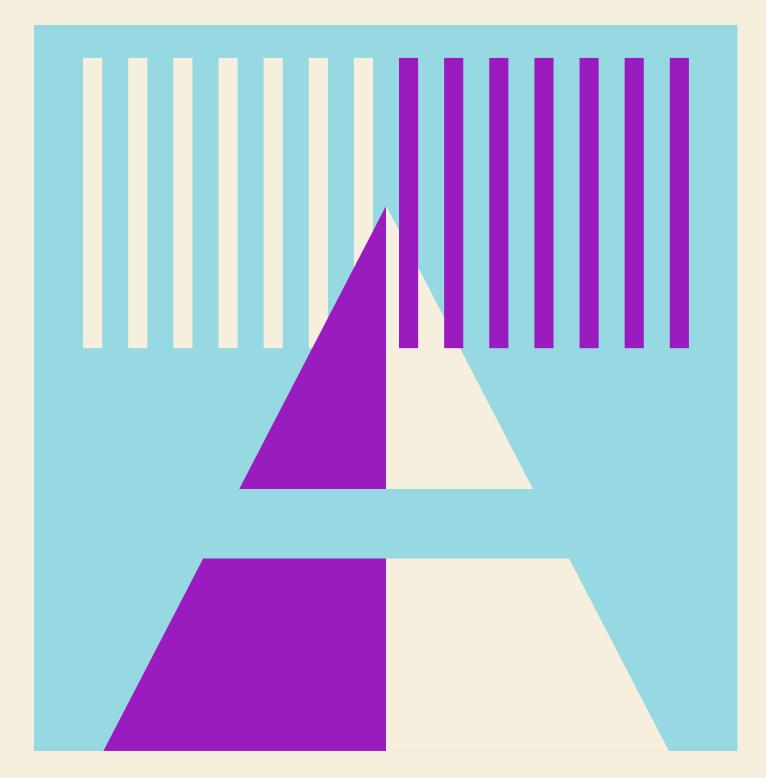


# **ALIGN**

AI & HUMAN BEHAVIOUR







# ▲ The need to understand how AI and humans interact is urgent

The rise of LLMs has created an explosion in human-Al relationships. For the first time, people can interact with Al 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 <u>"a giant real-world experiment"</u> that is creating sudden and unexpected results. An analysis of 1 million ChatGPT interaction logs shows that <u>the second most popular use of LLMs is sexual role-playing</u>. Features like <u>sycophancy</u>, <u>personalisation and unlimited content generation</u> can make LLMs addictive. Higher daily use of LLM chatbots is correlated with <u>"higher loneliness</u>, <u>dependence</u>, <u>and problematic use</u>, <u>and lower socialization</u>", although we are a long way from solid conclusions here.

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

This example shows how these powerful patterns of Al-human influence can bring both harms and benefits. The positive vision is that Al 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 Al who had been trained to play the game - moves that probably would have remained undiscovered. Other studies have shown that Al 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 <u>validate this concern</u>. For example, research from BIT has found that AI-generated financial scams were <u>more persuasive</u> 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 Al 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 Al agents "implicitly favor LLM-based Al 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 Al agent could come untethered from the human's wishes - without the human realising.

Yet the risks go much wider than deliberate attempts to manipulate. All 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 Al 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, <u>more than a third of people reported false memories when exposed to generative Al</u> - 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 Al 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-Albehaviours...

There's a growing consensus that we need <u>"machine psychology"</u>: the use of methods from behavioural science to analyse how Al behaves - in particular, how it interacts with humans. Machine psychology focuses on what Al 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 Al agents that do more things, in a wide range of environments, than chatbots can. That has led to calls for "Al agent behavioural science" that moves the focus away from an Al 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 <u>naive</u> <u>assumption</u> that "Al thinks like humans". Instead, they are helpful regardless of any differences. One reason is that Al is often designed using human roles and behaviours as guides - and humans perceive them in the light of those roles. <u>Virtual assistants are often inspired by human ones</u> 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, "Al 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.

### 

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 Al 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 Als 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.

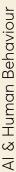
It's not yet possible to map reliably all the paths by which AI behaviours influence human ones. A massive range of factors influence human behaviours - and the psychology of generative AI is in its infancy. However, we think it is useful to think about four factors when considering how these influence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

**Valence**: How do we feel about the Al 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?<sup>1</sup>

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

<sup>1</sup> 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 Al-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 Al confidence, they see the Al as an effective tool that wants to help them. That leads them to unconsciously align with the Al (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-Al influence

#### For Al application builders & enterprises:

- → Measure what matters: Go beyond task-completion metrics. Develop methods to assess the psychological impact of your Al, such as measuring shifts in user confidence, decision-making or sentiment over time.
- → **Practise 'influence transparency'**: Where an Al 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 Al 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 Al 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 Al uses persuasive techniques:

Train employees and users to recognise the ways that Al can influence human users. Awareness that an Al'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.

#### Al 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 <u>anchoring</u> effects, <u>framing</u> effects, <u>availability bias</u>, <u>confirmation bias</u>, <u>perceptions of randomness</u>, <u>cause-and-effect</u> judgements and <u>many more</u>.

We know that this is a fast-moving field and <u>several biases have been</u> <u>eliminated</u>. 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 <u>LLMs are overconfident</u> and <u>struggle to adjust their confidence</u> based on past performance. Not only is overconfidence often seen as <u>"the most significant cognitive bias"</u>, we have seen that LLMs can <u>transmit it to humans</u>. The first step towards a solution is understanding how this transmission happens.

#### Biases get amplified in a feedback loop between humans and Al

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 <a href="makebiased judgements">make biased judgements</a> - 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 <a href="https://humans">humans</a> 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 <a href="slightly-biased">slightly-biased</a> 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 <a href="sycophantic">sycophantic</a> tendency can create a <a href="sycophantic">"chat chamber"</a>: 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 <u>LLMs may worsen present bias</u>, 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 <u>will start acting in a way that creates the exact behaviours we expect</u>.

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 <u>in both directions</u> 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-Al alignment

In the table below, we explain three main current ways of aligning humans and Al. 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 Al companies instil human values and psychological preferences into the model's behaviours after initial pretraining 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 <u>LLMs assume</u> 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 3.170B) 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 Al 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 Al 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 <u>mainly been seen as a technical issue</u>, 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 <u>stable and have been</u> <u>defined</u> before a person encounters the Al.	Our preferences can shift dramatically according to the choices available and how they are presented.	
Our <u>stated preferences reflect our revealed</u> <u>behