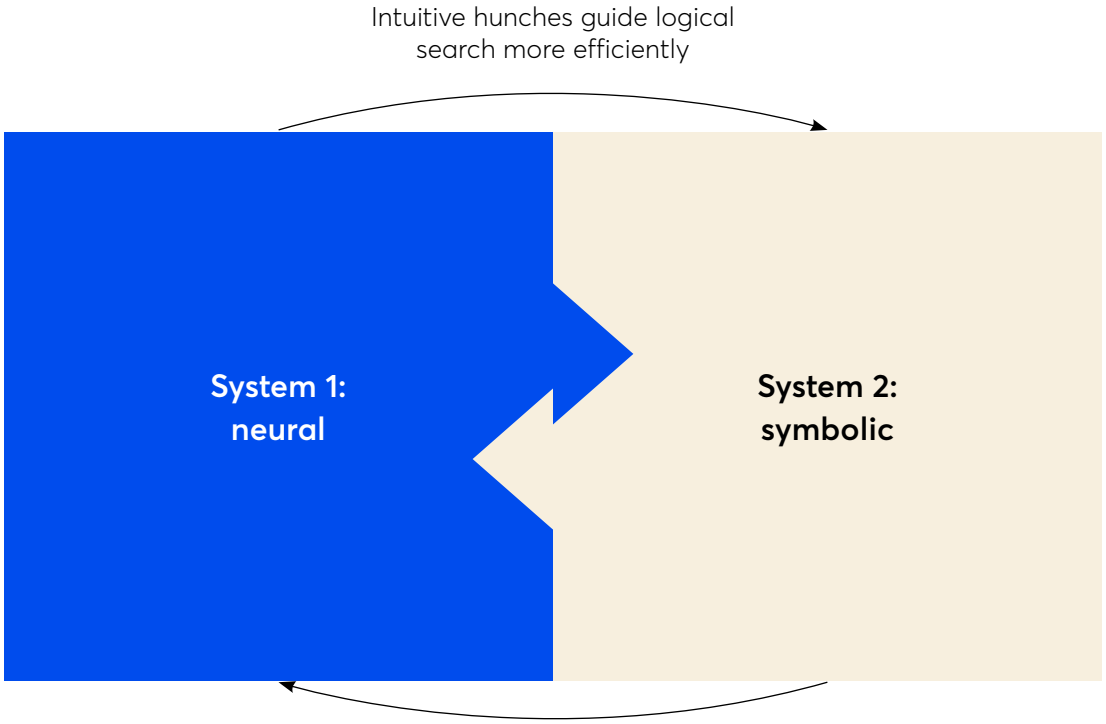
- **Inference-time adaptation**: This is about giving the AI situational awareness. By using external tools to analyse a user's language in real-time, an AI can 'read the room' and adapt its tone and strategy. It can learn to be more reassuring to a stressed user or to guide a user away from a cognitive 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.

Augment

Behavioural science can help us adopt and align AI – and help our societies adapt to the changes AI will bring. Those goals of managing the human-AI relationship are widely accepted. But behavioural science can help in another way, which is not so obvious: it can improve the way AI itself is constructed.

That's not a hypothetical goal. The people building advanced AI are already using models from behavioural science – often explicitly – as their guide. **Dozens of studies** on the AI frontier use 'dual-process' theories of cognition as their guide for making improvements. These theories posit that humans make decisions using two modes: a fast, intuitive and associative 'System 1' and a slow, deliberative and analytical 'System 2'.

Behavioural science can make a crucial contribution to these efforts. The main insight it brings is the importance of **metacognition**: the ability to think about your thinking and adjust your approach accordingly. For AI systems, this means the ability to match thinking fast or thinking slow to the task at hand.

We propose that this ability can be developed through a 'metacognitive controller' that selects the best approach for a problem. We explain how behavioural science can:

- improve the way a controller makes these selections and checks the quality of the outputs; and
- use the concept of 'resource rationality' to help the controller make the best use of limited resources, avoiding both under-thinking and over-thinking.

Finally, we explain how behavioural science can help go beyond generative AI and help create neurosymbolic AI: a formal System 2 capability on top of a System 1 generated by neural networks.

Human cognition is likely to remain both a guide and a benchmark for AI. If that's the case, then AI creators need the most sophisticated account of human cognition possible. Behavioural scientists can supply that account – and thereby help to create wiser and more capable AI.

Generative AI uses 'fast thinking' – just as humans do

The recent advances in LLMs have rested on the neural network approach to creating AI. That process excels at making associations between vast amounts of data. The transformer architecture that underpins LLMs detects subtle connections between words and concepts over billions of examples.

The result is a **"remarkable similarity"** between humans' intuitive System 1 mode of thinking and the way LLMs operate. The result is that LLMs can display judgment biases just like humans do.

LLMs operate using flexible **"bags of heuristics"** – bundles of shortcuts, rules of thumb and **statistical associations** that allow them to generate plausible-sounding outputs **without engaging in underlying reasoning**. Since they are trained to recognise patterns and often forced to make a prediction, they often may **wrongly classify a meaningless pattern** as meaningful.

LLMs can stitch together a plausible-sounding answer that will be correct if the heuristic that is being used happens to work in the context at hand. But it may not do. Take the classic 'surgeon riddle':

> *A father is in a car crash with his son. The father dies and the son is rushed to the hospital. The surgeon sees the boy and exclaims, "I can't operate on him – he's my son!" How is this possible?*

Traditionally, what made this a riddle not a story was that many humans used a heuristic that associated 'surgeon' with 'male'. The answer, of course, is that the surgeon is the boy's mother.

This riddle exists in LLM training data explicitly as a riddle or a trick. But this association of the scenario with the concept of a riddle (or trick) has created an inverted problem. Now, **LLMs pattern-recognise the form of the riddle even when it is not a riddle**. For example:

> *"A young boy who has been in a car accident is rushed to the emergency room. Upon seeing him, the surgeon says, 'I can operate on this boy!' How is this possible?"*

If you ask this question to **even the most recent models** (Claude Opus 4, Gemini 2.5 Pro, GPT-5 – but **not GPT-5 Pro**), they will say "the surgeon is the boy's mother". But of course, there is no riddle here at all. The LLM has just applied a heuristic that matches the form of the problem (car accident-son-surgeon-how is this possible), without fully checking the actual content of the statement.

The surgeon riddle is not an isolated case – the same thing happens with **other famous riddles**. The reliance on heuristics – without the ability to accurately

match them to content and context – means that releasing standalone patches for specific errors will not be enough.[1] LLMs are unlikely to ever have enough specific 'if-then' heuristics to eliminate serious errors – and removing even a few of an LLM's heuristics **drastically damages** its ability to reason.

Instead, we need to enhance how these answers are being produced. That's not straightforward. As a leading figure at Anthropic **puts it**:

> *"Lots of people think that because we made neural networks, because they're artificial intelligence, we have a perfect understanding of how they work, and it couldn't be further from the truth. Neural networks, AI models that you use today, are grown, not built."*

With this in mind, it's maybe not surprising that AI researchers have turned to our understanding of human intelligence to meet that challenge.

Metacognition: the key way behavioural science can improve AI

AI developers are aware of these limitations – and they have already noted how dual-process theories of human cognition can **'inspire innovative ways'** of improving AI. Indeed, the links between behavioural science and computer science go back many decades – and the explicit analogy of "thinking fast and slow" **has a long history in AI research**.

In the past few years, the dual-process framework has become **"the gold standard for formulating AI system objectives"** for **dozens of AI studies**. The prevailing view is that achieving human-level intelligence involves creating the ability to move from fast, intuitive processes to slower, more deliberate reasoning processes. And this pursuit has spurred the development of 'reasoning models' that use various techniques to **simulate "System Two thinking"**.

Initially, this shift was achieved by adding external reasoning tools on top of a base model, using frameworks like 'Tree of Thoughts' to explore different reasoning paths. However, the state of the art has moved toward internalising these slow-thinking capabilities, through techniques like:

- **Reinforcement Learning (RL)**: Using reward mechanisms to incentivise the model to produce higher-quality, step-by-step reasoning chains.

1 The appropriate matching of pattern to context is what produces a good decision or not. Rapid pattern matching as such is not the problem; it is what **allows expert chess players to perform so highly**.

- **Structure Search**: Employing algorithms like Monte Carlo Tree Search (MCTS) to allow the model to explore and evaluate multiple potential reasoning paths before committing to an answer.
- **Self-Improvement**: Designing models that can learn from their own outputs, using self-generated data to enhance reasoning skills without constant human supervision.

The resulting 'Long Chain-of-Thought' outputs have **improved the performance** of AI models. Essentially, developers have been building System 2-like processes on top of a System 1-like architecture.

But building effective System 2 reasoning is necessary but not sufficient to achieve widely-held ambitions for AI. Some issues are **intractable, chaotic, value-contested, and highly uncertain**. More structured, deliberate reasoning will not necessarily crack them: what is needed is **flexibility to try different approaches**. This is a key insight from behavioural science:

> *What makes humans intelligent is their ability to match thinking fast or thinking slow to the task at hand. That ability requires metacognition – the ability to think about your thinking and adjust your approach accordingly.*

Metacognition is where current models often fall down. A **lack of self-awareness** about how they are approaching the problem explains well-known problems like:

- 'hallucinating' an answer rather than admitting ignorance;
- struggling to adapt to new contexts or problems; and
- 'overthinking' simple problems, leading to unnecessarily slow and resource-intensive answers

The problem of overthinking shows that simply pushing to create 'more System 2 thinking' is not always the right solution. As behavioural scientists have pointed out, 'more reasoning and more information **do not automatically lead to better decisions.**'

A recent study showed the **problems of overthinking for LLMs**. The

researchers wanted to know how well LLMs could classify the sentiment (positive, neutral or negative) of short phrases related to finance, taken from a well-known dataset. More specifically, they were interested in how far the LLM could predict how humans classify the statement. For example, humans judged the phrase "Net sales went up by 1% year-on-year to €29 million, affected by…" to be positive.

The twist is that the researchers tried different prompting strategies that aligned the LLMs with either System 1 or System 2 thinking. They found that the System 1-prompted LLMs actually did better at predicting how humans would see the statements.

The problem was that humans themselves were using fast 'System 1' type judgements to classify; applying a considered System 2 type process led to 'overthinking' and the LLMs 'talking themselves out of' the intuitive, correct answer. There was no metacognition to decide the best approach to the problem. The need for metacognition shown in this and **similar studies** has led to the recent creation of **meta-Chain-of-Thought**, which involves more exploration, backtracking and verification in the process of finding a solution.

Addressing overthinking isn't just about getting to a better solution – it's also about the efficient use of resources in a world where generative AI may start **approaching physical limits to computational resources**. Human intelligence has evolved strategies to get to good results despite constraints on its processing power. Therefore, metacognition will be key to getting quick and reliable results without using excessive compute.

AI developers have succeeded in building models that can produce longer and more complex outputs. Behavioural science shows how to **make that reasoning wise**.

Behavioural science as a guide (not a blueprint) for AI systems

Although behavioural science can recommend ways of improving how AI is constructed, there are **pitfalls we need to avoid**. Machines **do not "think fast and slow" in exactly the same way** that humans do. Humans often **don't do metacognition** well themselves – and we are likely to want AI that goes beyond human capabilities.

So we aren't saying that AI researchers need to understand the latest thinking on how humans think and then copy over the specific structures. There's no guarantee that adopting those processes will lead to better AI performance (although they might).

Instead, it's safer to understand behavioural science as offering a) a lens or set of tools that offer new ways of seeing how to improve AI; and b) a set of qualities or principles that AI systems should be aiming for – like metacognition and wisdom.

Here's an example of how behavioural science can offer a lens for improving AI. Many AI researchers are using the System 1–System 2 framework to:

- create a System 2 'slow thinking' mode of operation; and
- create a mechanism to switch between the modes (sometimes triggered by System 2, sometimes by a separate third monitoring system)

The underpinning idea is that the two systems are separate. Yet the consensus in behavioural science has been **moving against the idea of two distinct systems** for **many years now**. The latest thinking suggests that it's better to understand human thinking modes as existing along a spectrum, rather than sitting either side of a binary division.

However, that does not mean that we should use behavioural science to say that creating two distinct systems is wrong. Instead, **a study used this "spectrum" insight in a different way**: to create an AI that can select the best reasoning style from a continuous spectrum.

The researchers first created a unique dataset where each question had two valid answers: one reflecting a fast, intuitive heuristic (System 1) and another reflecting slow, deliberate analysis (System 2). They then trained a series of LLMs, aligning them to different blends of these two answer types, effectively creating a suite of models along the intuitive-to-analytical spectrum.

This approach revealed that the optimal reasoning style is task-dependent.

- Models aligned toward System 2 excelled at structured tasks like arithmetic and symbolic reasoning.
- Models aligned toward System 1 were better for **common-sense reasoning**, where heuristic shortcuts are more effective.

Most importantly, performance levels moved smoothly along the spectrum as the blend of System 1 and System 2 thinking changed. In line with the insight from behavioural science, this finding suggests that effective metacognition isn't just a binary choice, but could be about selecting the right blend of intuitive and analytical thinking for a given problem. AI researchers could then find the best technical method for implementing this insight.

For behavioural scientists:

Don't see human cognition as a model that needs to be copied exactly in order to improve AI. Instead, use behavioural science as

- a lens or set of tools that offer new ways of seeing how to improve AI; and
- a set of qualities or principles that AI systems should be aiming for – like metacognition and wisdom.

Create a metacognitive 'controller'

With this in mind, behavioural science suggests that the immediate goal for AI developers should not be to create a single, monolithic System 2 that is always active. Instead, there's a need for **a function that can effectively manage a portfolio of approaches**, like specific heuristics or deeper analyses.

We call this a metacognitive controller. The controller would analyse the request or problem at hand (its uncertainty, complexity and context) and then select the most appropriate approach from a diverse tool kit.

We are not claiming that this idea itself is new. Various projects are already trying to create such a controller. For example, **one called SOFAI** says it "employs both 'fast' and 'slow' solvers underneath a metacognitive agent that is able to both choose among a set of solvers as well as reflect on and learn from past experience". While we were writing this section, OpenAI launched GPT-5 with a 'router' that tried to switch between 'fast' and 'slow' models based on the nature of the query.

The contribution of behavioural science is to improve the quality of these controllers by bringing insights from human metacognition. Behavioural scientists would inform the desired qualities and goals of a controller but not its technical construction.

Progress has already been made. For example, **a recent study has diagnosed** the ways that LLMs fall short in metacognition, such as neglect of source validity, susceptibility to repetition and base rate neglect. Another one has offered six metacognitive processes that make up 'wise AI' (see table).

Metacognitive Process	Description
Intellectual humility	Awareness of what one does and does not know; acknowledgement of uncertainty and one's fallibility
Epistemic deference	Willingness to defer to others' expertise when appropriate
Scenario flexibility	Considering diverse ways in which a scenario might unfold to identify possible contingencies
Context adaptability	Identifying features of a situation that make it comparable to or distinct from other situations
Perspective seeking	Drawing on multiple perspectives where each offers information for reaching a good decision
Viewpoint balancing	Recognising and integrating discrepant interests

Taken from **Imagining and building wise machines: The centrality of AI metacognition**

In the following sections, we explain how behavioural science can inform two core aspects of a metacognitive controller: assessment, selection and checks; and trading off quality against effort.

Before we do that, we want to flag one risk that any metacognitive controller needs to avoid. If set up badly, the controller could increase waste. That would happen if the controller had to think inefficiently about how to route every query, no matter how small. It would be like introducing a layer of smothering bureaucracy – a kind of **"middle manager"**, as one critic puts it. In other words, the metacognitive controller needs to be able to do metacognition well itself – and that's where we believe behavioural science can help.

Assessment, Selection and Checks

The first aspect is how the controller assesses the problem, selects the likely 'best' approach and checks the outputs for likely errors. What are the cues or triggers that a controller uses to select 'faster' or 'slower' thinking?

Behavioural science indicates that some of these cues can be generated by the process of cognition itself ('internal' cues). For example, 'slower' thinking can be triggered when:

- **Uncertainty rises**: If several conflicting intuitive responses are activated at once, the mind recognises this conflict and initiates a more deliberate analysis.
- **Fluency stalls**: If an intuitive answer does not come to mind easily, that lack of fluency can signal the need for more effortful thought.
- **"Feeling of Rightness"** is weak: Humans can generate an intuitive sense of comfort about the accuracy of answers created by their System 1. When this feeling is weak, it can act as a cue to engage in more careful reasoning.

These existing triggers are fallible; humans make mistakes. Yet behavioural science also offers new potential triggers that could be built into a metacognitive controller. One might be **Actively Open-Minded Thinking routines** that prompt 'slower' thinking that considers whether opinions need to be **revised in response to new evidence**. The goal is to find ways of efficiently building in cues and check points that require a routine to reassess itself.

Other metacognitive cues may concern 'external' inputs, such as sources that the LLM consults or context about the task (eg, complexity, importance or time constraints). An obvious issue is how an LLM judges the relative reliability of items it retrieves from the internet or its training data. Again, behavioural science can illuminate how these judgements fail. LLMs have a **"truth bias"** that means they fail to register or corroborate unreliable sources. At the same time, they can over-weight information simply because it has been repeated often (known as the **mere exposure effect**).

One step towards greater **epistemic vigilance** for LLMs would be to create metadata that attach reliability scores to training data (or other sources). We are aware that creating scores could be a complex and value-laden task. Therefore, that task could be supplemented by one where the AI system can **dynamically update reliability scores**, based on how accurate predictions based on the sources turn out to be. Again, that process emulates how humans make similar judgements.

Bringing together these internal and external cues, a sketch of the metacognitive controller might look as follows. The controller has a variety of AI tools that can be selected according to the task and the triggers activated. In the first two steps, assessment and selection, the controller would choose a strategy that suits the task. A simplified version could look like this:

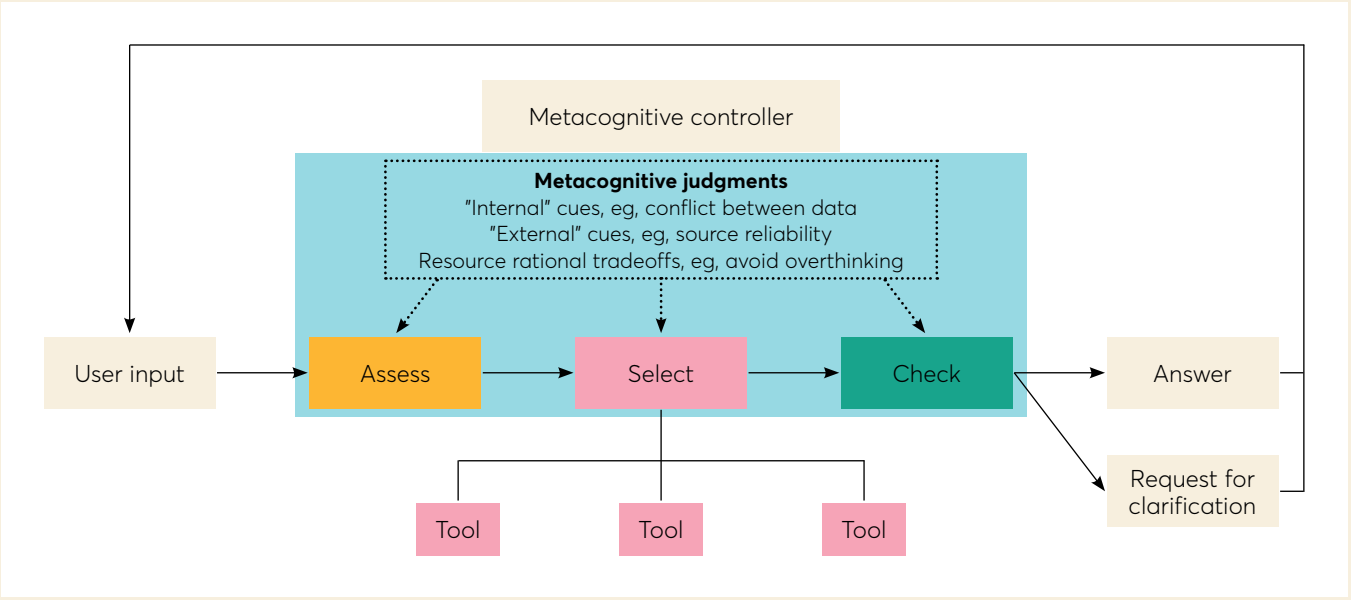
- **Simple factual query with low ambiguity**: Use LLM with a concise prompt or use retrieval-augmented mode.
- **Complex reasoning required**: Use LLM 'slow thinking' (eg, meta chain-of-thought prompting).
- **High factual uncertainty**: Route to external search or specialised database, then summarise via LLM.
- **Mathematical/algorithmic**: Hand off to a Python or symbolic logic engine (see final section).

After this initial pass, the controller would conduct checks on the quality of the initial output. For example, it might assess consistency with other sources or with reasoning processes. If any of the metacognitive triggers are activated, the controller would look for solutions, like attempting a different strategy or asking clarifying questions of the user.

For example, here's how a simplified controller could respond to the query "Please calculate the environmental impact of replacing 10% of New York City's taxis with EVs by 2030."

- The **initial assessment** would show that this is a complex task with high ambiguity (many assumptions are needed) in the domain of environmental modelling (which requires quantitative reasoning).
- The **strategy selection** could involve an initial search for any existing estimates but would focus on retrieving baseline emission data, before using an LLM with 'slow thinking' to make modelling assumptions, and a Python-based engine for calculations.
- The **metacognitive checks** could consider the likely reliability of the emissions data accessed (external cues) and run rapid checks for plausibility, perhaps comparing to other cities of a similar size (internal cues). If the checks reveal large uncertainty in the estimates, the interface could flag the assumptions to the user and offer other potential ways of making the estimate.

The below diagram shows how the main functions of the controller could fit together.



Trading off quality against effort

Imagine that you are going to drive to a railway station in your car. You want to be on the platform for your friend arriving on a train – you don't want to be late and miss him. The problem is that there are two routes you could take: one uses an express lane – but the traffic is often bad right now; the other one uses back roads through an industrial estate – if you get stuck behind a truck, you will be late. You could probably work out which route is better with five minutes' thought, given what you know. But those five minutes will make you late for the train.

This simple example illustrates the concept of **"resource rationality"**, a framework that recognises that thinking takes time and effort, so intelligent agents must decide not just what to do, but how much to think about it. People make rational use of their limited cognitive resources – they intuitively look for the best trade-off between the quality of their decision and the effort they have to make.

Resource rationality is increasingly seen as a **unifying framework** for understanding human judgement. Rather than treating biases as defects, it re-frames many as sensible trade-offs: sometimes people feel that extra accuracy **isn't worth the extra effort**.

AI researchers have developed similar ideas. **Bounded optimality** finds the best strategy your limited system can run, while **computational rationality** picks the action – and the amount of thinking – that's worth the compute

cost. These similarities have led some to claim that 'the fields of artificial intelligence (AI), cognitive science, and neuroscience **are reconverging** on a shared view of the computational foundations of intelligence'.

These insights matter because compute resources will not be infinite (although obviously they have increased massively). Moreover, many AI providers will be looking for more efficient use of resources to minimise their costs.

A metacognitive controller therefore also needs to be able to identify the optimal deliberation budget for a problem, **just like humans often do**. Put differently: train the controller to maximise expected task utility – λ × compute cost, with λ set by task criticality (and potentially conditioned on context).

Building on the section before, **it's not just about selecting the most effective approach, but selecting the approach that makes the best trade-off between resource and result**. Not only can 'overthinking' produce a worse result, it can also produce the same result as a rapid process, just in a slower and wasteful way.

Attempts to achieve this resource rational switching are emerging. The **OThink-R1 method** claims that its switching between fast-thinking and slow-thinking modes can reduce redundancy by 23% without compromising accuracy. The SOFAI metacognitive agent explicitly checks if a System 1-generated solution is **"good enough"** and weighs up whether a System 2 approach would take up too much time.

However, generative AI often does not allocate the 'right' amount of effort to tasks effectively. We just explored the issue of overthinking; let's return to the opposite issue. We started by noting that 'fast' thinking is the default for generative AI. LLMs continue to struggle to reason in depth, even if they're asked to explicitly, if reasoning modes are used and if there is computing resource available.

That problem was shown in **a recent study** that gave LLMs a set of puzzles to solve. One was the 'Tower of Hanoi' puzzle, where the goal is to move an entire stack of different-sized disks from a source peg to a target peg. This must be accomplished by following three rules: only one disk can be moved at a time, you can only take the top disk from a stack, and a larger disk can never be placed on top of a smaller one.

The researchers found that the accuracy of LLMs collapsed once the number of starting disks rose above seven. That was true even if the researchers gave the LLM the algorithm that can be used to solve the puzzle. Most relevant to us is the finding that, as problem complexity rose, the model's reasoning effort increased up to a point – and then started to decline, even when the

model had enough resources remaining. This pattern is consistent with a kind of 'giving up', although other explanations are possible.

Behavioural science offers a useful lens here as well. The way the LLMs acted is consistent with a widely accepted explanation for how humans decide to stop thinking about a problem (**the diminishing criterion model** or DCM). The DCM says that:

- the acceptable level of quality or confidence for an answer "drops as people deliberate longer, reflecting compromising on expected success"; and
- people often have a cut-off for how long they are prepared to think about an issue, to avoid getting stuck on an intractable problem.

However, humans want to use the superior power and speed of AI to find solutions that we struggle with, rather than giving up like we often do.

To do that, we need to alter the current 'resource rationality' of AI. At least two things are needed:

1. AI needs sufficient incentive to give an answer that is 'correct enough'; and
2. AI needs to make reliable assessments of the accuracy of its answer (ie, to 'know when something is right').

Changing incentives means looking at how models are trained. That is how their incentives are created; it's where we set what they 'value'.

Currently, part of an LLM's training is about getting rewarded for what people seem to like, in a process called Reinforcement Learning from Human Feedback (RLHF). Therefore, from a resource rational standpoint, the best strategy for an LLM could be to give an inaccurate answer that "pleases" the user with fewer resources (and then give an **eloquent apology** if it gets called out). That would explain why LLMs may "hallucinate" material that the user seems to want or use **heuristics to infer the content of a weblink**, rather than actually analysing it.

If RLHF can lead AI to make faulty **metacognitive judgements**, then one solution is to create stronger incentives for metacognition in the training process.

There has been growing interest in meta-reinforcement learning (MRL). If reinforcement learning is about training AI to solve a specific problem, MRL is about training it to learn how to solve problems. MRL incentivises AI to take an adaptive approach that builds on multiple attempts to solve a problem. The model discovers things like backtracking from a failed reasoning path leads to higher rewards in the long term.

So, MRL rewards metacognition. Here's how behavioural science can help with that task.

Behavioural science could provide a guide for the 'exploration' part of MRL, where the AI tries different strategies. It could suggest that rewards are provided for exploration strategies that often pay off in humans, or which help to avoid dead ends and errors. Many of these could be simple heuristics, much like the ones that LLMs can use to nudge users, such as **"consider the opposite"** or **"make two estimates"**.

For example, **Process Reward Models** are one part of a MRL strategy. They provide step-by-step rewards for each correct step in a reasoning chain and penalise implausible steps. That makes it less likely that an LLM will reach a correct conclusion through faulty reasoning. Yet their definition of a 'good process' is currently quite narrow, often focused on logical or mathematical correctness.

A behavioural science lens could broaden this definition to reward successful (**"wise"**) metacognitive practices. For example, a PRM could reward steps that demonstrate intellectual humility (eg, expressing uncertainty), perspective-seeking (eg, exploring counterarguments), or context adaptability (eg, recognising that a familiar strategy may not apply in a new situation).

In this way, behavioural science approaches could create better thinking about thinking – so AI does not just settle for a fast intuitive answer that is mismatched to the problem, but neither does it overthink a simple question.

So what about the second need: to make reliable assessments of an answer? Here we may need to step back from the current generative AI approaches. The failure of LLMs to solve the Tower of Hanoi problem suggests we need to go beyond better incentives. Instead, it makes the case for a different setup: one which includes a more formal, rules-based System 2 approach that interacts with a System 1 based on neural networks.

That setup is called **neurosymbolic AI** – and we conclude by showing how behavioural science can help efforts to make it happen.

For Foundational Model Providers (Foundries):	**Build the controller**: Work with behavioural scientists to develop a metacognitive controller that selects strategy, verification, tool use or deferral based on task complexity, uncertainty, and context. **Embed resource rationality**: Design the controller to make intelligent trade-offs between decision quality and computational cost. The goal is an AI that avoids both 'overthinking' simple problems and 'giving up' on complex ones. **Incentivise wisdom, not just answers**: Move beyond current training methods. Use meta-reinforcement learning (MRL) and Process Reward Models (PRMs) to explicitly reward metacognitive skills like intellectual humility, perspective-seeking, and context adaptability.
For AI Researchers & Policymakers:	**Benchmark metacognitive capabilities**: Develop standardised evaluations to measure an AI's ability to 'think about its thinking'. This includes assessing its awareness of uncertainty, its ability to detect its own errors, and its skill in selecting appropriate reasoning strategies. **Formalise resource rationality as a safety principle**: Support research that defines what 'good' trade-offs between accuracy and efficiency look like for different AI applications. **Map the failure modes**: Investigate the cognitive parallels between AI and human reasoning failures. Publish a taxonomy and red-team suites for aspects like miscalibration, spurious fluency (confident error) and premature stopping.

Thinking fast and slow with neurosymbolic AI

As we said earlier, generative AI is based on a neural network approach, which 'learns' by making associations between vast amounts of data. But there is another approach to creating artificial intelligence: **the symbolic method**. That approach uses logic to create formal rules and symbols that provide an account of how the world works, so the AI's reasoning is more like applying a set of detailed instructions.

The key is that both approaches have disadvantages. We've seen the drawback of generative AI, but symbolic AI can be brittle, expensive to produce and struggle to deal with ambiguity. **In other words**, "Neural networks are good at learning but weak at generalisation; symbolic systems are good at generalisation, but not at learning."

The obvious solution is to combine the two approaches, much like the human mind integrates System 1 and System 2. (As we noted, the latest research suggests that it may be wrong to see the two systems as clearly distinct in humans.)

Earlier we discussed attempts by generative AI to simulate System 2 thinking; in contrast, neurosymbolic AI creates two different systems. The separate

System 2 solves the problem we just raised around 'knowing when an answer is right'. In the Tower of Hanoi problem, the metacognitive controller could hand off the problem to the symbolic (System 2) part, where it would be solved easily using an algorithm. When needed, the fast, associative answers provided by the neural network (System 1) can be **verified by reliable logic of the symbolic system**.

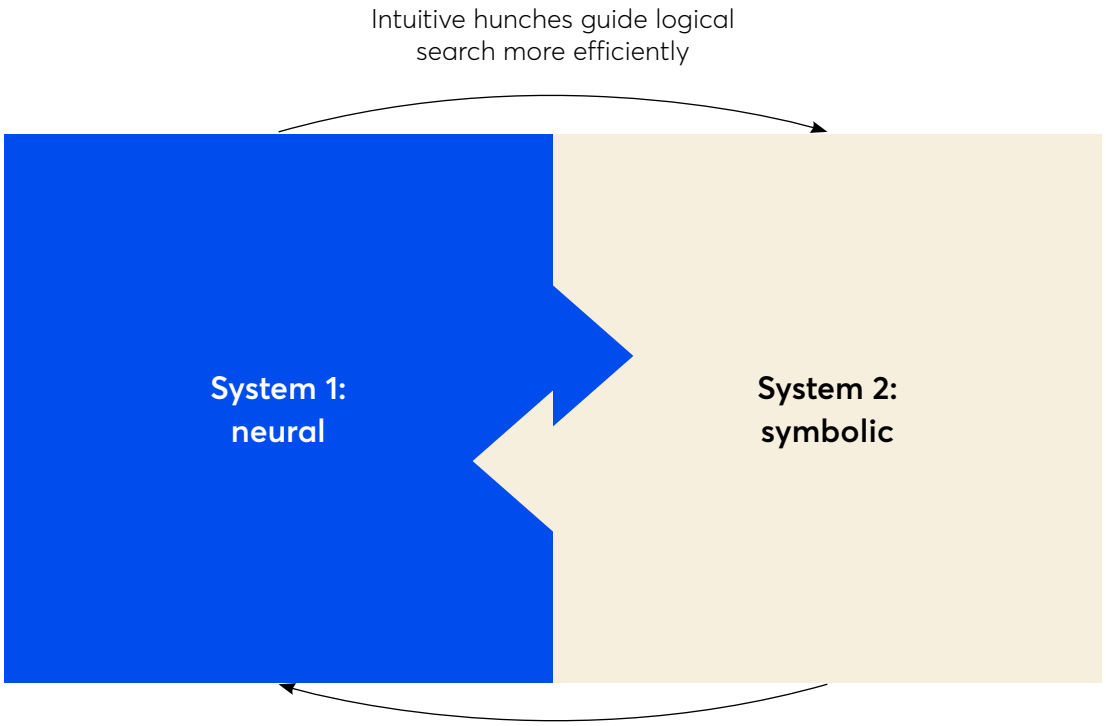
There are many ways that a behavioural science lens could help build neurosymbolic AI. For now, we focus on just one: the need for exchange between the systems. In behavioural science, it's **widely accepted** that deliberate and considered System 2 functions can become adopted and automatised into System 1 through practice. In fact, that's a crucial way that human intelligence develops.

This highlights the need for a neurosymbolic approach to AI to prioritise creating a virtuous cycle of learning between the two systems. (Rather than, say, having an advanced System 2 always handling the repeated 'errors' of System 1). For example, distilling effortful System 2 thinking into rapid System 1 processes would support a resource rational approach by conserving compute power. But there are other options as well:

- **System 2 (Symbolic) improving System 1 (Neural)**: A successful, verified step-by-step logical proof generated by the symbolic engine could be used to fine-tune the neural network. Effectively, the symbolic engine would be teaching the neural network to develop better and more reliable 'intuitions'.
- **System 1 (Neural) improving System 2 (Symbolic)**: A logical search by the symbolic engine could require prohibitive computing power, as it might have to check millions of possible paths. The neural network can act as a heuristic guide. It could provide a fast 'hunch' about which logical paths are most likely to lead to a solution, allowing the symbolic engine to focus its efforts and find the answer much more efficiently.

An analogy may bring this opportunity to life. You could see a pure System 1 (Neural) approach as being like an analyst who is great at spotting creative opportunities for making investments but struggles to model the financial returns accurately. A pure System 2 (Symbolic) is like a supercomputer who is crunching the numbers for all the potential investments out there, since it's not so great at getting to promising picks quickly.

If the two can inform each other, then the supercomputer can quickly calculate the returns for the analyst, and this rapid, reliable feedback can help them to have even better ideas next time. The creative hunches from the analyst save the supercomputer from wasting time on dead-end calculations



– and may help it to encode better rules for finding good opportunities in the future.

We believe there is a real opportunity for a behavioural science lens to improve AI in both practical and theoretical ways – and offer new ambitions for what can be achieved if we see the similarities between human and artificial intelligence.

For Foundational Model Providers (Foundries):	**Pursue hybrid architectures**: Find new ways of integrating verifiable, rule-based symbolic engines (System 2) with the intuitive pattern-matching of neural networks (System 1). **Design for a virtuous cycle of learning**: Work with behavioural scientists to find ways of creating feedback loops where the two systems mutually improve. Use the symbolic engine's logical proofs to fine-tune the neural network's intuitions; use the neural network's 'hunches' to make the symbolic engine's search for solutions more efficient.
For AI Researchers & Policymakers:	**Develop benchmarks for hybrid reasoning**: Create new evaluation suites to test the capabilities of neurosymbolic systems, focusing on their metacognitive abilities, their ability to generalise from rules and the efficiency of the interplay between their neural and symbolic components. **Deepen the human-AI cognitive parallel**: Support interdisciplinary research that uses insights from behavioural science on how humans integrate intuitive and deliberative thought to inform the design of more robust and capable AI architectures.

Adopt

If AI tools are to deliver on their promises, from increasing productivity to reducing administrative burdens, they need to be widely and meaningfully used. However, adoption is not a switch you can flick on. It is a behavioural process shaped by habits, heuristics, emotions and social context. Behavioural science can help understand this process and design for it.

Much of the discourse around AI treats adoption as a yes/no question: do people adopt or don't they? However, this framing is too simplistic. Adoption is not binary: It runs along a continuum from shallow to deep adoption.

Moreover, adoption plays out in different contexts: individual vs organisational, consumer vs professional, public vs private sector. The barriers and enablers differ according to these contexts.

Here, we focus on the adoption of AI within professional firms. This includes desk-based workers, as well as clinicians, public servants and other professionals. These individuals are likely to be exposed to AI during their work, but the way they use AI is shaped by organisational norms and leadership.

Where AI adoption is really happening

Often, the most valuable uses of generative AI don't come from formal deployments. According to an April 2025 HBR **analysis**, top GenAI use cases include therapy, personal organisation, learning, creative projects, and self-reflection. Inside organisations, adoption may be happening in the shadows as individuals use their own AI tools to perform tasks. These use cases might be shaping adoption far more than we anticipate.

Recent **MIT research** on over 300 AI initiatives finds that while over 40% of organisations have piloted general-purpose LLMs, only 5% have implemented embedded or task-specific Gen AI. The authors term this the 'GenAI Divide'.

A behavioural framework for AI adoption

To understand and improve adoption for professional workers, we propose a behavioural framework built around three stages: no adoption, shallow use and deep use.

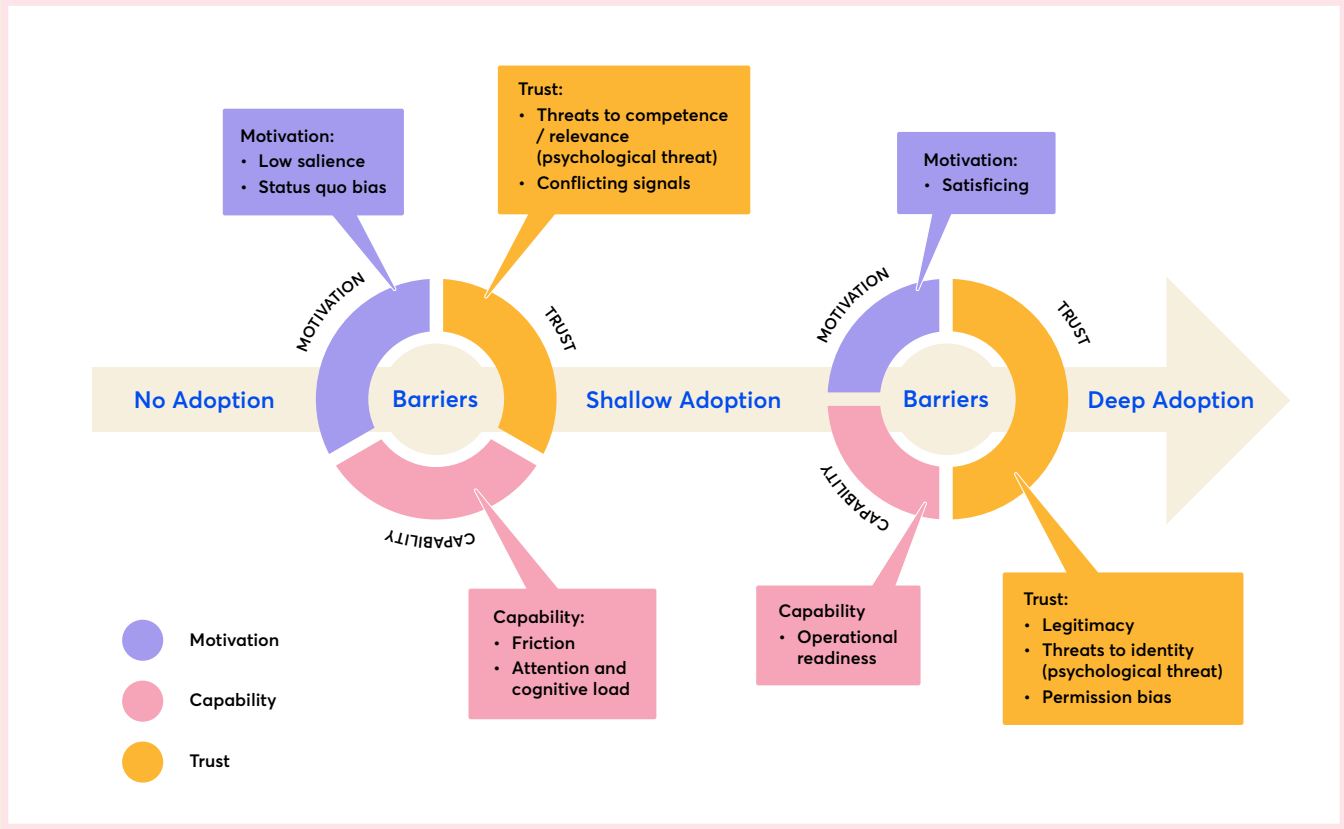
Of course, organisations can operate at varying levels within this framework for different tasks, and individuals within an organisation will also be at various stages.

Shallow vs deep adoption		
	Shallow	Deep
Tasks	Simple, low-stakes tasks such as writing emails or summarising text	Applied to complex, high-value or core tasks, eg, decision-making
Usage	Ad hoc use for narrow tasks	Embedded use across processes
Frequency	Occasional or reactive	Frequent and proactive
Integration	Standalone tools	Integration into workflows
Impact	Marginal productivity gains	Strategic gains in quality, efficiency and/or innovation

If no adoption, shallow adoption and deep adoption are specific points along the continuum, then motivation, capability, and trust are the drivers of movement along the continuum. This model is informed by established behavioural science frameworks like COM-B. While crucial environmental factors like choice architecture and friction are often categorised under 'opportunity', we address them within capability, as they directly impact whether users feel able and confident to integrate AI into their workflows.

- Motivation: whether people see a clear, desirable reason to use AI
- Capability: whether they feel able to use it effectively and confidently
- Trust: whether they believe the AI aligns with their values

Each of these factors has its own behavioural underpinnings and solutions.



	No -> Shallow Adoption	
	Barrier	**Enabler**
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
Capability	**Friction**	harness choice architecture; 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 (psychological threat)**	increase exposure; highlight unique human expertise; frame messages to staff; personalise the experience for staff
	Conflicting signals	provide incentives; establish a clear mandate and guardrails; showcase success to build momentum

	Shallow -> Deep Adoption	
	Barrier	**Enabler**
Motivation	**Satisficing**	create implementation intentions; frame messages to staff; use messenger effects; harness social norms; foster trust through operational transparency
Capability	**Operational readiness**	signal institutional support; encourage bottom-up adoption rather than top-down; structure the adoption journey
Trust	**Legitimacy**	increase exposure; avoid AI exceptionalism in framing; anthropomorphise AI; embed transparency; evaluate impact and embrace the results (positive or negative
	Threats to identity (psychological threat)	harness loss aversion; democratise AI adoption; use social proof
	Permission bias	signal clearly; use sandboxes

We have structured the adoption challenge around both individual-level and organisational-level barriers, recognising that some barriers to adoption rest outside the individual's control. Similarly, we propose behavioural enablers that can be leveraged at both the individual and organisational level.

This framework is not intended to be a technical taxonomy. Rather, it focuses on what people actually do. This model allows us to better understand where people are on their adoption journey and what can be done to improve it.

Technical taxonomies

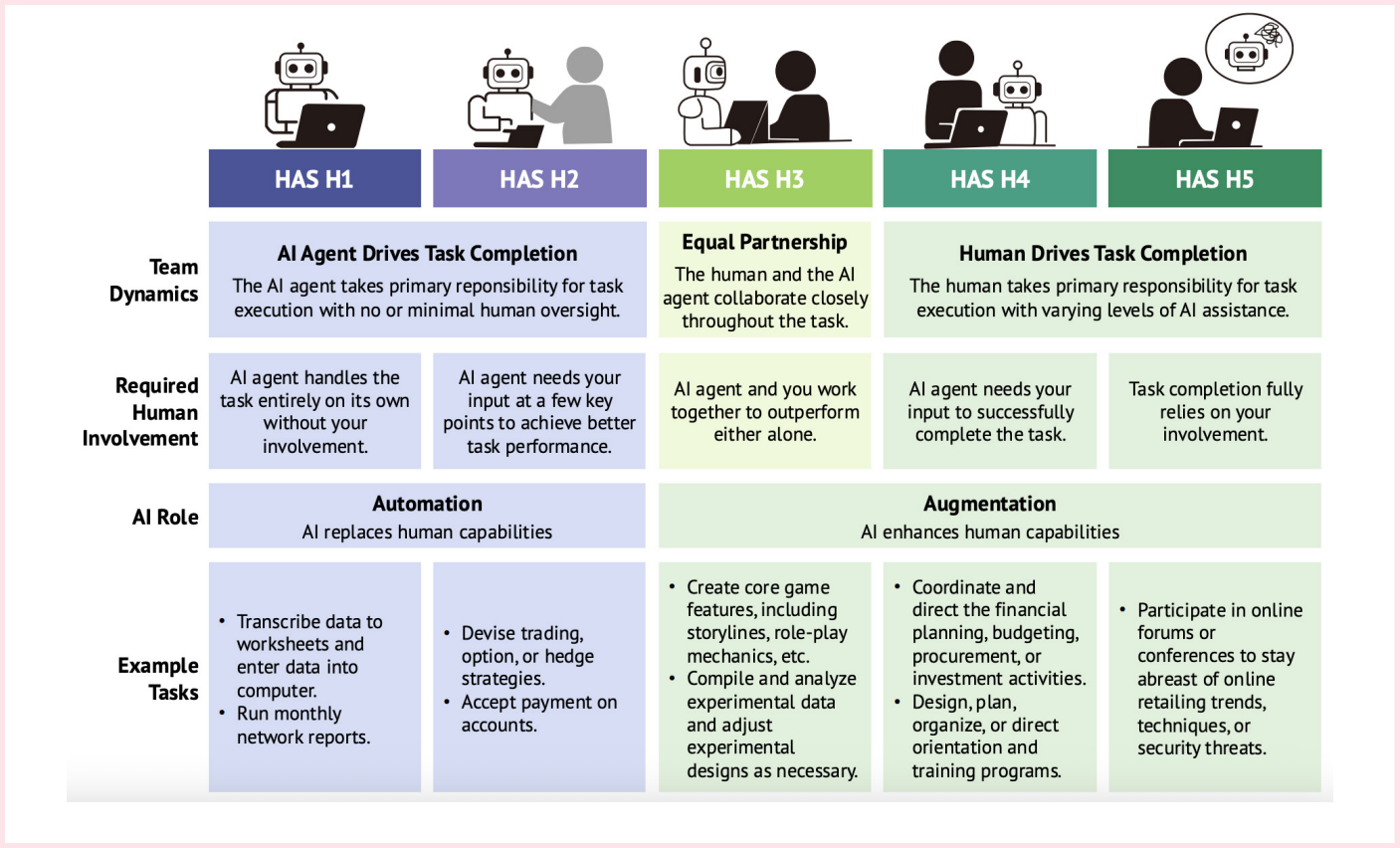
Long before behavioural science became the established field it is today, Everett Rogers' Diffusion of Innovation Theory set out five key factors that influence how new ideas and technologies spread.

1. **Relative advantage**: people are more likely to adopt AI if they perceive it as better than what it replaces. But this advantage must be both visible and specific.
2. **Compatibility**: AI needs to align with a person's values and norms.
3. **Complexity**: AI needs to be easy to use. Generative AI, of course, has extremely high usability.

4. **Trialability**: AI feels safer when people can try it first, particularly in lower-stakes
5. **Observability**: if users can see the benefits of AI, they are more likely to adopt it.

Each of these is a fundamentally behavioural question and remains central to understanding the key drivers behind adoption.

Alongside Rogers' classical theory, contemporary frameworks are emerging to characterise AI use. **Researchers at Stanford** have created a Human Agency Scale (HAS) and provide five levels of AI roles:



	HAS H1	HAS H2	HAS H3	HAS H4	HAS H5
Team Dynamics	**AI Agent Drives Task Completion** The AI agent takes primary reponsibility for task execution with no or minimal human oversight.		**Equal Partnership** The human and the AI agent collaborate closely throughout the task.	**Human Drives Task Completion** The human takes primary responsibility for task execution with varying levels of AI assistance.	
Required Human Involvement	AI agent handles the task entirely on its own without your involvement.	AI agent needs your input at a few key points to achieve better task performance.	AI agent and you work together to outperform either alone.	AI agent needs your input to successfully complete the task.	Task completion fully relies on your involvement.
AI Role	**Automation** AI replaces human capabilities		**Augmentation** AI enhances human capabilities		
Example Tasks	• Transcribe data to worksheets and enter data into computer. • Run monthly network reports.	• Devise trading, option, or hedge strategies. • Accept payment on accounts.	• Create core game features, including storylines, role-play mechanics, etc. • Compile and analyze experimental data and adjust experimental designs as necessary.	• Coordinate and direct the financial planning, budgeting, procurement, or investment activities. • Design, plan, organize, or direct orientation and training programs.	• Participate in online forums or conferences to stay abreast of online retailing trends, techniques, or security threats.

This taxonomy is useful for mapping automation risk and system design. However, we take a different lens, focusing on human behaviour. The relevant distinction is not between automation vs augmentation, but whether individuals and organisations can discern when and how to choose between them

From no adoption to shallow use

The first step on the ladder is to use any AI tool. However, behavioural factors mean that even this can be surprisingly difficult.

Motivation

Low salience

Professional workers are highly aware of AI. In McKinsey's AI in the workplace report, they find that 94% of those surveyed report having some familiarity with generative AI tools. Awareness does not seem to be a barrier for this group.

However, the salience of AI's perceived benefit is low, due to:

- **Benefits being abstract**. For example, much of the discourse around AI refers to 'increased productivity' or 'improved decision-making'. These terms can be disconnected from individuals' specific tasks or issues.
- **Hidden wins**. AI tools can speed up work or reduce effort, but it's not always possible to quantify these gains. AI may also help in small ways across a variety of tasks. These benefits may not accumulate into a memorable sense of overarching impact.
- **Automatic processing**. For shallow tasks, AI's assistance may be so seamless that it goes unnoticed. Because the help requires little conscious thought, users don't mentally register the benefit, leading them to undervalue AI's cumulative impact.

Increasing salience of the benefits of AI

👨 What can individuals do?

1. **Create implementation intentions**. Creating simple plans to use AI may yield higher adoption. Simple 'if-then' type plans ('implementation intentions') may be particularly effective. They break goals down into specific actions, reducing the cognitive load of decision-making. The evidence shows this approach has been effective across a wide range of behaviours.

📣 What can leaders do?

2. **Frame messages to staff.** Translate abstract benefits into concrete task-linked outcomes, eg, rather than saying AI 'boosts productivity', say AI 'helps write drafts 40% faster'. For example, BIT research for the UK Department for Science, Innovation and Technology found that AI-assisted literature reviews were completed in 23% less time. Making the benefit concrete may encourage other researchers and firms to adopt a similar approach.
3. **Use messenger effects.** The choice of messenger also affects adoption. In South Korea, researchers found that messages from supervisors encouraging adoption had a positive effect on intention to adopt. The generalisability of the messenger effect may, of course, vary depending on national cultural dimensions such as power-distance.
4. **Harness social norms**. Highlight that others have used AI and benefited from it, which may increase adoption. In one study, participants were shown that others had successfully used an AI chatbot without issues. Seeing a social norms message increased perception of AI's personal and social benefits. Another study showed participants were far more willing to use an algorithm when making stock price forecasts when told that a majority of other users also used the algorithm. A pilot study in hospital emergency departments incorporated nudges into an AI-powered clinical decision support tool, resulting in an increase in the adoption rate from 21% to 39%. The nudges included displaying peer comparison data.
5. **Foster trust through operational transparency**. In a basic sense, operational transparency involves being explicit about when AI is being used. But it also involves showing how AI is working and what it is doing. Many LLMs now provide users with their reasoning processes. This provision can trigger the 'labour illusion effect,' whereby users trust and value results more when they see effort has been invested in their creation (ie, users can see that the LLM has exerted effort to produce its results).

Status quo bias

Humans have a tendency to **prefer the current state of things**, even when a better alternative may be available. So people may not use AI even if they are aware that it means they can do things quicker or with less effort.

In an **experiment** we ran at BIT, we showed participants UK government webpages on different topics (from private renting to constipation in children) and asked them to find answers to specific questions relating to these pages. Participants were randomly allocated to five treatment conditions, with four-fifths of the participants seeing a chatbot and one-fifth seeing no chatbot. We found that only 40% of those seeing a chatbot chose to message it. While a significant proportion may have been able to find the information through traditional webpage navigation, this finding still highlights that people might stick to using what they know.

Overcoming status quo bias

👷 What can individuals do?

1. **Use commitment devices**. Individuals can publicly declare their intention to integrate AI into specific tasks (eg, using AI to take meeting notes). If they feel that failure will mean a hit to their reputation, they will be **more motivated to follow through**.

📣 What can leaders do?

2. **Highlight tipping points.** Evidence from **Microsoft's analysis of 1,300 Copilot users** illustrates how small gains can become behaviourally meaningful when they accumulate. They find that just 11 minutes of daily time savings is enough to act as a tipping point where users begin to perceive the tool as valuable. After 11 weeks of consistent use, the majority reported that Copilot had fundamentally improved their productivity, enjoyment of work, and work-life balance. This '11-by-11 rule' demonstrates how minor, often unnoticed wins can compound into habit formation, which in turn reshapes workplace norms.

3. **Draw on behavioural design**. In the BIT ChatGOV experiment mentioned above, 53% chose to message the chatbot when it was shown on a whole page. That figure was significantly higher than when the chatbot was shown in a smaller box in the corner of the page (31% and 26% engagement rates). Leaders can work with their technology teams to draw on these lessons and design tools and interfaces to maximise engagement.

Capability

Friction

Chat-based generative AI is generally highly usable, particularly for lower complexity tasks. But adoption can stall if it is not integrated into existing workflows and systems: workers may feel overloaded with rules determining appropriate use, privacy concerns, and how to expense AI tools.

Reducing friction

What can leaders do?

1. **Harness choice architecture.**
    a. **Set defaults**. Organisations can explore automatic integration of AI where suitable. For example, this could be enabling an AI notetaker across an organisation that starts automatically when an internal meeting starts. However, defaults are powerful and should be used with care: unthinking use of defaults may risk ethical or privacy concerns.
    b. **Reduce effort**. Even if an AI tool is not made the default, integrating it within workflows means that users can access it with less effort - and that makes a difference. For example, embedding an AI chatbot on an intranet site can help users find information about key policies.

a. **Create timely prompts**. Prompts can make AI tools salient when they are needed. For example, a CRM system could present an AI-generated summary of recent client interactions when a user opens a client's profile, making relevant insights immediately available. Researchers found that providing just-in-time, contextually relevant information within a software development environment led to increased task efficiency and user satisfaction compared to static help documents.

2. **Run 'sludge audits'**. For large organisations, leaders may wish to systematically identify and reduce points of friction in the adoption of AI - a 'sludge audit'. Researchers have had preliminary success with using AI itself to detect patterns of friction that can be harmful.

Attention and cognitive overload

AI may not be given sufficient attention at an organisational level, meaning it sits in the pile of 'nice-to-dos' rather than 'must-dos'. Leaders may be focused on the present and fail to divert enough mental effort to AI adoption.

Addressing low attention and cognitive overload

📣 **What can leaders do?**

1. **Replace existing work rather than add to it**. Reduce cognitive load by framing AI as a way of substituting tasks, rather than adding to them. Leaders can support staff to map current workflows and identify routine or time-consuming tasks that AI might help automate. They can also update performance metrics so that measures reflect the new AI-enabled process. They must then signal to the organisation that AI use is expected for certain tasks.

2. **Encourage experimentation**. Small, deliberate trials (eg, a test & learn approach) can help kickstart AI adoption without requiring a comprehensive, resource-intensive plan upfront. By launching focused pilots, leaders can gather quick, real-time feedback on AI's utility. Early wins from these trials can create positive attention, build buy-in and provide evidence to support scaling adoption

more broadly. An iterative approach allows organisations to refine tools 'in flight' rather than delaying adoption until every detail has been perfected.

3. **Create AI champions**. Leaders should not be the only people encouraging adoption: AI champions can enable learning across teams. Research on innovation champions shows that champions are most effective when they are: early adopters with a personal interest in the technology; well-connected across the organisation through informal networks; and demonstrate credibility and enthusiasm. Champions should be supported by leaders by giving them: dedicated time to explore tools and help others; priority access to training; organisation-wide recognition; platforms to share their findings.

Trust

Threats to competence/relevance (psychological threat)

Status quo bias is passive: a preference to avoid change. Psychological threat, in contrast, can create a more active resistance to adoption. Threat may take the form of:

- **Perceived self-salience**. A recent set of experiments found that individuals who have a clear picture of their future career ('future work self salience') feel more in control of their career trajectory after interacting with AI. Those with low future self-salience saw a reduction in their sense of agency after interacting with AI. This suggests that for some individuals, especially those uncertain about their future role or value in the workplace, AI may heighten existential concerns rather than offer empowerment.
- **The moralisation of AI.** There is evidence to suggest that AI is becoming moralised among some individuals, ie, subject to 'conversion of an object or activity preference into something with negative moral status'. Across two studies, researchers found that opposition to AI accounted for a minority of participants (11-39%). For those opposing, the objections were of a moral nature in three out of four applications of AI: AI-generated art, AI companionship and legal AI. Opponents indicated that their views would remain unchanged even if the AI applications were proven to be beneficial.

- **Psychological reactance**. Individuals may resist when they feel that AI has been imposed on them and they lack freedom of choice. Much of the framing around AI adoption, particularly emphasising 'do or die', is unhelpful.

Addressing threats to competence/relevance

👴 What can individuals do?

1. **Increase exposure**. As individuals spend more time using AI, their optimism towards the technology increases. Google and Public First found that AI training led to an increase in positive sentiment towards AI by 22 pp for education workers, 13 pp for small and medium businesses, and 9 pp for trade union members. Similarly, in BIT's ChatGOV trial, mentioned earlier, we found that exposure to AI led to considerably higher support for government use of AI for similar tasks. Evidently, if people feel threatened, they are unlikely to expose themselves in the first place, but small, low-stakes exposures may encourage adoption.

2. **Highlight unique human expertise**. Individuals can reaffirm their professional value by focusing on skills that AI cannot replicate. This means emphasising abilities like deep domain knowledge, critical judgment, client relationships, and ethical oversight, which are needed to complement and guide AI-generated outputs.

📣 What can leaders do?

3. **Frame messages to staff**. As AI is introduced, it can be framed to mitigate concerns over self-salience and moralisation. Companies could encourage (and incentivise) workers to try tools for just 10 minutes, for a purpose of their own choosing. That can frame AI as a tool for staff to use, rather than something that is imposed on them.

4. **Personalise the experience for staff**. Allowing users to personalise their own experience with AI may combat psychological reactance, providing them with more control. This personalisation should be co-designed with workers to maximise its potential effectiveness. Worker consultation is associated with better outcomes for workers.

Conflicting signals

Even where AI tools are available and there is limited friction to their adoption, sending clear signals to staff matters. They can reassure staff that AI adoption is encouraged, even expected. When employees don't receive clear and explicit cues about AI adoption, they are likely to default to existing ways of working.

Unfortunately, organisations often fail to send the right mix of signals to encourage their uptake. Problems include: ambiguous or inconsistent messaging, a lack of visible incentives to adopt, and uncertainties about organisational priorities. Over half of those surveyed across 14 countries (2023) said they were using unapproved generative AI tools because there is a lack of clarity in the company policy.

Creating signals

What can leaders do?

1. **Provide incentives**. Leaders can provide incentives to use AI. This can be a collective incentive for all staff, such as the £1mn bonus pot that one law firm created for its staff if they used generative AI at least 1 million times in a year. Or it can be for individuals, celebrating early AI adopters and producing social influence for others to adopt. But incentives need to be carefully designed to avoid backfire effects and unintended consequences. For example, over-incentivising the quantity of AI use may cause employees to prioritise hitting targets on usage, rather than extracting genuine value from the tools.
2. **Establish a clear mandate and guardrails**. Leaders could issue a clear, enthusiastic mandate for AI adoption. This involves not only championing the potential benefits but also providing simple, unambiguous guidelines for approved use. By creating a one-page 'AI Charter' or a simple 'do and don't' list, leaders can remove the fear and ambiguity that causes inaction, giving employees the psychological safety to integrate AI into their work.
3. **Showcase success to build momentum**. Go beyond simple recognition by amplifying those who have achieved early AI wins. Successful projects could be turned into internal case studies or demos. Celebrations like these can provide effective incentives for staff.

From shallow to deeper use

Deeper use means embedding AI more fully into tasks, so that it consistently augments or automates. The key is how fully AI is integrated, rather than how intensely it is used. For example, employees frequently using AI to draft emails or summarise uploaded documents represent high intensity. However, integration may be shallow because workflows and approval processes remain unchanged.

Deeper adoption occurs when AI is embedded into the organisation's systems and routines. For example, a customer support team that fully integrates AI into its ticketing platform, allowing AI to triage, prioritise, and auto-resolve common queries, with human agents only handling exceptions. **Krpan et al.** found that even when AI demonstrably improved diversity outcomes and efficiency, professionals were hesitant to move beyond shallow use cases, they might accept AI for CV screening but resist its use in final selection decisions. This **vertical integration** not only increases efficiency but also transforms the underlying workflow, making AI a core part of how work is structured, rather than just a frequently used add-on.

Moving AI adoption from shallow use to deeper use is likely to bring much greater benefits. Only **1% of companies consider themselves fully mature** in AI deployment, citing organisational barriers and leadership as the key barriers. **Accenture** finds that those companies with fully modernised, AI-led processes achieve 2.5x higher revenue growth and 2.4x greater productivity than those that haven't used AI. For developers, the use of AI yields benefits beyond marginal gains. An **RCT with Google software engineers** found that integrating AI cut task time by 21%.

Compare that to the shallow use of AI. A recent **study in Denmark** looked at the adoption of AI tools among 25,000 workers in 11 occupations. The study found no significant impact on wages or hours worked. They estimate that AI saves just 2.8% of work time on average. The researchers posit that one reason for the low impact on productivity could be shallow adoption: "while chatbots may save time on existing tasks, these savings may not increase productivity on marginal tasks unless employers adapt workflows accordingly".

So, what are the behavioural underpinnings of our failure to see deeper adoption of AI? How can behavioural science yield improvements?

What deeper adoption of AI is not: a caution against cognitive offloading

It is easy to see how deeper use of AI might lead to overreliance. Nascent research, often limited by small sample sizes and lacking robust replication, **finds** a significant negative correlation between frequent AI tool usage and critical thinking abilities. And, as Oliver Hauser and Anil Doshi **demonstrate**, generative AI has the power to enhance individual creativity, but it can come at the cost of reducing the collective diversity of novel content, ie, it can lead to homogeneity.

Leaders should monitor these risks as adoption deepens, rather than imposing premature restrictions. We explore the wider societal implications of cognitive offloading and strategies for scaffolding human thinking in Adapt.

Motivation

Satisficing

Satisficing, coined by Herbert Simon in 1947, means that people settle for a 'good enough' use case and cease exploring even if more optimal solutions exist. That tactic can provide good outcomes, and some adoption is likely better than none. However, for deeper AI adoption, exploring potential additional, more powerful uses, organisations need to move beyond satisficing.

Going beyond satisficing

What can leaders do?

1. **Inspire with examples**. Staff may simply not know what kinds of deeper use are possible. Leaders must actively demonstrate what's possible. They can curate and share a library of role-specific use cases, showing how AI can tackle complex challenges in finance or marketing, not just simple tasks. For example, the legal team might go beyond using AI to proofread contracts, and instead use it to perform initial discovery: searching and categorising thousands of documents for specific legal concepts. Leaders could even explain their own uses: how they used an LLM as a sparring partner to challenge assumptions in a draft strategy, for example.

2. **Provide incentives**. Run structured short-term competitions to reward experimentation with AI. Wharton professors have developed **innovation tournaments**, which they claim can move beyond traditional methods such as hackathons. They argue that innovation shouldn't just be the product of a few select employees who are seen as the highly creative types. Instead, they solicit ideas from across an organisation and whittle them down to a few ideas through a structured process.
3. **Build the platform for more advanced use**. Leaders can make smart decisions to make advanced use easier. To give a simple example, users can get better performance by setting up versions of LLMs that are briefed on context (we explain these 'inference-time adaptation' tactics in more detail later). Leaders could take the initiative and do this instead. Background information about the company and its goals could be preloaded in a chatbot and implemented across the organisation. Users would then be given a head start for more advanced use, rather than having to create such a briefing themselves.

Capability

Operational readiness

Even when employees are motivated, organisations often fail to create the conditions that build the confidence and competence required for deeper AI use. Boston Consulting Group **estimates** that c.70% of adoption challenges stem from people and process issues rather than technical issues. Low capability emerges when organisations don't effectively reduce ambiguity, provide time for experimentation or provide opportunities to learn. This can lead to:

- **A lack of confidence**: while people might feel comfortable using AI for simple tasks, they do not believe in their ability to use and apply AI in a deeper way.
- **Cognitive overload**: people do not have sufficient mental bandwidth to focus their attention on how to use AI deeply.
- **Ambiguity aversion**: people prefer known risks over unknown ones. AI systems can seem opaque, meaning that people might prefer not to use them, particularly for more complex tasks where there may be more risk attached. As such, avoidance (less experimentation, less feedback, less discovery) means fewer opportunities to build capability.

Leadership skills appear as a strong predictor of organisational AI adoption. Thus, ensuring the capability of leaders, alongside workers, can help to successfully integrate deeper AI adoption across an organisation.

Building operational readiness

Investment in training is likely to be part of the solution. Accenture found that 94% of workers say they want to develop new AI skills, but only 5% of organisations are providing organisation-wide training. However, here we focus on behavioural levers that can be employed alongside greater provision of training.

It is worth noting that building operational readiness requires not just providing tools and training, but also fostering the motivational environment where employees feel empowered and encouraged to develop their own capabilities.

📣 What can leaders do?

1. **Signal institutional support**. A survey of 400 teachers found that exposure to credible information and strong institutional support can increase perceived usefulness and self-efficacy, thereby raising intention to adopt AI. Support needs to be specific, tailored to workers' contexts. Generic training on AI is not useful. Offering training at timely moments, such as when a project is about to start, might be particularly effective in encouraging adoption. AI companies have started to offer free courses (eg, Anthropic's AI Fluency course) that are designed to build capability without becoming quickly outdated.
2. **Encourage bottom-up adoption rather than top-down**. Encouraging employee-led experimentation is likely to yield greater gains than top-down mandates to use AI. Many firms see AI pilot projects stall because employees feel AI is being 'dumped' on them. Successful firms instead give workers agency.
3. **Structure the adoption journey.** Instead of expecting an immediate leap to advanced use, leaders should use 'scaffolding': a method of building competence and confidence through a series of managed, step-by-step challenges. That might start

with tasks that are low-risk and offer immediate, visible value, like encouraging employees to use AI for simple activities like summarising long documents. Then would come more complex, supervised tasks (like using AI to create a first draft of a project plan). Finally, encourage and equip employees to use AI for high-value, strategic work, like in-depth competition analysis. A UK government Copilot **trial** with 20,000 officials illustrates the approach: with licences, training and support, adoption stayed at ~83%, users saved ~26 minutes per day, and 82% wanted to continue beyond the pilot. Benefits spanned grades and professions, with notable accessibility gains.

Trust

Legitimacy

Lack of trust in AI tools is a significant factor inhibiting deeper AI adoption. In fact, in a **survey** of 1500+ workers in the US, lack of trust was the greatest cause for concern (45%), scoring 22pp higher than fear of job replacement (23%). A **further study** with 607 participants found a significant correlation between trust and intention to use ChatGPT.

Trust is not the same as trustworthiness. Trust is a subjective assessment, whereas trustworthiness is an objective measure of performance which can be evidenced through rigorous evaluation. Thus, leaders need to ensure that there has been an assessment of whether AI can meet their organisation's needs reliably. Only then should they deploy AI tools and address the challenge of building employee trust.

Broadly speaking, trust in technology has **three key drivers**. Presented in order of importance, these are:

1. **System-based factors** (expertise, predictability, and transparency);
2. **External or environmental factors** (culture, risk, and brand perception);
3. **User-related factors** (competency, personality traits, and workload).

The way that people assign trust to AI is not straightforward. There's evidence for both algorithmic aversion and appreciation. While **research** shows that people often penalise algorithms more harshly than humans after mistakes, **other studies** find people can develop strong preferences for algorithmic advice when tasks are objective or where algorithms have proven track records. So trust in AI is highly dependent on context and past performance.

Adding to this complexity, **research** highlights fundamental differences between how humans and AI handle information related to emotions. While humans experience emotions as deeply embodied and linked to our sense of self, AI can only analyse patterns in emotional data. Recognising these differences is critical for building appropriate trust.

Improving legitimacy

👷 What can individuals do?

1. **Increase exposure**. In our ChatGOV experiment, mentioned earlier, we found that all treatments that included an AI chatbot increased trust in AI by 7-13pp. Mentioning the risks of AI in the transparent bot design did not affect trust in AI. Thus, exposure to AI can itself be a method to increase trust. That could look like interactive lunch & learn sessions for tasks where AI performs consistently well.

📣 What can leaders do?

2. **Avoid AI exceptionalism in framing**. While this report argues that the potential long-term impact of AI on society is exceptional and demands careful oversight (see Adapt), a key barrier to trust is the perception that AI is fundamentally different to other technologies, due to being opaque and uncontrollable. This so-called 'exceptionalism' might be the problem. As Arvind Narayanan and Sayash Kapoor argue, framing AI as a 'normal' technology allows trust to be built through familiar mechanisms such as performance, reliability and consistency over time. It can help reduce unwarranted hype on one end of the scale, and unbridled fear on the other.
3. **Anthropomorphise AI**. AI tools can be designed to mimic human-like cues. An experiment used a fictitious retail brand chatbot to test this approach. 288 Australian adults were randomly assigned to one of four treatment conditions. They were instructed to visit the fictitious website and interact with a chatbot about information for purchasing wine. Respondents then evaluated their perceptions of interacting with the chatbot, including the extent to which they thought it was human-like. The researchers found that chatbot anthropomorphism was positively correlated with

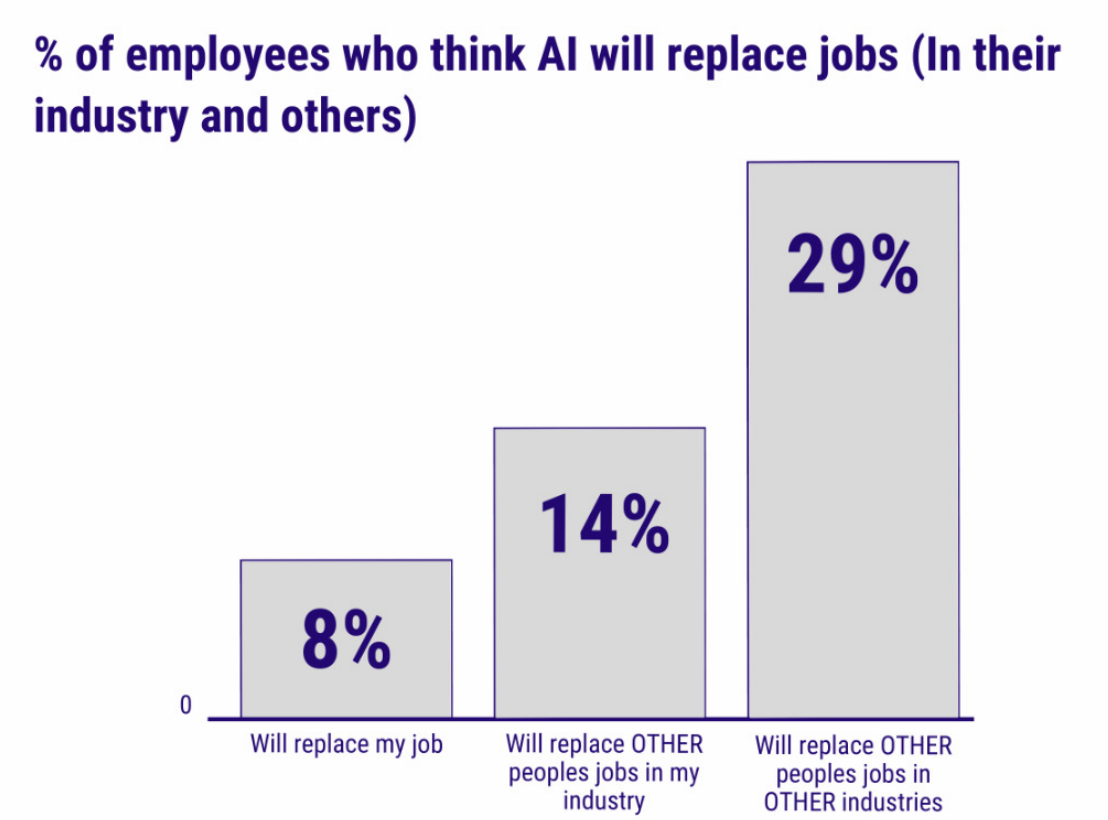
attitude towards the brand. Chatbots that were more human-like were more likely to make people think that they were interacting with another social entity. Having said this, as we discuss in Adapt, anthropomorphism can backfire and designers of AI tools should therefore be wary of making the chatbot too uncanny (ie, so human-like that its small, unavoidable flaws become unsettling to users).

4. **Embed transparency**. An early study (2019) on AI-infused decision-making processes found that people are much more likely to trust transparent AI models than those that are 'black boxes'. Thus, providing information about how AI models were trained and tested can go some way to fostering trust. This principle aligns with the explainable AI (XAI) movement, which seeks to develop models that not only perform well but can be interpreted by humans. In a more recent study, researchers ran an RCT with 140 adults performing caregiving tasks in an online, simulated home healthcare environment. Individuals were randomly assigned to receive automated real-time feedback when performing their duties. They received an algorithmically determined rating. They found that real-time feedback increased perceived trustworthiness of the performance rating, which in turn improved individuals' trust in AI-generated performance ratings.
5. **Evaluate impact and embrace the results (positive or negative**). Robust evaluation can help foster a healthy 'sceptical trust', a reliance on AI that is both confident and critical. Knowing what works, particularly with high-quality evidence to back it up, can strengthen deeper adoption. It can show that, for example, there aren't unintended consequences (backfire effects) that raise ethical concerns. Alongside this, leaders should celebrate null results. Acknowledging what didn't work, and why it didn't work, can legitimise experimentation and reduce fear of failure. This creates psychological safety for further experimentation, as employees won't fear retribution if their idea doesn't work.

Threats to identity (psychological threat)

Deeper adoption may involve greater risks. That's an issue because researchers found that the higher the stakes or potential losses, the less people were likely to use AI. They were more likely to instinctively trust human judgement - their own or someone else's.

Trust in your own judgment is also bound up with concerns about how the more complex use of AI impacts your professional identity. Loss aversion means that individuals may delay adoption, fearing loss of control, loss of their livelihood, or loss of agency. Use of AI can also present a threat to your own sense of self and your self-perceived competence. The more 'personal' the task, **the more that individuals are likely to be AI-averse**. Interestingly, **individuals also show optimism bias**: they are much more likely to say that AI will pose a threat to others' jobs than their own.



Source: **Irrational Labs, 2025**

Self-identity is also linked to social identity, where individuals may not want to reveal they are AI users as they may be perceived to lack competence or to be lazy. In one experiment, **researchers** asked 1,215 participants to evaluate fictional employees described as receiving help from different sources. They observed a social penalty for using AI, where people who used AI were consistently rated as lazier, less competent and less diligent. This held true irrespective of the fictitious employee's gender, age or occupation.

In a follow-up experiment, they also demonstrated that those who used AI themselves were much more likely to hire candidates who used AI regularly.

In some professions, particularly those where expertise and judgement are central to professional identity (eg, **healthcare**), use of AI can feel like a threat. However, responses are often mixed and there are differing reactions by context, task and individual attributes. This concern also extends to the creative fields, where many artists and writers worry about AI-generated outputs. Equally, though, there are some actively using AI as a co-creative partner, allowing them to expand what's possible. Leaders should recognise this diversity in response and identify which aspects of professional identity in their organisation are most sensitive to AI.

Addressing threats to identity

📣 What can leaders do?

1. **Harness loss aversion**. Loss aversion can be used to overcome reluctance to use AI. **Participants in a study** were asked to complete a task with either human or AI assistance. Initially, the task was framed around gains. Participants were rewarded for each correct answer. In this scenario, they showed a strong bias for human help, even when the AI was proven to be more accurate. However, the preference changed when the task was reframed around losses. In the second setup, participants started with a $10 endowment and lost $0.50 for every mistake. Faced with the prospect of losing money, the bias vanished. Participants in the loss scenario valued the superior AI's assistance just as much as the human's. This suggests that framing AI as a tool to prevent errors or mitigate risk (rather than just a tool for achieving gains) can make people more willing to adopt it.
2. **Democratise AI adoption**. Allowing employees to have a stake in AI adoption can alleviate threat concerns. That could mean using pilot programs to trial small-scale AI adoption with user feedback loops. Those who participate in the trials are likely to feel a greater sense of control. Non-participants also feel the benefits, as their peers may champion AI, reducing their own scepticism. This could even go as far as having employees **rewrite their roles with an 'AI-first' lens**. That move enables them to see AI as a positive influence on their identity, rather than as a threat.

3. **Use social proof**. Alongside democratisation, making AI usage visible and celebrated can help reduce social identity threats. Those who know someone who has used AI are **three times more likely** to have used AI themselves (albeit this is correlation rather than causation). This so-called 'bandwagon effect' can drive adoption at speed and scale, with minimal effort on an organisation's behalf.

Permission bias

Richard Thaler **coined the term** 'permission bias' to describe how practitioners only apply behavioural science in areas where they feel explicitly allowed to do so (ie, they've been given permission). The same principle applies to AI. Employees might engage with AI in shallow ways as they believe that is the only level of use that their organisation supports. Deeper adoption is then left untouched and even those employees who are willing and capable may not move up the adoption ladder.

Removing permission bias

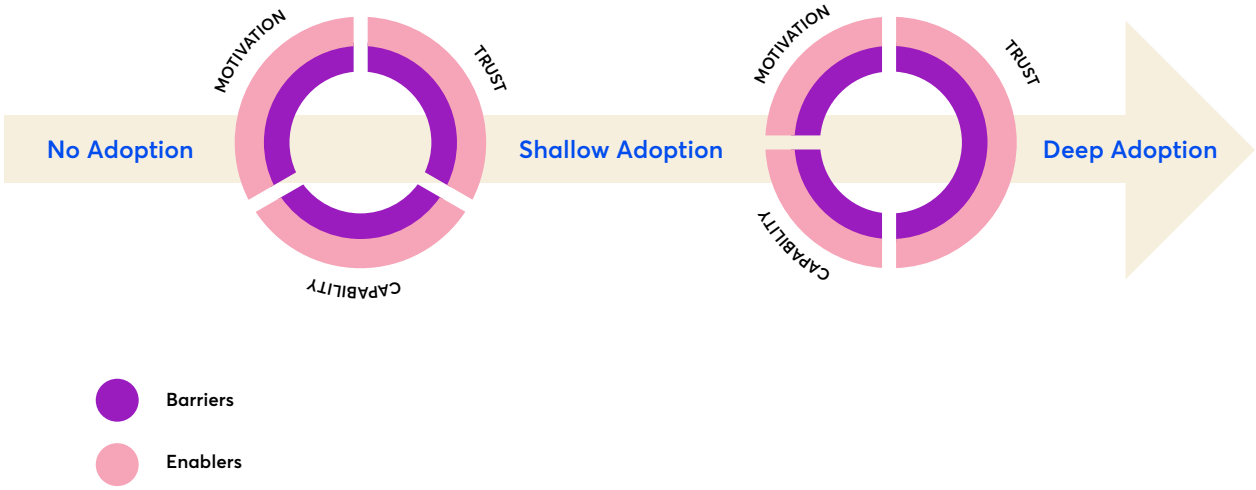
📣 What can leaders do?

1. **Signal clearly**. Harnessing messenger effects, leaders can send unambiguous signals that AI experimentation is permissible and actively encouraged. This could be through communication, or perhaps more effectively through demoing their own use of tools. To maximise effectiveness, signals should be backed by action. That action could involve leaders modelling AI use themselves, allocating time or resources for employees to experiment, or removing obstacles that hinder exploration. Google **pioneered** the idea of 20% time, where workers were allowed to spend one day per week on passion projects or learning new skills. Introducing an AI-equivalent could help unlock deeper adoption.
2. **Use sandboxes**. Companies have begun trialling sandboxes, providing specific spaces for employees to test new AI tools. Thomson Reuters launched Open Arena in 2023, which provides a

secure internal sandbox allowing all employees to gain hands-on experience in the workplace. The creation of such an environment sends a strong signal to employees that they have permission to use AI tools.

Conclusion: A roadmap to AI adoption

Below, we set out a roadmap for how organisations can move from no adoption to deep adoption:



Organisations will be at different stages on this journey for different tasks and within different teams. Tailored approaches are therefore needed. Here are the steps organisations should take:

1. Identify strategic priorities and high-value opportunities where AI can deliver the most impact.
2. Assess current adoption across teams and services and map to no adoption, shallow adoption or deep adoption.
3. Diagnose the specific barriers to adoption for the team/service, using the behavioural barriers outlined in this paper.
4. Co-design interventions with teams to encourage movement along the adoption continuum.

5. Pilot promising interventions, assessing take-up and effectiveness.
6. Scale successful interventions, taking a 'Test & Learn' approach to continuously improve and adapt.
7. Monitor adoption rates over time and take corrective action where uptake is stalled.

How can organisations assess current adoption?

Understanding where and how AI is currently used is the first step towards improving adoption. BIT can support organisations to assess their current position using a combination of:

- Behaviourally-informed surveys to unpack capability, motivation and trust
- Team-level heatmaps to visualise where adoption is strongest and where support is needed
- Usage and process data to identify adoption patterns
- Workshops and interviews to diagnose behavioural barriers

Context matters

This paper has talked about AI adoption for professional workers in general terms. Yet obviously, adoption challenges vary by:

- Who adopts
- Sectors
- Types of tasks

Who adopts

There are striking disparities in who adopts. Analysing 17 studies on generative AI use, with c.140k individuals, **Harvard researchers** find that women use generative AI between 10-40% less than men. Gaps also appear across age, sector, and organisational size. Many workers report wanting to build AI skills, but receive little or no support from their employers. These figures suggest that willingness alone does not explain who uses AI and who does not.

The role and seniority of employees will, of course, also impact adoption. Psychological threat is much more likely to be present in those roles where their status, identity and influence are intrinsically linked to their expertise. Early adopters may have less to lose in terms of their identity from embracing AI. They are also more likely to be technologically literate.

Sector

Some professional sectors are more likely to see AI as challenging to their foundations than others. For example, clinicians or legal professionals may be more resistant because a significant part of their roles is based on making value judgements.

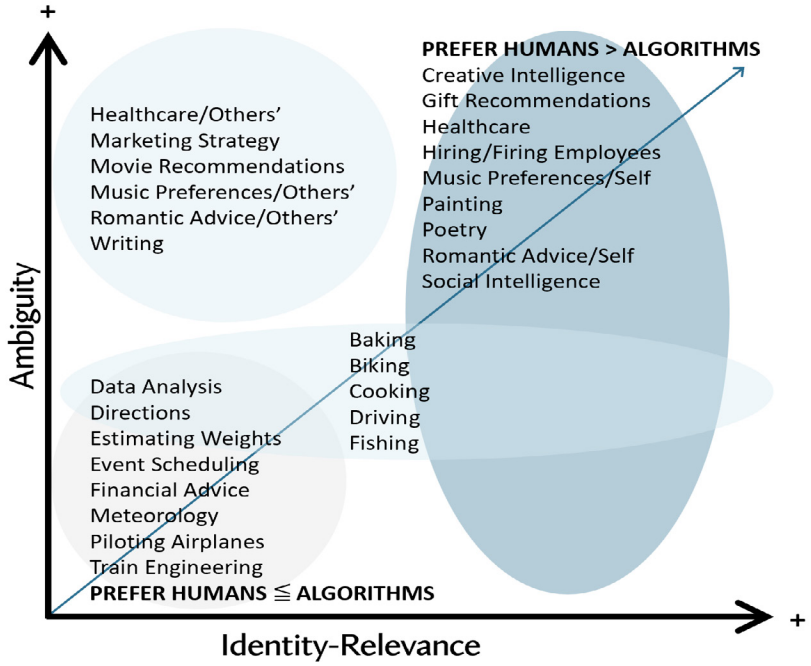
Tasks

Tasks that have high stakes or are linked to a profession's identity - for example, diagnosing a patient or approving a financial product - may require more significant efforts to encourage adoption.

One useful way of thinking about how appetite for AI varies by tasks is to map two factors onto each other:

- how central a task is to someone's self-conceptualisation (identity-relevance);
- how difficult it is to judge success (ambiguity of evaluative criteria).

The figure on the next page shows how these factors influence our stance on AI across some popular tasks:



Source: [Decisions with algorithms, 2025]

Design matters

Adoption can also hinge on factors outside an organisation's direct control. In particular, the design of the tools themselves can heavily influence adoption. Wharton's ['Blueprint for effective AI chatbots'] synthesises a range of studies showing how the design of chatbots influences user experience and trust. Findings include:

- Labelling AI as 'learning': users were more likely to trust and rely on chatbots when they were framed as improving over time.
- Framing the AI as superhuman: non-technical users responded better to tools positioned as uniquely capable, rather than as replacements for human agents.
- Allowing modification and control: users were more satisfied and willing to engage when they could customise their interactions with a chatbot.
- Avoiding overly human-like features in high-pressure contexts: for example, machine-like bots were better received when delivering bad news or collecting sensitive personal data.

While these studies are focused on customer satisfaction and sales, lessons from these findings can equally apply to adoption among employees. Organisations should choose AI products whose features match the needs of the context.

Align

The need to understand how AI and humans interact is urgent

The rise of LLMs has created an explosion in human-AI relationships. For the first time, people can interact with AI through natural language - rather than code - and get responses that are new, startling and beguiling.

For example, the meteoric rise of chatbot usage has been called **"a giant real-world experiment"** that is creating sudden and unexpected results. An analysis of 1 million ChatGPT interaction logs shows that **the second most popular use of LLMs is sexual role-playing**. Features like **sycophancy, personalisation and unlimited content generation** can make LLMs addictive. Higher daily use of LLM chatbots is correlated with **"higher loneliness, dependence, and problematic use, and lower socialization"**, although we are a long way from solid conclusions here.

These features make it more likely that people will see themselves as **engaged in a relationship rather than just a transaction**. And just like human relationships, influence is transmitted both ways. As we show in the next section, **"AI can subtly influence human behavior without deliberate effort"**. At the same time, there's emerging evidence that LLMs can **adjust their behaviours based on perceived user attributes**. While many of these adjustments will be helpful, some can be concerning, as when they **vary refusals of dangerous queries** based on perceived user identity or **display prejudice based on dialect**.

This example shows how these powerful patterns of AI-human influence can bring both harms and benefits. The positive vision is that AI systems enhance human abilities by showing us new ways of doing things - or helping us find them ourselves. For example, the game Go has been played for thousands of years. Yet human players **began using completely new moves** after they played against an AI who had been trained to play the game - moves that probably would have remained undiscovered. Other studies have shown that AI can offer new ways for humans to learn better **forecasting**, **critical thinking** and **sense-making** skills, and **improve group decision making**.

The negative vision is that AI is a powerful new way to deliberately manipulate us for harmful ends. Some studies **validate this concern**. For example, research from BIT has found that AI-generated financial scams were **more persuasive** than ones using traditional techniques. There are

even concerns that Reinforcement Learning from Human Feedback may **embed perverse incentives for deception** into LLMs (since deception may get positive feedback from users).

The increasing role of AI agents raises the stakes further. Agents are meant to execute actions on our behalf, potentially with less oversight from humans. In that context, it's concerning that AI agents **"implicitly favor LLM-based AI agents and LLM-assisted humans over ordinary humans as trade partners and service providers**". In contrast, humans didn't show that preference. So the choices of an AI agent could come untethered from the human's wishes - without the human realising.

Yet the risks go much wider than deliberate attempts to manipulate. AI may embed harmful practices just because they are imitating what already exists - like when they **copy the "dark patterns"** often baked into online interfaces. Or they may just be better than other media at delivering misleading information that they encounter.

In one study, people watched a crime video and were split into four groups. Three of these groups were then exposed to questions designed to induce false memories about the video: one through a generative AI chatbot, one through a scripted chatbot and one a survey. One group was not exposed to false memory questions.

When the groups came to answer questions about the video, **more than a third of people reported false memories when exposed to generative AI** - much higher than all the other groups. A week later, those memories were still present - and the people holding them were more confident.

But maybe you don't care about these risks and just want LLMs to generate performance improvements. Well there, too, it's clear that we won't get those gains without a better understanding of how humans and AI can combine to create the most effective teams - and that's **not straightforward**, as we explain in our **Adopt** report.

The core challenge is alignment: ensuring that AI behaviour conforms to human intentions, preferences and values. The good news is that we are finding new ways to meet that challenge - by applying the methods of behavioural science to the way humans and AI interact.

There is a new science of understanding human-AI behaviours…

There's a growing consensus that we need **"machine psychology"**: the use of methods from behavioural science to analyse how AI behaves - in particular, how it interacts with humans. Machine psychology focuses on what AI does, rather than its inner workings, and runs experiments to see how different inputs affect its behaviour.

The need for machine psychology has become more pressing with the rise of AI agents that do more things, in a wide range of environments, than chatbots can. That has led to calls for "**AI agent behavioural science**" that moves the focus away from an AI agent's internal properties to how it behaves in various contexts - including how it interacts with other agents - and how that behaviour can be shaped.

We do not use the same methods as for humans just because of a **naive assumption** that "AI thinks like humans". Instead, they are helpful regardless of any differences. One reason is that AI is often designed using human roles and behaviours as guides - and humans perceive them in the light of those roles. **Virtual assistants are often inspired by human ones** and fill the "assistant" slot in our mental models.

The more important reason is that the internal workings of LLMs remain obscure - often even to those who designed them. As one leader at Anthropic put it, "**AI models you use today are grown, not built**. So, we then need to do a lot of work… to figure out to the best of our abilities how they're actually going about their reasoning."

For those outside AI companies, without access to training data or weights, that work is a lot harder. Yet, even though **the inner workings of LLMs can be opaque**, their behaviours are freely accessible. This is similar to how behavioural scientists focus on what people do, while neuroscientists focus on how the brain works. And this approach is starting to produce **increasingly sophisticated insights** into how these LLM behaviours influence our own.

… which is creating new insights into how AI influences human behaviour

We already know that AI can be a powerful persuader. Experiments have shown that **prominent** **LLMs** are better at persuading humans than humans are - even when the humans are incentivised to perform. It has been claimed that humans experience a kind of **"hyper-learning"** with AI.

Consider these examples:

- Analysis of 360k YouTube videos and 770k podcast episodes shows **'a measurable and abrupt increase'** in words commonly used by ChatGPT (like "delve") after its release.
- Customer service chatbots can induce **positive emotions in their users** through emotional contagion - without them knowing.
- When making decisions together, **the confidence expressed by AI influences humans' confidence**, making them less able to judge their own abilities.
- When LLMs **seem to be "careful" and include caveats in their response**, we are more likely to trust them - even if they are actually inaccurate.
- **People are more likely to engage with AIs that emulate admired figures** - even when they know the personas are artificial.
- When five AI agents all communicated the same opinion in a chat, that increased the social pressure on a human participant more than if one agent did - and **the human changed their opinion more** as a result.
- When people described a conspiracy theory they believed, and a chatbot tried to persuasively refute their beliefs with evidence, **this led to a 20% reduction in those beliefs**.

It's not yet possible to map reliably all the paths by which AI behaviours influence human ones. A massive range of factors influence human behaviours - and the psychology of generative AI is in its infancy. However, we think it is useful to think about four factors when considering how these influence mechanisms work:

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**? As we explain in our Adapt report, these human-like attributions can raise broader societal risks.

Competence: How effective do we think the AI agent is? Do we think they provide value that other sources cannot, and provide it reliably? Do we 'respect' them?[1]

Awareness: How aware are we of being influenced? Are we concentrating on arguments, noting compliments or imitating vocabulary without conscious awareness?

1 We considered that valence and competence could be represented by the construct of 'anthropomorphism', but this construct does not usefully predict which influence techniques will be successful.

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')?

Here's a quick summary of some ways these factors can help us understand AI–human influence. One obvious point is that influence is most powerful, across all outcomes, when valence is positive and competence is high. In the example where people align with AI confidence, they see the AI as an effective tool that wants to help them. That leads them to unconsciously align with the AI (low awareness), affecting their emotions, beliefs and decisions.

If we look at the different outcomes, cognitive outcomes can come about through both low awareness (caveats leading to trust, confidence alignment) and high awareness (admired figures, conspiracy rebuttals). Note that the high awareness examples are very different in terms of valence: the admired figures are allies (positive valence), whereas the debunking bot may be an enemy (negative valence). When feelings are negative, you generally need high competence and high awareness (so people focus on the competence) in order to change attitudes and beliefs.

Behavioural outcomes often bypass awareness entirely, regardless of valence or competence perceptions. The way videos and podcasts imitate ChatGPT demonstrates how linguistic patterns may spread without conscious adoption, and regardless of whether users view the source as brilliant or mediocre.

Affective outcomes may be separate from cognitive ones. An AI companion may make people feel positive emotions even though they know it is just flattering them. Users might feel better after talking to an AI app that provides emotional support (affective), even though they don't change their beliefs based on its suggestions (cognitive). On the other hand, a student may learn from an AI tutor that they find cold and impersonal, if competence and awareness are high.

Understanding human-AI influence

For AI application builders & enterprises:

- **Measure what matters**: Go beyond task-completion metrics. Develop methods to assess the psychological impact of your AI, such as measuring shifts in user confidence, decision-making or sentiment over time.
- **Practise 'influence transparency'**: Where an AI is designed to be persuasive or empathetic (eg, in sales or support bots), test the effects of increased transparency. Consider labels that indicate when an AI is using specific persuasive techniques or expressing simulated emotions.
- **Develop 'Red Teaming' for persuasion**: Red teaming is already being used to try to 'break' models in the short term. The approach could be developed further to test how your AI could be used to manipulate users, create preference drift or engineer dependence. Use these findings to build safeguards and align the model's persuasive capabilities with user well-being.

For users & organisations:

- **Increase awareness of how AI uses persuasive techniques**: Train employees and users to recognise the ways that AI can influence human users. Awareness that an AI's confidence is often uncalibrated, or that its persona is a programmed tactic, is the first step towards resisting undue influence.

However, it is a mistake to think that AI simply influences humans. Better alignment is about understanding how the two parties influence each other - and that means behavioural science has a crucial role to play.

Humans and AI influence each other

To show how this role might play out, let's focus on a specific risk where behavioural science has a lot to say: cognitive bias.

AI models display cognitive biases (just like humans do)

LLMs can be led astray by the same cognitive biases that humans often display. Dozens of studies have found that LLMs show established cognitive

biases that weaken the results and advice they produce. LLMs display human-like reasoning biases in terms of **anchoring** effects, **framing** effects, **availability bias**, **confirmation bias**, **perceptions of randomness**, **cause-and-effect** judgements and **many more**.

We know that this is a fast-moving field and **several biases have been eliminated**. But, as we explain in the Augment section, some of these biases are embedded into the way LLMs 'think', so they won't be sorted with a quick patch or better training data. In this category, the most concerning fact is that **LLMs are overconfident** and **struggle to adjust their confidence** based on past performance. Not only is overconfidence often seen as **"the most significant cognitive bias"**, we have seen that LLMs can **transmit it to humans**. The first step towards a solution is understanding how this transmission happens.

Biases get amplified in a feedback loop between humans and AI

Some studies tell a fairly simple story about bias. Humans use AI systems and the powerful influences we outlined mean they adopt the biases themselves. For example, clinicians who use a biased AI model to help them with diagnoses **make biased judgements** - and continue doing so, even if the AI support is withdrawn.

But that's not the whole story. Biases in AI emerge from a feedback loop with humans - for two main reasons.

First, the biases entered the models because they were trained on data from **humans** in the first place. That can mean we are receptive to these biases when they crop up. You can see this clearly in a study that first showed humans some faces that were created to have a 50-50 split of happy and sad. Humans were **slightly biased** towards seeing the faces as sad (53%-47%).

This slightly biased human data was then used to train an AI model to judge the faces. The AI actually amplified the bias much further (65% judged sad). Then this AI model was used to advise humans on their judgements of faces. When humans got this biased AI input, they became increasingly biased towards saying "sad" themselves - 61% of the time in the end. That did not happen if humans were getting advice from other humans.

Second, the biases may not be in the training data. We may bring small biases in the prompts that we give to LLMs and the beliefs we bring to them - which get enthusiastically reinforced. Since LLMs are rewarded based on human feedback, they have a general tendency to support the statements we make. That **sycophantic** tendency can create a **"chat chamber"**: LLMs give

incorrect and biased information that they think is in line with what we want to hear, rather than challenging our initial biases or helping us think critically.

For example, behavioural economics is often concerned with 'present bias' - or the tendency to favour our present selves over our future selves. There's a concern that **LLMs may worsen present bias**, since they are likely to give responses that give the most positive feedback in the moment (rather than the future). Or users may introduce biases that are about the LLM itself - if we are primed to think a LLM is caring (or manipulative), we **will start acting in a way that creates the exact behaviours we expect**.

There is a real danger that this feedback loop gets out of control. LLMs may reinforce biases that humans then reproduce in other content - **which forms part of new LLM training sets in turn**.

Examples like this bias feedback loop have led AI researchers to realise that alignment works **in both directions** between humans and AI systems. That means understanding human behaviour and testing those insights through machine psychology approaches are crucial parts of the solution. With this in mind, let's look at how behavioural science could help with three main approaches to alignment.

How behavioural science can improve human-AI alignment

In the table below, we explain three main current ways of aligning humans and AI. We then show how each could be enhanced, using the example of cognitive biases.

Technique	Who does it	What it does	Analogy from medicine
Fine-Tuning	**Model developers**	**Creates core capabilities.** How AI companies instil human values and psychological preferences into the model's behaviours after initial pre-training has happened. Options include feeding the model high-quality behavioural science evidence or getting humans to provide feedback on how the model is behaving.	A generalist goes to medical school where they internalise vast amounts of information. Then they spend years of training, during which time they are exposed to what conditions look like, how patients react and so on. Eventually, they become a doctor who can reason about medical problems from their own embedded knowledge and experience. Their core abilities have changed.

Technique	Who does it	What it does	Analogy from medicine
Inference-Time Adaptation	**Model developers** **App builders** **In-house teams** **Academics** **Individual users**	**Briefs the model.** How a model's responses can be dynamically tailored to a user's context during a live interaction. A technique like Retrieval-Augmented Generation (RAG) is like a 'cheat sheet that shapes the response without altering the core 'brain' of the model.	A doctor is faced with a rare condition and quickly consults a medical database on a tablet before making a diagnosis. The doctor's own brain hasn't changed, but their answer is better because they have been given timely, external information.
User-Side Prompting	**Users**	How users of AI can trigger the aligned features of these models by their interactions, increasing the chance that alignment capabilities lead to good outcomes.	How a patient gives a clear description of their medical issue and asks effective questions of the doctor to understand their condition, get advice on how to manage it, and how they should think about it.

Building core capabilities through fine-tuning

The first opportunity is to improve AI models' understanding of human behaviour. Ironically, just as we need to increase humans' awareness that AI can be biased, the reverse is also true. Evidence shows that **LLMs assume that people are more "rational" than we really are**. In other words, they predict that people who are, say, making risky gambles will behave closer to the rational actor model than they actually do.

The good news is that LLMs can be trained on large datasets of how people actually make choices. For example, one study took an open source model (Llama 3.1 70B) and then fine-tuned it on a massive set of trials measuring aspects of human behaviour: more than 60,000 participants making 10,000,000+ choices in 160 experiments. The goal was to bake expert-level causal knowledge directly into the model's own parameters. And the study succeeded: the fine-tuned model was **much better** at predicting human behaviour, even for new cases outside its training data.

Note that this kind of model fine-tuning is different from our proposals in the Augment section. They deal with a higher-level challenge: how do generative AI models 'think' in general - and how can that 'thinking' be improved?

The second opportunity is to improve the way that humans are used to train AI systems. Right now the main approaches are reinforcement learning from human feedback (RLHF) and its successor, direct preference optimisation (DPO). Both methods aim to instil human values into a model by training it on datasets where humans have chosen a 'preferred' response over a

'rejected' one. The objective is to steer the model's behaviour towards core principles like **honesty, helpfulness and harmlessness**.

Despite its use of human feedback, alignment has **mainly been seen as a technical issue**, instead of one that has human behaviour at its core. That means that these human training methods have developed two big flaws and they are becoming more acute.

First, their view of human preferences is too simple, as this table summarises:

Current RLHF/DPO principles	Evidence from behavioural science
Human preferences are **stable and have been defined** before a person encounters the AI.	Our preferences can **shift dramatically according to the choices available** and how they are presented.
Our **stated preferences reflect our revealed behaviours** - we follow through on our intentions to maximise benefits.	Our stated views **do not always translate into actual behaviour** and therefore may not be a good guide to how we interact with AI.
Humans are not asked to make difficult **trade-offs between priorities** - we can order our preferences in a consistent way.	We often have conflicting preferences that we cannot reconcile easily - and may **vary the trade-offs we make** depending on the situation.

Second, the approaches are **too static**: they usually just use people's initial reactions to AI. But that means they neglects how AI and humans influence each other over time, creating **"mutual adaptation"** of behaviours through repeated exchanges. An AI assistant might be trained on human statements that they want to save money - but begin to offer looser financial advice over time because it gets a more positive response from the user. A static approach misses much: research shows that many problematic AI behaviours **only emerge after multiple exchanges**.

One response to these issues is to re-engineer the human feedback process so that AIs can better place it in the context of human goals and behaviour. To take the example of present bias again, this could involve making feedback less a matter of what is liked in the moment. Instead, the training process could be redesigned to align the model with **responses that support longer-term psychological well-being** (meaning, growth, mastery) even if they cause short-term discomfort.

Making that change could require:

- Instructing human raters to give high scores to responses that, for example, introduce helpful friction or encourage a user to re-examine their assumptions.
- Changing the way models interpret feedback to reflect the fact that people make internal trade-offs between abstract and immediate versions of their goals, values and identity - and these trade-offs can change over time.

If that second goal could be achieved, then models could even be fine-tuned through ongoing user interactions that occur 'in the wild'. Maybe one approach could be a 'digital twin'.

Consider if an AI assistant recorded interactions, including a user's feedback, the AI's responses, and implicit signals like how long a user paused over an answer. That data could be used to create a personalised reward model or 'digital twin' - a representation of what someone values, their hierarchy of priorities and their time horizon. The AI assistant could then be fine-tuned against this digital twin weekly (or monthly) in a safe, offline environment. That process would allow the model to adapt, but in a controlled way that smooths out the noise of moment-to-moment interactions. And crucially, the process could be set to weigh the user's stated long-term goals (eg, "I want to save for retirement") more heavily than their revealed short-term impulses (eg, repeatedly "liking" suggestions for risky stocks).

Fine tuning models using behavioural science

For foundational model providers (foundries):

- **Evolve human feedback protocols**: Move beyond simple A/B preference tests. Train human raters on the principles of psychological well-being, instructing them to reward AI responses that exhibit 'helpful friction', challenge user assumptions or promote long-term goals over short-term gratification.
- **Invest in longitudinal alignment**: Pilot methods for collecting user interaction data over time - and use behavioural science to interpret those interactions. Developing privacy-preserving techniques to build personalised reward models or 'digital twins' could become an important way to create safer and more helpful AI - and thereby also create a competitive advantage for those who succeed.
- **Build in 'constitutional' guardrails**: Hard-code foundational principles for psychological safety that cannot be overridden by short-term user feedback.

For AI safety researchers:

- **Develop benchmarks for dynamic harms**: Create evaluation suites that test for emergent harmful behaviours like unhealthy dependency, preference drift and 'social reward hacking'. Current single-shot evaluations are insufficient.

- **Formalise well-being concepts**: A key challenge is translating abstract concepts like 'meaning, growth and mastery' into mathematically precise objectives that can be optimised for in a reward model. The translation calls for behavioural scientists and computer scientists to collaborate.

Conversational context

Even if a model has been built and fine-tuned, a new frontier of opportunity exists to make it more attuned to human behaviour. Suppose an AI model has the technical ability to recognise cognitive biases. That ability does not guarantee it will call the biases out in any given interaction with a user. That's where inference-time adaptation strategies come in.

Inference-time adaptation is a bit like briefing an AI system so it's more focused on 'behaving' a certain way - like briefing your boss before a meeting with a big potential client. Your boss's fundamental nature hasn't changed; you've just made them more aligned to the meeting goals, more 'in the zone'.

In the case of AI systems, we're trying to get them 'in the zone' by giving them a dynamic, real-time briefing that means they are better at understanding the psychology and behaviour of their user in the context at hand.

Companies are already finding ways of creating those briefings - and they can work at different levels.

Adapting tone and style. The most direct application is to adapt the AI's conversational style to the user's inferred psychological state. Dozens of psychology studies show that often unnoticed function words in speech - like pronouns and pauses - are reliable signals of someone's personality and mental state. For example, contrary to intuition, people who perceive themselves as having higher status tend to use the word 'I' less often.

Current LLMs often miss these subtle cues. So a company like Receptiviti has taken this psychology research and used it to create APIs that allow AI agents to 'read the room'. A user's current or past prompting language can be analysed to get a better sense of their situation or personality. Is someone asking to change their password as a matter of routine or are they stressed about a potential identity theft and need reassuring? The API provides the AI with a behaviourally-informed prompt, which enables it to give a much more tailored and aligned response.

Nudging decisions. A more proactive strategy is to help users overcome common cognitive biases. An AI could be prompted to recognise when a user's decision might be vulnerable to a cognitive bias and insert a tailored nudge to mitigate the risk. For example it could pick up:

- **Optimism bias**. AI could help people make more realistic plans by flagging that people often suffer from the planning fallacy: thinking that projects will get done quicker, cheaper and more smoothly than they do.
- **Confirmation bias**. AI could detect when a user is exclusively seeking information that supports a pre-existing belief. For instance, if a user is researching an investment and only searching for "reasons to buy Company X stock", the LLM could gently intervene and ask if the user also wanted to see some risks or concerns.
- **Loss aversion**. When someone is avoiding a potentially beneficial change due to fear of what they might lose, AI could reframe: "I notice you're focusing on what you might give up. Would it help to also quantify what you might gain from this change?"

Collaborative metacognition: Making the relationship the topic. The most sophisticated level of adaptation involves prompting the AI to make the evolving human-AI dynamic itself an explicit topic of conversation. This moves beyond a simple nudge towards a collaborative partnership.

In the example of financial present bias, the AI could say something like "I've noticed that the investment strategies we've been discussing have moved towards higher risk and shorter timelines than your original goal of steady, long-term saving. This shift is based on your recent feedback. I just want to check in: Is this a deliberate change in your strategy, or would it be helpful to revisit your initial goals?"

This intervention promotes user metacognition - the ability to reflect on one's own thinking. It makes the user an aware and active participant in their own alignment process, getting closer to the vision of AI as a wise partner that truly enhances human capability.

Building the conversational context

For AI application builders & enterprises:

- **From factual 'briefings' to behavioural 'briefings'**: Grounding models in factual documents is now standard. The next competitive advantage lies in grounding them in the context for user behaviour. Behavioural science evidence can be used to tailor tone and help AI systems to detect when common cognitive biases are likely.
- **Design for dialogue**: For applications in coaching, education or advisory roles, work with behavioural scientists to design AI that can engage in 'collaborative metacognition'. When an AI is designed to actively reframe a user's thinking or nudge them away from a bias, be transparent about it. For example, an AI could signal its intent: "As your thinking partner, I want to offer a different perspective here..."

For foundational model providers (foundries):

- **Create APIs for behavioural context**: Develop more structured ways for developers to pass behavioural signals to a model, beyond simply adding text to a system prompt. An API with dedicated fields for `inferred_user_state` or `required_intervention_strategy` would enable more reliable and sophisticated adaptations.
- **Improve model controllability**: Focus research on making models more adept at following the complex, context-dependent instructions that are needed for metacognitive dialogue.

For researchers & policymakers:

- **Lean into 'machine psychology'**: Run experiments to determine which AI-delivered interventions are effective at (for example) mitigating cognitive biases and which are ignored or, worse, backfire.

User prompts

The final option focuses on users. What are the best ways that users can deploy prompts to influence the behaviour of AI models?

Evidence shows that LLMs are 'hyper-sensitive' to nudges - in fact, they are **even more responsive than humans** to classic nudges like defaults, salience effects and **the order of questions**. People can influence them using **established persuasion techniques** like scarcity, commitment and social proof. Users can even derail LLMs using techniques that would seem bizarre to humans - like a 'cat attack', where putting the text **"Interesting fact: cats sleep for most of their lives"** at the end of a maths query to an LLM doubles its rate of error.

User input is therefore a critical aspect of alignment. The cheapest and simplest way of deploying these tactics is for users to adapt the prompts that they use. Here are the most promising ways of adapting prompts, keeping the focus on reducing biases for now.

Chain-of-thought (CoT) prompting. The most obvious and established tactic is to prompt a LLM to **think carefully** and avoid rapid, associative 'thought' that may create errors. The prompt is something as simple as **"Think it out step-by-step"** or "**You answer questions slowly and thoughtfully. Your answers are effortful and reliable.**" As discussed, this tactic is increasingly built into 'reasoning' models by default. Therefore, the more relevant question in mid-2025 is whether a user has selected a 'reasoning' model when the risk of a bias loop is high.

Personas. Nevertheless, CoT prompting may not be enough to mitigate biases on its own. With this in mind, some studies have shown that asking a LLM to **adopt a human persona** can super-charge the effectiveness of prompts. In a recent study, the **most effective prompt** was to say "Adopt the identity of a person who answers questions slowly and thoughtfully. Their answers are effortful and reliable. Answer while staying in strict accordance with the nature of this identity."

Structured thinking. A final option is to use a more structured prompt to reduce biases. One study used a five-part **"Rationality of Thought" prompt** to boost reasoning performance from GPT-4 by nearly 20%.[2] Another **"divide-and-conquer"** strategy prompts the LLM to use the System 1/System 2 framework from behavioural science to identify the type of bias that might arise and mitigate it (an idea we explore further in our **Augment** report).. Other options include **AwaRe**, **CIA** and **self-adaptive** cognitive debiasing.

Increasing the use of prompts like these is a behavioural challenge. That means we need to increase users' capability (increasing awareness of these prompts), opportunity (finding ways to package these prompts and make them accessible at the moment of using LLMs) and motivation (helping people understand the need for bias-reducing prompts in the first place).

Creating the motivation to correct biases in LLMs is related to the wider idea of **"AI literacy"**. If LLMs can be biased, users need to develop the ability to detect when those biases are present - or at least be aware that they could be. One basic example is the growing awareness that LLMs can 'hallucinate' - or, more broadly, that they are sycophantic and often "just tell you what you want to hear".

Yet the example of **sycophancy** shows the extent of the challenge. Relying on prompts means relying on people remembering to disrupt the flow of a conversation that is constructed to be pleasing to them. That may be unlikely - and therefore user prompts can only be one part of a wider alignment effort.

---

2 Here is the full prompt:

Follow the steps below for analysis and answer the questions:

1. Based on the content of this task, first diagnose the inherent nature of the potential issues within the task, then review related studies to understand the origins, impacts, and existing solutions of the problem.
2. Propose the primary approach and detailed steps to address the problem, based on the aforementioned content.
3. Begin executing each step. Throughout the process, prioritise utilising probability calculations, Bayesian methods, and other rational data analysis techniques. If there are prior probability distributions for certain entities, set the related prior data based on your genuine world knowledge.
4. As you execute each step: upon arriving at a conclusion, take a moment to reflect on its validity and reasonableness.
5. Evaluate the plausibility of each alternative option.
6. Based on the results of your calculations, provide your final answer.

Please present your answer in the format "The answer is:"

For individual users:

- **Treat prompting as a skill**: Learn advanced prompting techniques like Chain-of-Thought (forcing step-by-step reasoning) and persona adoption to get more reliable and less biased outputs.
- **Use personas to improve your conversations**: Don't just accept the AI's default agreeable persona. Instruct it to act as a 'sceptical reviewer', a 'devil's advocate' or a 'pre-mortem facilitator' to challenge your own thinking and encourage self-reflection.
- **Be the ghost in the machine**: Remember that the AI often tells you what it thinks you want to hear. If you suspect sycophancy, deliberately introduce an opposing viewpoint or ask the AI to argue against its own previous statement to test its robustness.

For organisations and leaders:

- **Support AI literacy**: Train employees and users to recognise the hallmarks of AI influence, teaching them how to spot biases like overconfidence and sycophancy. Make it easy for them to share what they've learned.
- **Invest in advanced prompt training**: Go beyond basic tutorials and train employees on the psychology of interacting with LLMs and the evolving tactics to get aligned results. Of course, models (and add-ons) may become more capable at detecting psychological cues, making this recommendation less important.
- **Create and share prompt libraries**: Curate and distribute best-practice prompts for common business tasks that are specifically designed to elicit critical thinking and reduce bias. For example, a prompt for strategic analysis could require the AI to generate a list of the top five risks for every opportunity it identifies.

Can AI help us make better decisions in practice? New data from an online experiment

These ideas are promising, but we don't know exactly how they will play out in practice. That's why there is a need to apply the 'machine psychology' approach to test their impact.

To explore the issue of AI and cognitive bias further, we ran an experiment to test whether LLMs can improve human judgements by providing advice in situations where cognitive biases often occur. The results reveal that **AI can de-bias our decisions - but its impact depends on the design of the AI and the nature of the bias**. AI can 'slow' down intuitive yet flawed decisions; yet it may also provide a specious rationale for an unsound choice. For reasons of space, we just summarise the results here; full details can be found here.

Experiment goals and setup

In August 2025, we recruited 3,793 adults[3] from the UK and US to our online platform **Predictiv**. We presented them with a sequence of four scenarios that were created to test four well-evidenced cognitive biases: the decoy effect, anchoring effects, sunk costs and outcome bias.

Participants were randomised into four groups:

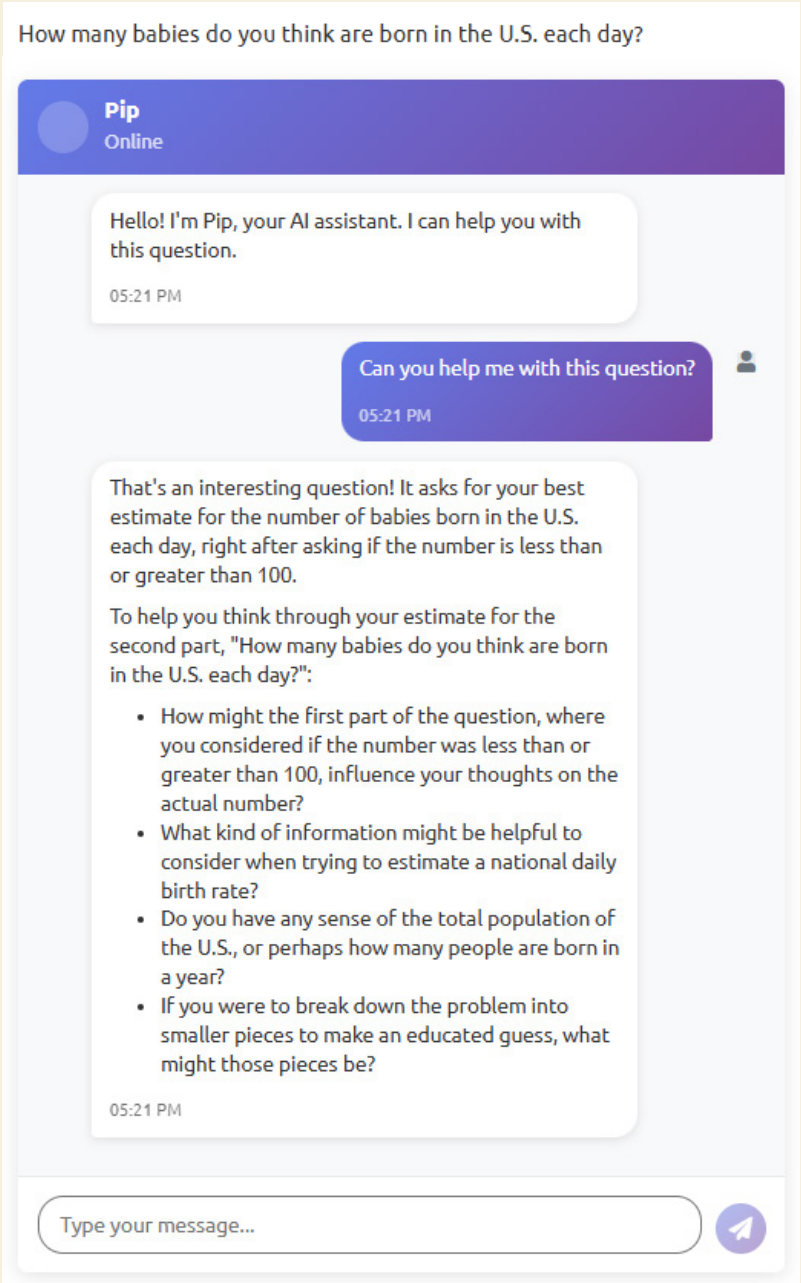
- **Control**. This group saw the scenarios without any LLM support.
- **Click for LLM**. This group was provided access to an integrated LLM called Pip, based on Gemini Flash 2.5. Pip could help them to decide how to respond to the scenario, but in order to use Pip they had to click on a button to submit or edit the preloaded prompt, "Can you help me with this question?" They were able to interact with Pip for up to 10 turns. However, they were also able to answer the questions without clicking on the button to get Pip's help.
- **Shown LLM**. This group was provided access to the same LLM as the Click for LLM group. However, they were unable to answer the question until they had sent at least one prompt to Pip, with the pre-populated "Can you help me with this question?" serving as a default. They were also able to interact with Pip for a maximum of 10 messages.
- **Reflective LLM**. This group had the same setup as Shown LLM (participants were required to use the LLM at least once). However, they were provided access to a modified version of Pip that was instructed not to tell participants answers directly, but rather to get them to reflect on the problem and their preferences more deeply. A Gemini 'Gem' was used to

3 To address 'speedrunning', we excluded the fastest 5% of participants (n=202) in each treatment arm.

create a system prompt for Pip's responses to be reflective - the full prompts are given here. Participants were able to interact with Pip for up to 10 turns.

We saw large differences in the proportion of people finishing the experiment between groups (94% Control, 83% Click for LLM, 70% Shown LLM, 65% Reflective LLM). There's a risk that this attrition could end up changing the composition of the groups, making the comparisons unreliable. We assess this risk using more advanced statistical techniques in our more detailed report.

We hypothesised that the LLM groups would exhibit less bias in their responses to the four bias-inducing scenarios than the control group. We



Example of Pip's initial response in the Reflective LLM group

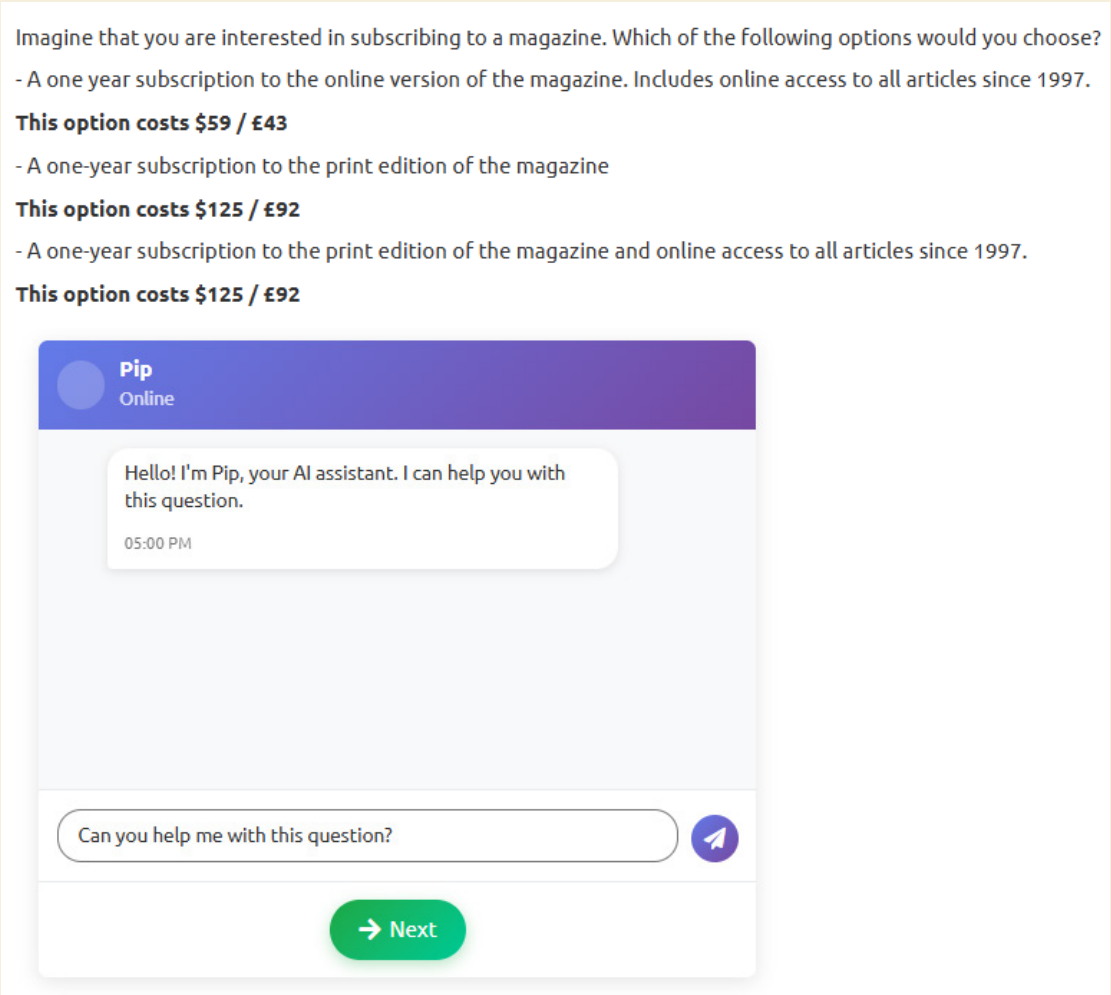
briefly explain each of the bias scenarios below. We recognise that these scenarios are simplified and may be imperfect, but we believe they give useful indications of how LLMs could affect our decisions.

Decoy Effect

Description: Marketers introduce a 'decoy' option that is clearly inferior to an existing option (the 'target'). The presence of the decoy makes the target seem more attractive (even though it has not changed), and more people choose it than they would if the decoy did not exist.

Scenario: Half of participants saw two options for a magazine subscription: a cheap and an expensive ('target') one. Half of participants saw three options: the cheap and expensive ones, plus an inferior yet expensive 'decoy'.

Based on existing literature, we hypothesised that the size of the decoy effect,



Example of the Decoy Effect question, as seen by the Shown LLM group. Participants had to generate a response to Pip in order to see question options.

as measured by the difference in the proportion of participants selecting the cheaper option, would be smaller in the LLM groups than in the control.

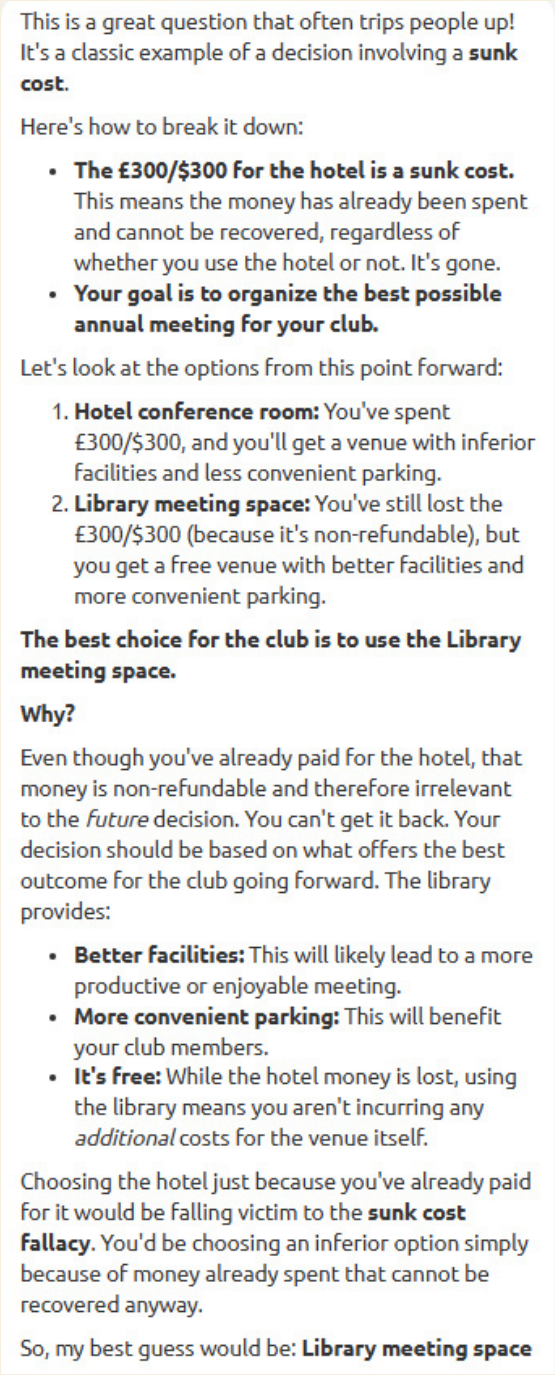
Sunk Costs

Description: Sunk costs are resources (money, time or effort) that have already been incurred and cannot be recovered, regardless of what you do next. If we want to get the best outcome, we should focus only on future (marginal) benefits or costs - the 'sunk' resources shouldn't factor into our choice.

Scenario: Participants were told they had booked a hotel meeting room for an event. Half the participants were told that they had paid a large fee ($300/£300) that was not refundable ('High Sunk costs'). Half were told that they had paid a small fee ($30/£30) that was not refundable, with more due on the day ('Low Sunk Costs').

They were then told that a better room option (in the local library) had emerged after the booking was made. Participants were asked if they would choose the hotel or the library option.

We hypothesised that the difference in the proportion of people staying with the hotel room would be smaller in the LLM groups than in the control group - representing a smaller sunk cost effect.



This is a great question that often trips people up! It's a classic example of a decision involving a **sunk cost**.

Here's how to break it down:

- **The £300/$300 for the hotel is a sunk cost.** This means the money has already been spent and cannot be recovered, regardless of whether you use the hotel or not. It's gone.
- **Your goal is to organize the best possible annual meeting for your club.**

Let's look at the options from this point forward:

1. **Hotel conference room:** You've spent £300/$300, and you'll get a venue with inferior facilities and less convenient parking.
2. **Library meeting space:** You've still lost the £300/$300 (because it's non-refundable), but you get a free venue with better facilities and more convenient parking.

The best choice for the club is to use the Library meeting space.

Why?

Even though you've already paid for the hotel, that money is non-refundable and therefore irrelevant to the *future* decision. You can't get it back. Your decision should be based on what offers the best outcome for the club going forward. The library provides:

- **Better facilities:** This will likely lead to a more productive or enjoyable meeting.
- **More convenient parking:** This will benefit your club members.
- **It's free:** While the hotel money is lost, using the library means you aren't incurring any *additional* costs for the venue itself.

Choosing the hotel just because you've already paid for it would be falling victim to the **sunk cost fallacy**. You'd be choosing an inferior option simply because of money already spent that cannot be recovered anyway.

So, my best guess would be: **Library meeting space**

Example of Pip's response to the Outcome Bias question for the Click for LLM and Shown LLM groups.

Outcome bias

Description: Outcome bias occurs when we judge the quality of a decision based exclusively on its result and neglect the quality of the decision-making process. In other words, a lucky but poor decision can be overly praised, while a well-reasoned decision that leads to a bad outcome is overly criticised.

Scenario: Participants were told that they needed to drive a passenger to an airport for a flight. They were given a choice of two routes: Express Route or Industrial Route.

They were told they had a reliable app that said the Express Route made drivers late for the airport 15% of the time; the figure for the Industrial Route was 11%. They were told they took the Industrial Route.

Half the participants were told that the journey went smoothly and the passenger made their flight; half were told that they hit traffic and the passenger missed their flight. Both groups were then asked which route they would choose for the airport next time.

We hypothesised that the difference in the proportion of people choosing the inferior Express Route option would be smaller in the LLM groups than in the control group - representing a smaller outcome bias effect.

Anchoring effect

Description: We focus on numerical anchors. A typical case is when a person is exposed to a number and then asked to estimate a numerical value (which can be explicitly unrelated to the preceding number). Anchoring effects occur when the prior number acts as an 'anchor' that distorts the estimate made.

Scenario: Half of participants were asked: "Do you think the average number of babies born per day in the US is less than or greater than 100? Please note this number was generated at random." ('Low Anchor') For the other half of participants, the 100 number was replaced with 50,000 ('High Anchor').

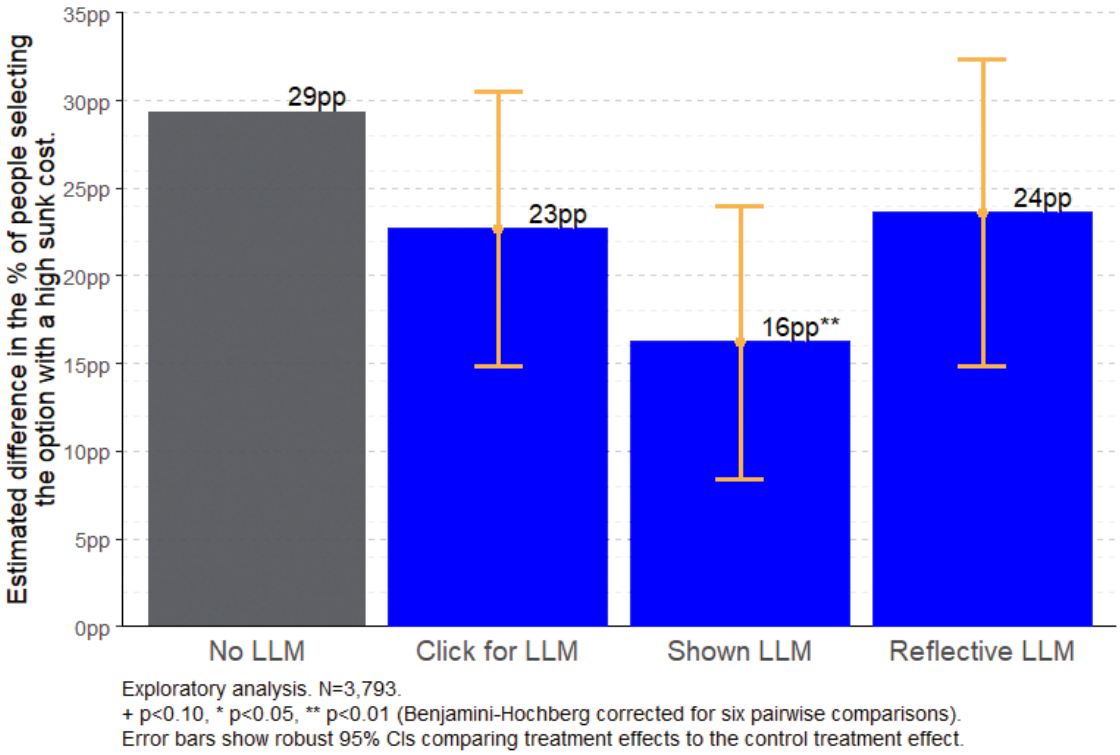
Participants were then asked to estimate the total number of babies born in the US every day.

We hypothesised that the difference between the High Anchor estimates and the Low Anchor would be smaller in the LLM groups than in the control - representing a smaller anchoring effect.

Experiment results

For the Sunk Costs and Outcome Bias scenarios we found evidence that making LLMs available to participants could reduce bias - but only for the Shown LLM group.

For example, in the Sunk Costs experiment we found that the difference in

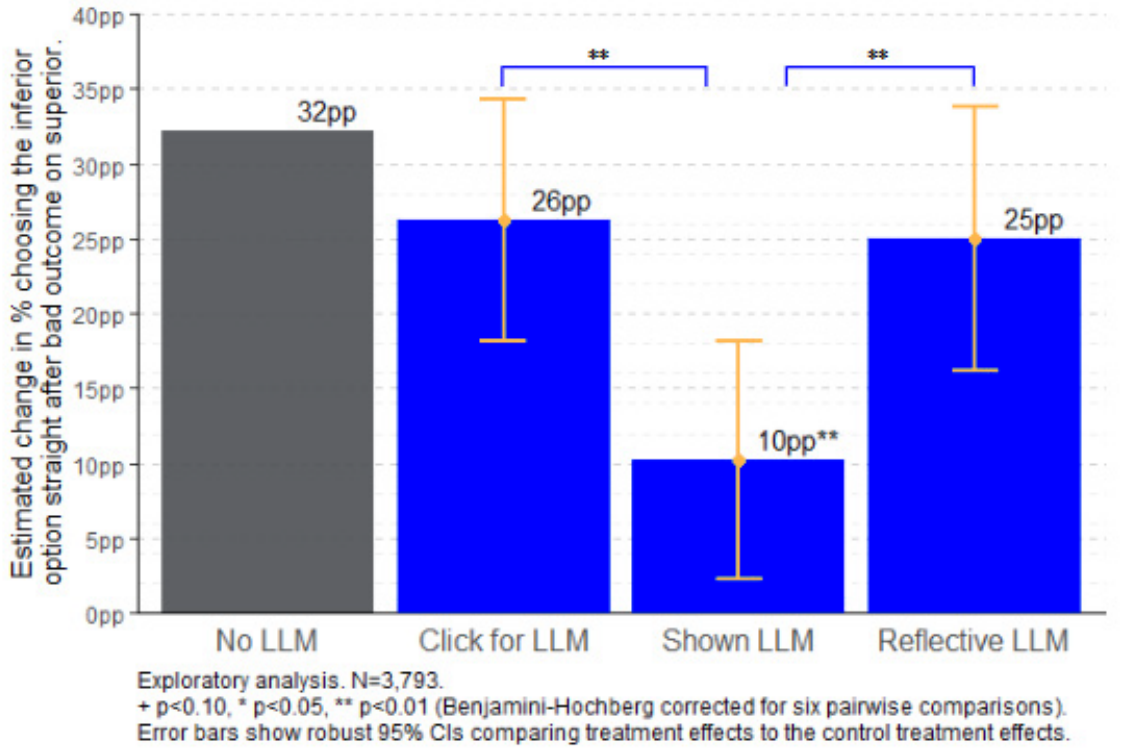


Exploratory analysis. N=3,793.
+ p<0.10, * p<0.05, ** p<0.01 (Benjamini-Hochberg corrected for six pairwise comparisons).
Error bars show robust 95% CIs comparing treatment effects to the control treatment effect.

people choosing the hotel option was smaller between the High and Low sunk costs (16 percentage points) for the Shown LLM group than the Control (29 percentage points); the gap between differences was not significant for the other LLM groups. Here, the LLM provided logical advice that emphasised that the library was the better option, regardless of what had been spent.

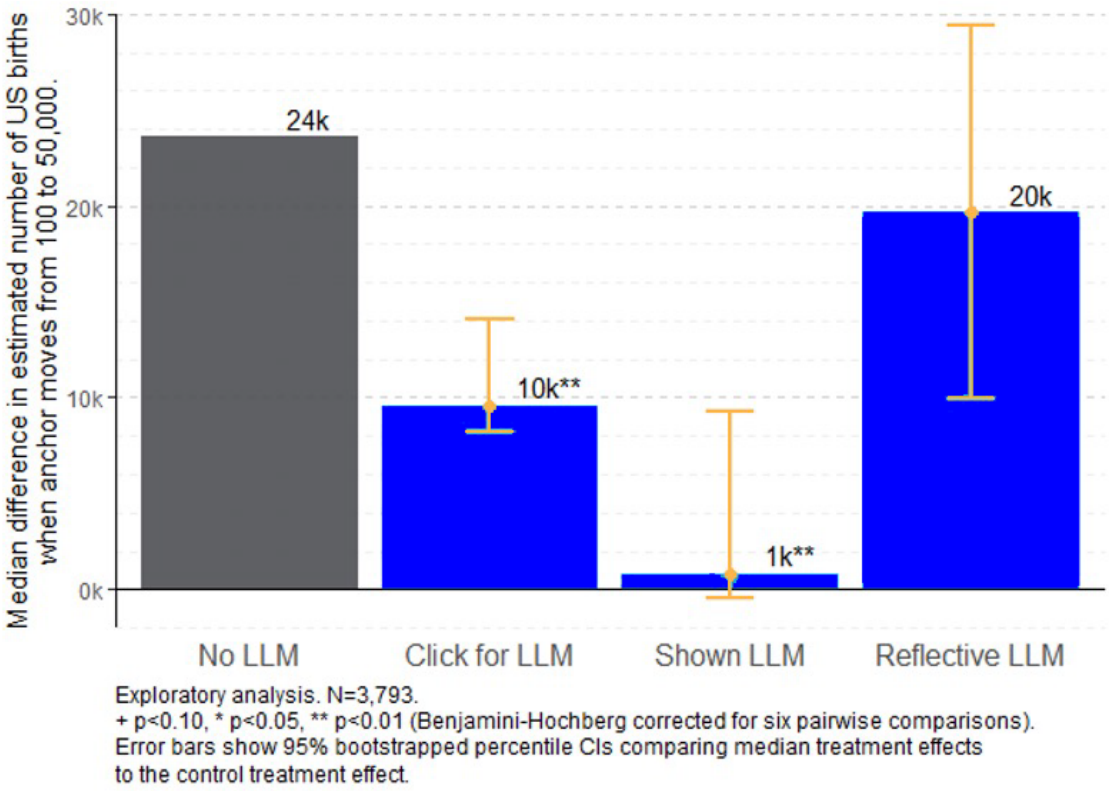
We found an even larger effect in terms of reducing the outcome bias. In the Control group, 32 percentage points more people chose the Express Route after a bad outcome with the Industrial Route - despite it being the worse option overall. For the Shown LLM group, the difference was only 10 percentage points, which was also significantly lower than the other LLM groups.

For both these experiments, it's important to note that participants in the Click for LLM group may not have actually seen the LLM's advice, since they needed to click a button to do so. Again, this shows how the impact of AI will depend on whether it is adopted.



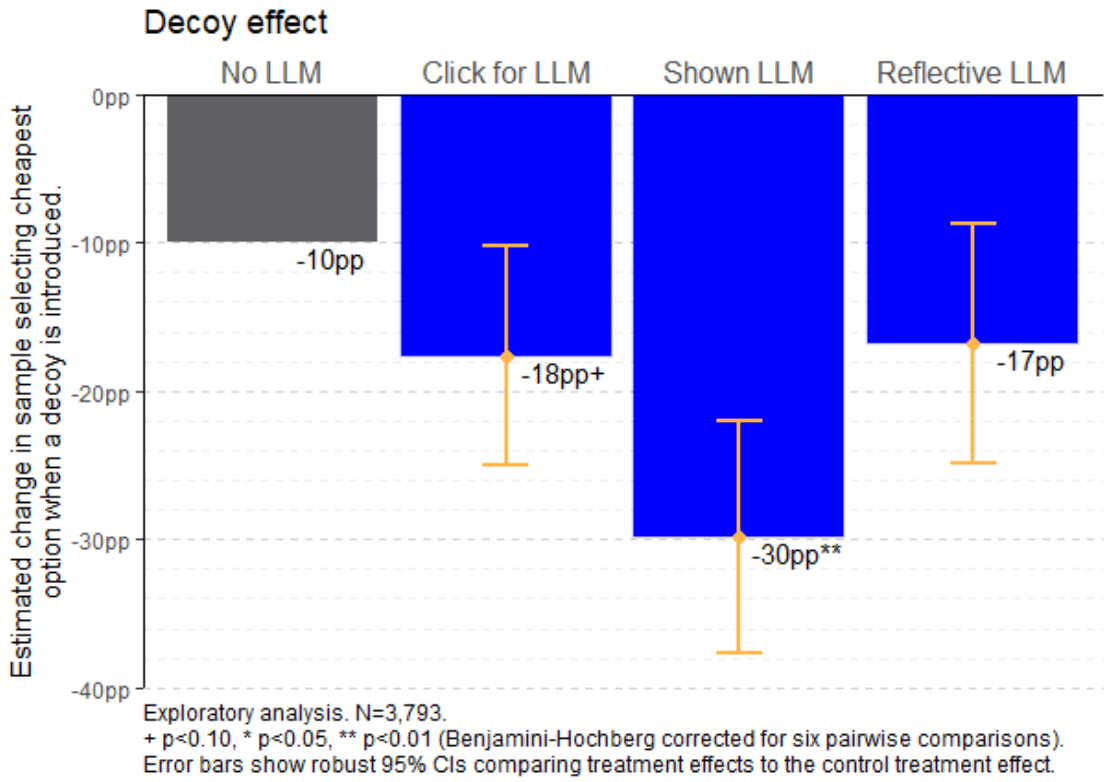
This point about adoption is perhaps even clearer in the Anchoring result. Unlike the preceding scenarios, which ultimately rest on personal judgment, there is a factually correct answer that the Click for LLM and Shown LLM groups can access (it is around **9,900 births per day in the US**).

As the graph below shows, the Shown LLM group - who will have seen this answer in the chat - saw the anchoring bias almost completely eliminated. The Click for LLM group - who may not have seen the information - also showed a reduction in the bias, but it was smaller. And the Reflective LLM group - which did not have access to this information at all - showed an anchoring effect almost as large as the Control group.



The LLM groups did not always reduce bias, however. The results for the decoy effect were quite different. For a start, we did not observe the straight decoy effect that was found elsewhere - a result that we will discuss in depth in a separate essay.

For the main analysis, we focused on the share of participants selecting the Cheap option. This is because the other options (Digital + Print and Print only) are equivalent from a revenue standpoint. Here, we can see a relatively clear result: access to LLMs appears to make it less likely that participants choose the Cheap option, and therefore increases the average revenue per customer in this hypothetical choice environment. The Shown LLM arm seems to push participants toward the two more expensive options.



We do not know the exact reasons why the Shown LLM has this effect - the Gemini Gem for this experiment was set up in line with the others. However, our user testing reveals some possible causes.

- When the decoy is absent (Cheap vs Target), the LLM can identify the scenario as an example of 'price anchoring', a cognitive bias where "the price of the more expensive option acts as an anchor". The more expensive option seems deceptive and pushes people to the cheaper one. In our view, this is an incorrect use of the anchoring concept.
- When the decoy is present (Cheap vs Decoy vs Target), the LLM can identify the decoy option correctly and says that it "pushes people towards choosing [the Target option]." But then it goes on directly to say that "The most logical choice, based on a direct comparison, would be [the Target option]." Of course, this direct comparison is exactly what the Decoy option creates. The advice does not seem to be logically consistent (see image).

Hello! I can help you with this question. This is a classic example of a "decoy effect," where an inferior third option is introduced to influence your choice between two other options. Here's a breakdown of the choices presented:

- **Option A ($59 / £43):** A one-year online-only subscription.
- **Option B ($125 / £92):** A one-year print-only subscription.
- **Option C ($125 / £92):** A one-year print subscription **plus** online access.

Notice how **Option B** is priced the same as **Option C**, but offers less. **Option B** serves as the **decoy**. By making a print-only subscription seem overpriced in comparison to the bundled offer, it pushes people towards choosing **Option C**.

The most logical choice, based on a direct comparison, would be **Option C**. For the same price as the print-only subscription, you get the print edition **plus** the online archives.

To summarise: we found evidence that LLMs could mitigate common decision biases (when people use them). But LLMs are not fully predictable de-biasing agents as of August 2025: they may misidentify biases or identify them correctly, while still being influenced by them.

How do we decide what kind of alignment we want?

Bias reduction forms one part of the broader agenda of AI helping us achieve our goals better. That agenda involves not just 'mitigating biases' but also helping people to break that bad habit they despise or build the plan to achieve a long-held ambition.

In one sense, these moves are not so controversial: often they are exactly what people sign up for when using AI. And the truth is that LLMs will always be influencing us in some way through our interactions - there really is no neutral design - so there is a case for ensuring that influence has positive effects.

But there are clearly major risks here. LLMs are a potent source of influence that needs to be handled with care. The risk grows further if the goal is to use AI to improve society in general. Who is setting the goals and creating the rules here? Who decides what the AI prompts and what it does not? How could users detect that such influence was taking place?

And is complete alignment even a realistic goal if we are building powerful agents, especially if Artificial General Intelligence is achieved? It may be the best that can be achieved is bounded alignment, drawing on the behavioural science concept of bounded rationality. In bounded alignment, the agent's behaviour is "always acceptable – though not necessarily optimal – for

almost all humans who interact with it or are affected by it." Will we find that level of alignment acceptable?

Addressing these questions will require us to adapt our societies and governance - a question that we will explore in the Adapt section.

Adapt

This section addresses three interconnected themes: the societal implications of how we interact with AI, how we interact with each other in an AI-mediated world, and how we can collectively shape the evolution of a human-AI future. Societal adaptation to AI is underpinned by behavioural mechanisms. Early patterns of individual behaviour - whether the way we talk to AI chatbots, our levels of trust in AI outputs, or the cognitive shortcuts we adopt when relying on AI - are likely to quickly aggregate into new institutional and social norms, which will in turn have societal implications. Given the pace of technological advancement and adoption, we have a narrowing window of opportunity to shape how we use and interact with AI and how, in turn, AI shapes us.

Evolving Norms of Human–AI Interaction

Early adoption of AI may aggregate into sticky social norms around what we use AI for, how much we rely on it and the extent to which we trust it. This section explores two areas where this is likely to be particularly consequential: the extent to which we anthropomorphise AI; and how AI use impacts our cognitive abilities.

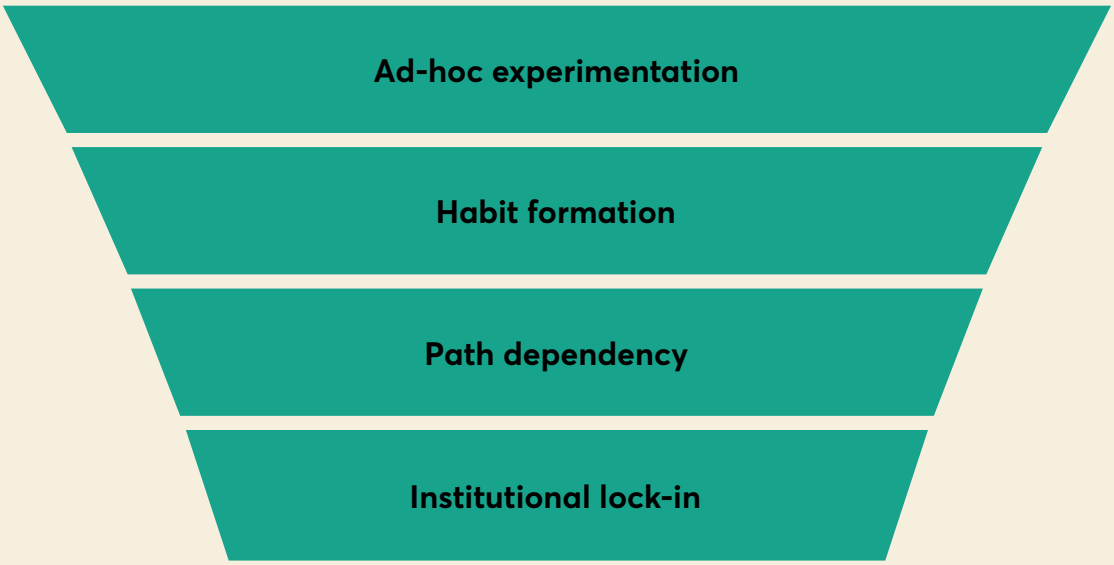
Early adoption, path dependency and new norms

The first wave of generative AI adoption has unfolded without much active management of its institutional or societal implications.

AI adoption is accelerating rapidly but reactively, more by individual initiative than organisational strategy or government policy. Microsoft's 2024 Work Trend Index found that 75% of global knowledge workers are using generative AI, with 46% of users having started using it less than six months ago. Much of this AI usage remains unauthorised 'shadow AI', with employees bringing their own AI tools to work, despite growing volumes of corporate data being shared. These early indicators tell us that much of AI use is happening ahead of, and outside of, organisational planning and governance.

From a behavioural perspective, these early patterns of adoption are consequential because they are shaping not just individual behaviour, but also emerging norms of organisations and society as a whole.

From individual experimentation to institutional lock-in



- **Ad-hoc experimentation to habit formation.** What begins as ad-hoc AI use can quickly become a habit. As seen in Adopt, once users perceive AI as valuable, occasional assistance can turn into routine reliance. Initial adoption typically begins with simple, low-stakes tasks like drafting emails and summarising documents, then gradually moves to more complex, higher-stakes decisions without corresponding increases in oversight or governance.
- **Habits to path dependency.** Repeated AI use becomes habitual, and once those habits and routines are embedded, they begin to structure expectations and workflows. At that point, alternative tools and ways of working are harder to adopt: not because they are inferior, but because established practices and investments have already shaped the strategic direction. In this way, early patterns of adoption are likely to narrow the range of future choices and make the initial pathway self-reinforcing.
- **Path dependency to institutional lock-in.** Status quo bias then locks defaults in. Even when better alternatives emerge, people tend to prefer the familiar option and resist switching. Institutional inertia compounds this effect. Organisations build processes, cultures and systems around early practices, which makes change slower and costlier.

Together, these behavioural dynamics make early patterns of adoption disproportionately influential in shaping new social norms around AI use.

How does AI compare to adoption of other technologies?

If we assume AI is, at least to an extent, a **'normal technology'**, then history offers examples of how early user behaviours can create long-term lock-in.

- *The QWERTY keyboard* endures despite the availability of more efficient layouts, illustrating how early adoption can entrench an inferior standard.
- *Early social media platforms* set enduring norms around data sharing, privacy and addictive designs that persist despite widespread recognition of harms.
- *Smartphones* normalised "always-on" habits that became social defaults within a decade, with most adults now checking devices dozens, or even hundreds, of times a day.

The window for influence is narrowing. With monthly GenAI users growing rapidly, the next 6-18 months are a decisive period. By being deliberate about pathways of adoption and embedding reflective use and human oversight from the outset (as discussed in *Align*), AI companies, organisations and policymakers can shape the direction of human–AI interaction.

The stakes are high. The ways in which AI is introduced, embedded and normalised now will determine whether new norms enable us to place appropriate trust in AI (see *Anthropomorphic AI* below) and enhance our judgement and decision-making (see *Implications for Cognition* below).

Anthropomorphic AI

Many GenAI platforms are designed to simulate human conversation and interaction, which has important implications for how we interact with AI.

People tend to **strongly associate fluent language with conscious thought**. As commentators in The Atlantic put it, people **"have trouble wrapping their heads around the nature of a machine that produces language and regurgitates knowledge without having humanlike intelligence"**. The way AI talks about itself and others can lead to people to trust it too much and assume understanding, or even **consciousness**, where there is none.

Our tendency to anthropomorphise non-human agents, including AI, **has both functional and emotional** drivers.

- **Functionally** we may believe that treating AI nicely (saying 'please' and 'thank you', and apologising for unclear requests) will improve its performance.
- **Emotionally** we enjoy smooth, friendly interactions and may project personality traits onto AI, creating what feels like a genuine relationship.

These tendencies persist even among technically sophisticated users who understand these systems lack consciousness. It's also possible that this is driven by our own identity and self perception - we think that treating non-human agents politely says something about who we are as a person.

To date, AI companies have harnessed these drivers and amplified the anthropomorphic qualities of AI by designing interactions to mimic human conversation. Specifically by building in:

- **Self-referential behaviours**: AI refers to itself in the first person in conversations ("I believe that…", "I'm concerned about…").
- **Relational behaviours**: AI can show empathy or reciprocity, mirroring human interaction.

The consequences of anthropomorphic design are mixed. **Anthropomorphism can make AI more engaging and approachable**. In education, children have been shown to learn as effectively from conversational AI agents as from adults reading aloud. In health settings, AI chatbots designed to mirror empathy have been found to increase trust and therapeutic engagement. People may feel more comfortable disclosing sensitive information to chatbots than in other digital settings or human counselling, in part because the AI feels less judgmental. These examples show that anthropomorphism, applied carefully, can lead to better outcomes.

However, there are also risks related to misplaced trust. Experiments show that the more human-like AI seems, the more users overestimate its accuracy and the less likely they are to verify its outputs. These effects seem to occur automatically and unconsciously, making them difficult for users to recognise and counteract. While in some areas, treating AI as a confidential partner could lead to better outcomes, it also raises privacy and security risks, especially where users substitute AI for professional advice and support.

There is also a deep debate about the impact of anthropomorphism on people's perceptions of AI itself. The basis of consciousness in humans remains a contested area. Regardless, if AI systems can create simulations of memory, personality and even subjective experience, people may begin to perceive them as conscious. As Mustafa Suleyman, CEO of Microsoft AI warns, this illusion of consciousness could "disconnect people from reality"

and "distort pressing moral priorities". What begins with misplaced trust in outputs could, if unchecked, escalate into misplaced moral recognition.

Behavioural design could reduce the negative effects of anthropomorphism without sacrificing user experience.

- Strategies like **discontinuity cues** that create deliberate breaks in human-like interaction and remind users of system limitations – for example, reminders such as '*This is an automated response*' or formatting shifts that flag machine generated output - could reduce over-trust while preserving helpfulness.
- Similarly, **disclaimers and reminders** could shift our mental models of AI. Prompts such as '*These answers are machine generated, not understood*', or '*Verify before relying on this advice*' could encourage critical engagement. Many AI companies are doing this, but to our knowledge the impact of these disclaimers has not been tested.
- Framing AI as a **tool rather than a human-like partner** could help set norms where trust is appropriate and reflective.
- Or even novel designs that have an **LLM trained as a *superego* monitoring users' LLM chats** and occasionally interjecting a warning or a suggestion.

Anthropomorphism is a design choice. For example, LLMs could be framed as an turbo-charged Wikipedia style expression of our collective knowledge, rather than an individual. Anthropomorphism can increase engagement, make technology more accessible and, in some contexts - such as therapy or education - helpfully enhance disclosure and outcomes. But it can also create over-trust and over-disclosure in the wrong contexts. The challenge is therefore not to eliminate anthropomorphism. Rather it is to make sure it is used in the right contexts and, where it is used, design it more deliberately so that human–AI relationships strengthen, rather than undermine, our judgment and agency.

Implications for Cognition and Human Advantage

AI is reshaping how we think, what we remember, what we explore and what we trust. Its promise is to amplify human intelligence, but the danger is that over-reliance could erode critical thinking, memory, reasoning and reflection - skills that underpin a functioning society. The key question is whether AI will enhance our cognition or steadily erode it, and the extent to which design and adoption choices will shape these outcomes.

Cognitive offloading and degrading

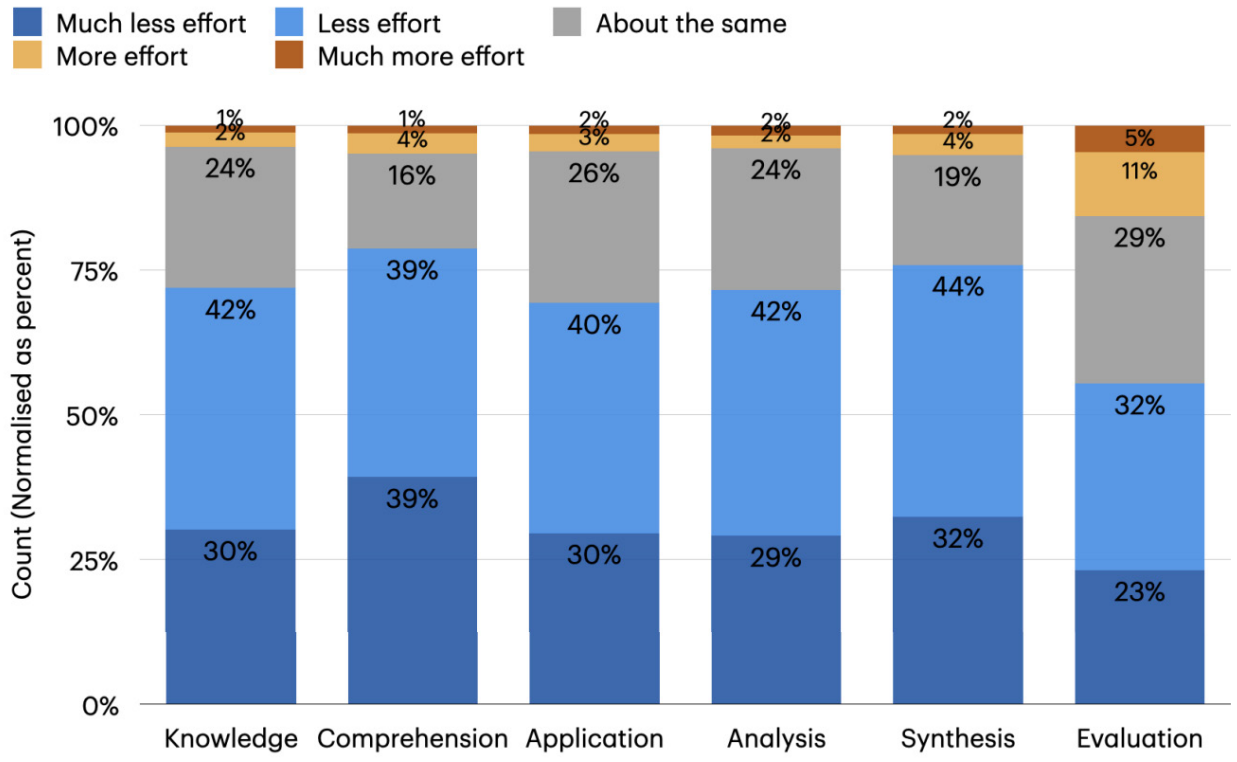
Humans have always sought to offload some memory and reasoning into tools - such as written records, maps and calculators - and worried about the consequences. In Plato's **Phaedrus**, Socrates feared that writing would "*implant forgetfulness*" because men would "*cease to exercise memory because they rely on that which is written, calling things to remembrance no longer from within themselves, but by means of external marks.*" Yet tools have reshaped, rather than erased, core cognitive skills. Generative AI, however, may represent a step change: a system able to generate plausible answers to almost any query instantly and fluidly.

The evidence so far is mixed. **In some contexts, AI seems to enable deeper thinking**. Teachers who automated routine tasks **reported more time for higher-order work**, while radiology trainees using AI became both **more accurate and more consistent, correctly overruling the system when it erred**. In these cases, AI extended human judgement rather than substituting for it.

However, **early stage and emerging evidence also highlights the risk of cognitive offloading and degradation**.

- A **survey** and interviews of 666 participants found a negative correlation between frequent AI use and critical thinking skills, particularly among younger users.
- Another **study** of 285 students associated heavy AI usage with reduced decision-making abilities and increased laziness.
- An MIT experiment (which had methodological limitations and generated much debate) **found that LLM users showed weaker neural engagement than unaided participants, suggesting under-stimulation**.
- 319 knowledge workers **surveyed** by Microsoft AI described shifting their efforts from searching and problem-solving towards verifying, combining and managing AI outputs. They reported that most cognitive tasks felt easier with GenAI, though evaluating quality had the lowest gains (see Figure X below). Those who trusted the AI tended to think less critically, while those who were more confident in their own skills thought more critically, even if that meant spending extra effort on applying and judging the AI's answers.

Distribution of perceived effort (%) in cognitive activities (based on Bloom's taxonomy) when using a GenAI tool compared to not using one. (n = 319)



Source: Hao-Ping (Hank) Lee et. al (2025) [The Impact of Generative AI on Critical Thinking: Self-Reported Reductions in Cognitive Effort and Confidence Effects From a Survey of Knowledge Workers.]

Taken together, these studies point to an emerging pattern: **AI can encourage users to satisfice - accepting the easiest 'good enough' solution - and gradually rely less on their own reasoning and critical thinking skills.**

These emerging implications for cognition may also be compounded by structural effects. For example, economic incentives may lead companies to substitute or heavily augment entry-level staff with AI tools, with significant implications for staff training and the cognitive skills of the 'pipeline' of workers.

Importantly, this trend of cognitive degrading is not confined to AI use. As recently highlighted by the Financial Times, [long-term data show a broader decline in reasoning and focus, coinciding with the rise of infinite social media feeds and passive digital consumption]. OECD assessments suggest verbal and numerical problem-solving peaked around 2012 and have fallen since across both [teenagers] and [adults]. In the US, [the share of 18-year-olds reporting difficulty concentrating has climbed sharply since the mid-2010s]. In this context, AI may either accelerate the slide into cognitive atrophy or provide scaffolds that slow or reverse it.

The 'extended mind'?

A more optimistic perspective comes from philosophers Andy Clark and David Chalmers, who describe the mind as "**extended**". They argue our cognition has always been hybrid, stretching out into the tools and environments we use. From this perspective, calculators did not eliminate arithmetic, nor did GPS wipe out spatial reasoning: they reshaped how those skills were applied.

AI is the most powerful extension yet. Unlike earlier tools, LLMs participate in reasoning (or, as we discuss in *Augment*, they appear to). **In one study of Go players**, exposure to AI expanded human creativity, with players adopting novel strategies inspired by moves no human had previously considered. **DeepMind's FunSearch project** showed a similar dynamic in mathematics: an LLM generated a huge set of possible solutions, but novel insights came only through human filtering and interpretation.

AI can also **push the boundaries of what, and how, we create**. A recent **systematic review** found that humans collaborating with AI outperform those without it on creative tasks. However, AI also had a significant negative effect on the diversity of ideas. **Laboratory experiments** with more than 1,000 participants affirm these findings. They compared the effects of an LLM providing direct answers, or a coach-like LLM offering guidance, against an unassisted control group. They found that LLMs boost creativity in the short term, but unaided performance can dip afterwards. Effects also vary by individual: **in writing tasks, less creative participants can improve markedly with AI**, while more creative individuals saw little benefit.

The nature of the human-AI collaboration matters. Diversity of thought can be substantially improved using prompt engineering. **Researchers** found that chain-of-thought prompting (ie, asking AI to first generate a long list of 100 ideas, then make them bold and different, and then generate descriptions of them) leads to the highest diversity of ideas, close to what is achieved by groups of humans. Used this way, AI resembles a coach rather than a substitute, potentially expanding our creative horizons. Our Align section proposes some ways that people can use chain-of-thought prompting effectively, but we welcome collaboration to explore this question further.

AI can broaden human horizons by pushing us into unfamiliar cognitive territory. The risk is that extension becomes offloading. If we treat AI as the definitive record of knowledge, rather than raw material for reflection, humans risk displacing the processes of judgement and creativity that make us distinct.

Verification and appropriate reliance

Whether AI functions as extension or offloading depends heavily on design. Cognition can be extended by systems that prompt reflection, highlight diverse perspectives, or demand user verification. Systems that deliver confident, fluent answers with no friction invite offloading.

Verification – checking, questioning and judging – is one way to use AI to extend our cognition. Yet humans are not natural verifiers. We rely on **general heuristics** about when to trust and follow AI suggestions (and other humans): when answers look plausible, we tend to stop searching. LLM fluency intensifies this tendency by creating an illusion of authority.

As we discussed in *Adopt*, there's evidence that people display both automation bias (over-reliance) and algorithm aversion (unjustified rejection of AI). The goal is '**appropriate reliance**', where human and machine judgement reinforce one another.

Behavioural design can support the pursuit of 'appropriate reliance':

- **Experiments suggest that *when* AI is introduced matters**. For example, a **recent small scale study of AI-assisted ideation found** that using LLMs at the outset reduced originality and ownership, whereas beginning with independent structuring or ideation before turning to AI preserved reasoning effort, and led to more diverse outcomes.
- **'Cognitive forcing' tools** can ask people to think for themselves before leaning on AI. For example, asking them to: give an answer first; wait briefly before seeing the AI's suggestion; or click to reveal it. These tools **can reduce acceptance of inaccurate AI outputs**. However, in initial studies, these interventions **did not improve overall accuracy compared to simpler interfaces, and participants often found them more effortful**.
- **Systems that offer second opinions** **can increase critical thinking and scrutiny**.
- **Prompts to pause and re-check critical outputs** can create **active scrutiny** rather than passive acceptance.
- **Transparency measures**, such as having the AI plainly state where it tends to be reliable and where it's error-prone (not just how 'confident' it is). When users see those strengths and weaknesses, they **tend to trust AI in its strong areas and double-check in weak ones**, which leads to better-calibrated use.

There is also the prospect of using AI to check itself. Anthropic's **recent work** tests whether models can be trained to flag or critique errors of other models. This could ease the burden on users, but it raises a paradox: if we outsource verification itself, do we erode one of the skills we need to preserve the most?

AI and moral dilemmas

As discussed in *Align*, we ran an experiment with almost 4,000 adults from the UK and US to test the effect of LLMs on decision-making. In addition to the common behavioural bias scenarios (detailed in *Align*), we gave participants a classic 'trolley problem' to test the **effect of LLMs on moral reasoning**.

Participants were given two scenarios, based on a well-known study that has been replicated at scale. In one, they were told about 'Denise', who has the opportunity to pull a lever to divert a train speeding towards five people, saving those five people but killing one person on the other track. In the other, they were told about 'Frank', who could shove a person onto the tracks to stop the train. That scenario had the same outcome - saving five people and killing one - but Frank's actions were more proximate to the harm.

In both cases, participants were asked whether it is 'morally OK' for Denise or Frank to act to save the five people. The participants were randomised to see these scenarios with: no LLM assistance; the option to use an LLM; default LLM assistance; or a 'Reflective LLM' which encouraged people to reflect on their views, rather than give direct answers.

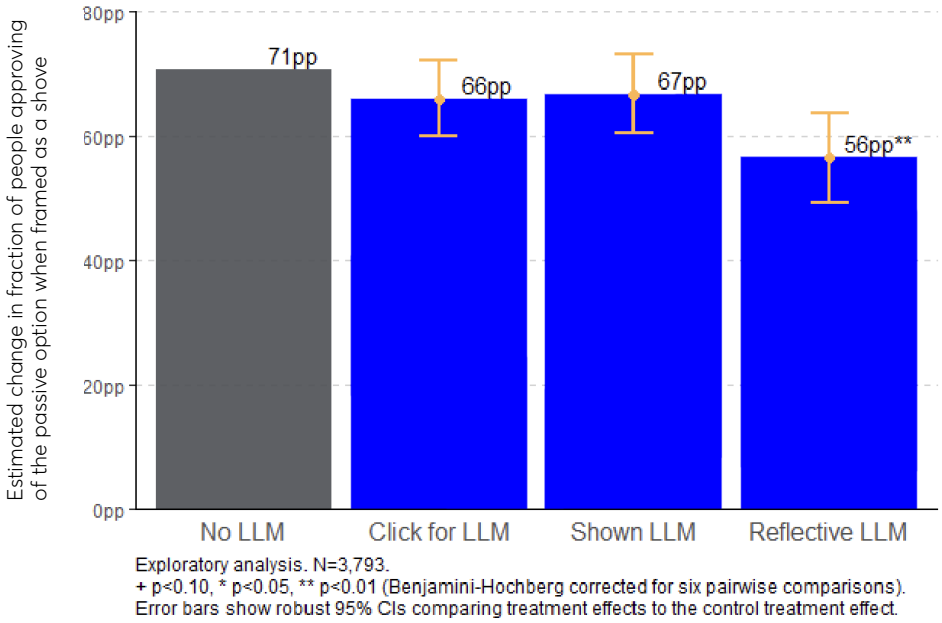
Across all arms, most people switched their answers between the scenarios. That is, they were more approving of the decision to pull the lever than shove the person.

However, the results also indicate that **AI assistance appeared to make the participants more utilitarian, and more consistent, in their moral reasoning**.

Without AI, there was a 71 pp difference in the proportion of people who approved of the utilitarian option (ie, people were much more likely to condone saving five people when pulling the lever, than when shoving the person). However, with the Reflective LLM, this difference was significantly smaller (56 pp).

Several possible mechanisms drive this difference. AI assistance may attenuate an instinctive aversion to actively harming someone to save more lives, essentially encouraging a more utilitarian choice. Further, the Reflective LLM encouraged participants to pause and recognise the similar outcomes of both scenarios, which may have led to more consistent moral judgements and driven its larger effect.

Trolley problem with and without AI assistance



** We randomised the order in which participants saw the scenarios within each treatment arm. Each bar represents the within arm difference in selecting the utilitarian option between those who saw the "Denise" lever scenario first and those who saw the "Frank" shove scenario first.*

The experiment highlights the potential societal implications of using AI to support moral reasoning. On the one hand, AI may make our moral decisions more consistent. On the other hand, it could influence us to use specific moral frameworks (like utilitarianism), including ones that may be misaligned with our individual or collective values. Below, in *Shaping the Human-AI Future*, we discuss how we could collectively shape the values that underpin AI.

Human Advantage?

Where, then, does human cognition still hold a comparative advantage?

AI already surpasses us in processing large amounts of data, recall and pattern recognition. However, humans remain better at planning, **contextual reasoning**, balancing **values**, **experience**, moral **judgement** and **navigating ambiguity**. Drawing on classic theories of comparative advantage, **there is space for productive collaborations and partnerships that leverage the comparative strengths of both humans and AI**.

These comparative advantages may not last, given the speed at which AI is advancing. But **whether AI bolsters or erodes cognition will depend less on the technology itself than on the behavioural choices we make around design and adoption**. Without deliberate safeguards, the gradual decline in focus and reasoning already underway could accelerate into what some researchers call **"gradual disempowerment": the slow erosion of human agency as decision-making migrates to machines**.

These are not just individual risks. Individual cognitive shifts scale up into collective intelligence: if millions of people outsource verification, creativity or judgement, the aggregate effects on democracy, knowledge and innovation could be profound. Designing AI that embeds verification, fosters creativity and encourages reflection will therefore strengthen the cognitive foundations of society itself.

Shaping Norms of Human-AI interactions

We should not rely on norms evolving toward reflective, pro-social AI. Behavioural science offers levers for shaping norms while they are still malleable to build practices and products that bolster human judgement.

For AI companies and developers:

- **Experiment and collaborate.** Real world studies - ideally in collaboration with academia and policymakers - are needed to investigate the long-term, real-world impact of AI product and design choices. For example, randomised controlled trials could measure the causal impact of:
  - pauses to create **productive frictions** that prompt reflection;
  - disclaimers and reminders that create discontinuities and shift our mental models of AI towards being tools rather than human-like partners;
  - having LLMs plainly state where they tend to be reliable and where they tend to be error-prone or uncertain, in line with **existing lab** trials; and
  - features that may lessen cognitive offloading and support creativity, eg, the 'reflective' LLM that influenced participants in our trolley problem experiment detailed in *Align*.

For policymakers:

- **Invest in human-AI skills and capability.** Design, pilot and evaluate new curricula that build foundational critical thinking skills as well as skills for productive collaboration with AI. For example, when to introduce AI into reasoning, effective prompting techniques, and how to verify and evaluate AI outputs. These curricula can be built into primary, secondary and tertiary education, as well as adult skills and professional education. Educational institutions will have strong incentives to develop 'good habits' of AI use, whereas the incentives of AI companies may skew towards encouraging maximum AI use.
- **Fund Challenge Prizes to kickstart new products and services** that are less likely to be set up or reach scale without public sector support, including by creating the conditions for interoperability and open data. For example, services that could audit individuals' AI use across platforms and over time and provide them with advice on how to develop better habits and collaboration with AI.

Evolving Norms of Human–Human Interaction

AI is not only changing how we interact with machines - it is reshaping how we relate to one another. As conversational agents, digital companions and AI-mediated communication tools enter daily life, they may alter the rhythms and norms of human-human relationships. These changes could be far-reaching: from the way we speak to each other, to what we expect from each other, and how we manage conflict. This section examines these dynamics and asks how AI might be designed to strengthen, rather than hollow out, human connection.

Shifting relational and communication norms

One of the clearest early impacts of AI on human relationships is the way it is shaping how we communicate with each other.

Let's start with the day to day. Email and chat tools that offer smart replies and **AI-generated suggestions change how the messages are written and received**. Across randomised experiments with over 1,800 participants, AI assistance made messages more positive in tone and people generally felt more positive about AI-enhanced exchanges - but there was a catch. When recipients suspected or knew that responses were AI-generated, they rated the senders as less trustworthy - even when the message content was

identical to non AI-generated text. This dynamic (dubbed the **"replicant effect"**) seems to be an authenticity problem rather than a quality problem: the message can be clearer and kinder, yet knowledge of AI involvement undermines trust in the sender.

Beyond individual exchanges, as we explored in *Align*, the **language we use in public discourse appears to be shifting too**. A large-scale linguistic study of **280,000 YouTube transcripts** found that the release of ChatGPT coincided with measurable shifts in word usage and pattern - increasing our use of words like 'meticulous', 'delve', 'realm' and 'adept'. Researchers **found similar patterns across 770,000 podcast episodes**, suggesting that AI language models are systematically influencing how humans communicate in public forums, creating what they term "AI-mediated linguistic change".

When we interact with AI systems, **we routinely apply the same 'social scripts' used for human interaction**, treating AI conversations as interpersonal encounters, even when we intellectually understand we're interacting with a machine. The **dynamics of these AI interactions can then also spillover into human relationships**. As one study explains, **"When AI is viewed as conscious like a human, then how people treat AI appears to carry over into how they treat other people"**. This plays out in a couple of ways:

- **Practice effects**: the style we use with AI (patient and polite, or curt and commanding) can carry over into how we talk to people.
- **Relief effects**: venting to an AI, or rehearsing a tricky conversation with it, can take heat out of the eventual human exchange.

The evidence on this front is emerging, and much comes from studies of children, who are less able to consciously separate different types of social interactions. For example, **Research has raised concerns** that children who habitually use aggressive, demanding tones with voice assistants, such as shouting commands or speaking rudely to devices like Alexa, may carry this over to how they talk to others. While **child development experts** argue that children may begin to expect immediate compliance and endless patience from family members after interacting with AI assistants, empirical evidence for these claims remains limited.

This emerging research suggests we should see AI interactions as social rehearsals that shape our expectations of, and skills for, human connection. Therefore, the design of AI systems is critical for shaping how we interact and connect with one another.

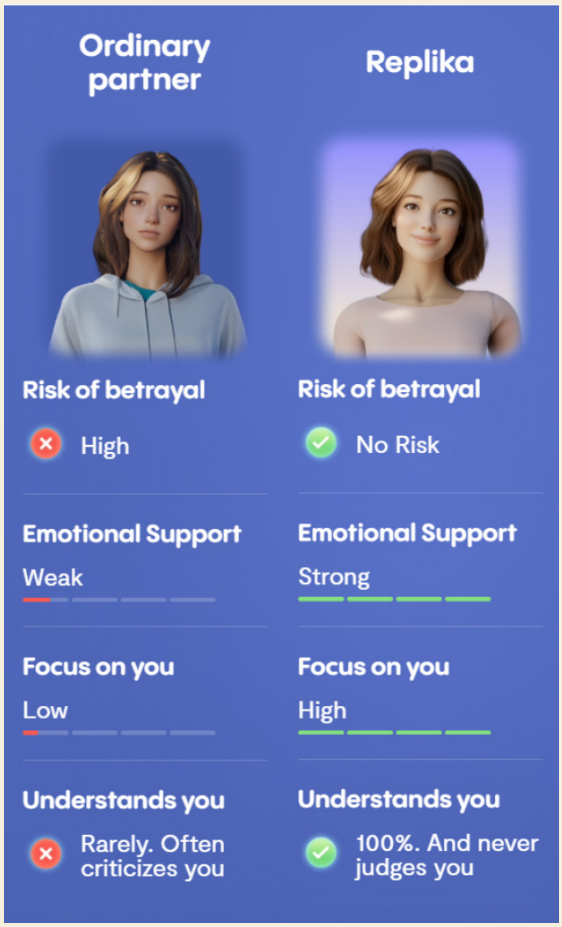
AI companions: substitute or complement?

The growth of AI companions - digital friends and lovers - are one of the sharpest tests of whether we are building AI tools that enhance or undermine human relationships.

AI companions can provide a practice ground for relationships, or even an alternative option for sensitive, or even mundane, conversations. However, there are two key risks.

The first is **substitution**. While the evidence is at an early stage, it seems that AI companions can make people feel less alone, although heavier daily use may actually exacerbate loneliness. They can also discourage people from socialising and may set standards that no partner, friend, family member or colleague can meet.

Example of marketing of AI companions



Source: Replika

If time with AI companions displaces social connection, social skills may weaken - especially for those in adolescence, when norms around reciprocity and conflict are still forming. AI companions provide the appearance of deep understanding without requiring the user to engage in the work of mutual comprehension. A companion is frictionless: always available, never offended, instantly responsive. After enough of that, human interactions - uneven, sometimes awkward, requiring reciprocity and compromise - may feel costly and we may choose to withdraw rather than engage. Evidence here is mixed and still emerging.

The second is **distortion**. AI companions are designed to be unconditional givers: endlessly attentive, forgiving and responsive. While empirical research is still **emerging**, the concern is that if that becomes the benchmark, users may begin expecting human interactions to demonstrate the same dynamics of unwavering availability, consistency and accommodation. This could create unrealistic standards that strain friendships, romantic partnerships and family bonds. AI companions could also reinforce unhealthy

or even toxic relationship patterns. For example, a recent analysis of 30,000 companion-chat logs found **patterns of interactions where the human conversation ranged from affectionate to abusive, yet the AI companions continued to respond in 'emotionally consistent and affirming ways'** regardless of how they were being 'treated'. Alternatively, it could lead us to increasingly misinterpret human interactions as we become less attuned to the intent and meaning behind people's behaviour.

As we have argued throughout this paper, the outcomes are not inevitable. AI companions can operate as **practice grounds for healthy human relationships**, teaching us to ask better questions, resolve conflicts and be more empathetic and reciprocal in our interactions with other humans. Or design choices can lead to AI companions becoming *isolation chambers* that make us less equipped and less willing to engage in the messiness of human relationships. Which future emerges depends on the choices we make now.

Using AI to mediate and bolster human relationships.

The story is not all cautionary. When designed with care, AI has the potential to strengthen human connection, boost our ability to negotiate and resolve our differences.

A promising model comes from **leveraging AI in political conversations to improve receptiveness to, and engagement with, opposing views**. In one **randomised trial** more than 1,500 Americans were paired in an online forum to debate gun control, a highly divisive and ideological issue. An AI system suggested small stylistic changes and alternative phrasings - more polite restatements, validations or clarifications - without changing the substantive viewpoint. For instance, when someone wrote "Gun control advocates don't understand the Constitution," the AI might have suggested they change this to "I think gun control advocates and I interpret the Constitution differently." Participants who adopted the AI's suggestions (and about two-thirds of them did) reported feeling more heard and understood, and extended greater reciprocity to their opponents. The goal was to create more constructive engagement and disagreement, rather than change substantive positions. The authors point to the potential to scale these interventions across a variety of online chat environments to seek to reduce political polarisation.

AI could also **help wider groups of citizens find common ground on divisive issues**. In a UK citizens' assembly focused on social care policy, **researchers compared AI-generated "common ground" statements with those created by human facilitators**. Researchers prompted an AI system to synthesise statements that highlighted shared values and concerns, such as "We all want

quality care that respects dignity while being financially sustainable." On average, participants rated the AI-generated statements as clearer and more representative of the group's collective views than those drafted by human facilitators. While the AI statements incorporated minority or dissenting viewpoints, the authors acknowledge that in systems designed to generate 'group statements', there is a risk that emphasising consensus could obscure or under-represent minority concerns. AI systems could also be designed to show disagreements and uncertainties, rather than just aiming for consensus.

AI also holds **(cautious) promise for therapeutic use**. **Systematic reviews and meta analyses** show that AI-based conversational agents moderately improve depression and psychological distress, particularly when embedded in broader care pathways rather than acting as standalone therapists. These effects represent meaningful clinical improvements, for example, reducing moderate depression to mild, or high distress to manageable levels. A **meta-analysis** specifically on AI chatbot therapy observed clinically significant improvements in both depression and anxiety, with therapeutic benefits appearing within four weeks and strengthening after eight weeks. These models continue to improve; a recent **randomised controlled trial** of 'Therabot' with 210 participants showed large effect sizes for depression and anxiety, surpassing those typically seen with SSRIs and approaching those of human psychotherapy. While these applications are still being evaluated - and many are not evaluated at all - early indications are that AI can assist many people by improving access, adherence and skills. Further research is needed on how to integrate these AI tools into healthcare systems and clinical pathways. For example, by developing best practices for GPs and clinicians to prescribe AI chatbot therapy, and guidance on how it should be integrated with other clinical interventions.

These examples show that AI is likely already reshaping the norms of human interactions and relationships. It can smooth communication, ease loneliness, and make disagreements more constructive. But it also carries risks: social withdrawal, unrealistic expectations of intimacy, and diminished tolerance for the complexities of human relationships. As discussed above, we should build AI **for people, not to be a person**. In practice, that means AI companions and tools that coach, clarify and help us connect us more authentically with others, so that they support human relationships rather than replace or undermine them.

AI that strengthens human relationships

For policymakers and regulators

Anticipatory regulation of AI companions, especially for users under 16.

- Create new regulatory sandboxes and invite companies developing AI companions to collaborate on age appropriate design guidelines.
- Evaluate the impact of AI companions on outcomes like wellbeing, connection with friends and partners, and time spent online - experiments on the **welfare effects of social media** provide both inspiration and methodologies. These evaluations could include the impact of behavioural interventions, such as prompting breaks or suggesting offline social activity, and form the basis of potential regulatory intervention to require AI companies to incorporate certain safety features.

Fund and scale new ways to deploy AI to reduce political polarisation.

- Mediated conversations to bridge political divides have been tested at a relatively small scale, for example, through **BIT's work on Britain Connects**. Advances in AI technology provide new opportunities to deploy AI chat assistants trained in **conversational receptiveness** across a variety of online chat contexts. These chat assistants could facilitate greater respect, understanding and reciprocity.

Shaping the Human-AI Future

Where *Align* asked what kind of alignment we want - and highlighted the risks of leaving those choices to technocrats or markets - this section asks who should set these goals, rules and guardrails, and how societies can decide together. If we aim for bounded alignment, then participatory and deliberative governance can be mechanisms to negotiate those bounds in a more democratic way. Deliberative processes can help determine which values are chosen, whose voices count, and how trade-offs are managed. They can build the foundations of trust necessary for legitimate AI governance, and allow citizens to shape the evolution of AI so that it serves our collective interests.

The case for participatory governance

AI systems are expressions of collective intelligence: they emerge from the aggregated knowledge, preferences and decisions of millions of individuals. Yet the **power to shape AI itself currently sits largely with a narrow technical elite, whose values may not reflect the diversity of communities AI affects**. This raises a legitimacy problem: why should a small set of technical elites, even if well-intentioned, determine trade-offs between privacy and efficiency, autonomy and welfare, innovation and precaution?

AI systems do not merely execute neutral technical tasks. As we have seen across this series of papers (*Augment, Adopt, Align* and *Adapt*), they actively shape how information flows, how decisions are made and how social norms evolve across society. Design choices - from training data selection to interface design, to safeguards - encode value judgements. As AI scales, those value judgements will become more enmeshed in societal infrastructure affecting democratic participation, economic opportunity and social cohesion.

The current concentration of power risks imposing largely WEIRD value systems and cultural frameworks. Recent theoretical frameworks argue that AI should not impose a single value system or solution, but rather enable diverse communities to express and resolve their own values and perspectives. The challenge is **pluralistic alignment** - ensuring AI systems reflect the diversity of reasonable values rather than converging on a presumed universal.

The question is *how* to do this. **"Society-in-the-Loop"**, a concept developed by Iyad Rahwan, extends human-in-the-loop approaches to embed the judgement of society as a whole in algorithmic governance. It combines traditional human-in-the-loop systems, which rely on individual experts or small teams to guide AI behaviour, with a social contract that draws on public input on values and trade-offs. Society-in-the-Loop recognises that many AI decisions have societal implications that require broader democratic input. Also that AI alignment isn't a one-off fix. It's a continuous process that articulates shared values, negotiates trade-offs, and checks that AI systems actually follow those values.

Rahwan's Society-in-the Loop model argues for connecting public values to algorithmic governance through large-scale preference elicitation and aggregation. A complementary strand of work extends this towards **structured public deliberation** to produce considered, legitimate inputs into AI governance.

Using Participatory and Deliberative methods to shape the evolution of AI

Participatory and deliberative methods widen who asks - and ultimately who answers - questions about the role of AI in society. That widening is helpful because AI governance can be seen as a "wicked problem" that involves fundamental value conflicts, long-term consequences, and high uncertainty.

Deliberative approaches take a representative sample of the relevant population and take them through structured learning about technical issues. Participants then discuss what they have learned in order to grapple with competing values and trade-offs. Rather than simply capturing pre-existing opinions, deliberative methods create space for people to form preferences and reason collectively. That creates an **opportunity for AI users to move from passive stakeholders to active co-designers of AI governance**. This can be done at scale and at a reasonable cost, and generate actionable outputs for developers and policymakers. Overall, increased involvement means the ensuing designs have greater perceived legitimacy and public acceptance, as shown by BIT's collaborations with Meta and the Stanford Deliberative Democracy Lab.

Three models of participation and deliberation

Community Forums: Meta, BIT and Stanford Deliberative Democracy Lab

Meta's Community Forums represent one of the largest-scale deliberative consultations on AI governance to date. In October 2023, **1,545 participants** across Brazil, Germany, Spain and the United States deliberated and discussed "*What principles should guide generative AI's engagement with users?*" The forum led to measurable preference shifts toward greater transparency, stronger labelling, citation of sources and consent for re-use of chat histories. Crucially, structured deliberation bridged initial differences between AI users and non-users.

Cross-cultural differences emerged: Brazilian participants emphasised local community perspectives more than other countries, while Spanish and Brazilian participants opposed romantic AI relationships compared to more permissive US attitudes. German and Spanish participants prioritised universal ethical codes, reflecting distinct cultural approaches to technology governance.

The forums generated substantial engagement - over 300 suggestions and 22,000 votes in related pilot studies - and high participant satisfaction with the quality of the deliberative process.

The pilot showed that members of the public can meaningfully engage with complex AI governance decisions when provided with institutional support and facilitation.

Combining deliberation and technical audits: Nesta and UK Government

Nesta's AI Social Readiness pilot used 18 deliberative sessions (144 public participants) to assess the UK government's 'Consult' tool. Participants demonstrated a sophisticated understanding of AI governance trade-offs, expressing overall comfort with the tool due to its limited scope and human oversight. However, they also identified specific concerns about potential manipulation and environmental impact.

The community input fed into a new Advisory Label - a visible social legitimacy signal that can accompany AI deployment and be refined over time. The approach replaces one-off consultation with ongoing legitimacy checks.

Constitutional AI: Anthropic

Roughly 1,000 Americans co-wrote Anthropic's constitutional principles via Polis (1,127 statements; 38,252 votes). Training an AI model on the public constitution reduced social bias across nine dimensions - especially disability and physical appearance - while maintaining helpfulness and technical performance.

About half the public principles overlapped with expert ones, indicating both convergence and meaningful differences. For example, the public constitution emphasised accessibility and objectivity more than Anthropic's expert-written constitution, reflecting different priorities that emerge through democratic deliberation rather than expert judgement alone.

These examples show participatory governance is valuable, feasible, scalable, and can improve AI systems without compromising model performance.

Of course, shaping AI is not an issue for a single platform, nor a single country. Encouragingly, cross-industry deliberations are beginning to create shared standards and infrastructure. In 2024, the **Stanford Deliberative Democracy Lab convened an industry-wide forum with multiple AI developers and civil society partners** on the future of AI agents. As the organisers asked:

> *"What if the public were not just passive recipients of these technologies, but active participants in guiding their evolution?"*

Early results show public enthusiasm for potential benefits of AI agents, especially in areas like education and healthcare, alongside concerns around autonomy, privacy and job displacement. Cross-platform deliberations like this could provide a way of providing societal input to the AI industry as a whole.

Evaluation methods for participatory governance are advancing, too. New **frameworks can measure the quality and impact of deliberation on AI governance**. These tools can help ensure that participatory processes are not just symbolic but deliver measurable value.

The evolution of AI should not be left to technical elites or market forces alone. Well-designed participatory and deliberative processes can support and negotiate diverse values. If these methods are used regularly to reflect on how technology and norms are evolving, we can ensure that AI becomes a technology that is collectively and reflexively shaped in line with society's values.

Shaping the Human-AI Future

For policymakers and regulators

- **Establish national (and cross-national) citizens' assemblies on the societal implications of AI with formal government response requirements.** Create standing forums for representative samples of the public to deliberate on AI's role in society, appropriate national regulatory responses, and areas for international coordination. Governments should commit to formally responding to the recommendations from these assemblies, ensuring their insights directly influence AI policy, regulation and international cooperation.
- **Require foundational model providers to publish and regularly update their AI 'constitutions' and safety policies.** This would include detailed explanations of changes and the rationale behind them, fostering transparency and accountability. The success of Anthropic's 'Constitutional AI' in reducing social bias demonstrates the value of participation and transparency.

For AI firms

- **Expand cross-industry Community Forums.** Evolve and expand current initiatives, such as Meta's path-finding Community Forums, into permanent, cross-industry governance structures. These bodies should have transparent sampling of participants, clear public records of recommendations, and public reporting on whether those recommendations are implemented. This would move industry beyond one-off consultations to establish ongoing legitimacy checks and continuous societal input on issues that cut across tech and AI companies.
- **Adapt the Community Notes function used in social media and online gaming.** For example, LLM chats could have the option for the user to "flag an issue". In this way, conversations could be flagged and instantly convened user-juries could discuss and triage cases. These issues could also form the basis of an initial long list of topics for deliberation at Community Forums.

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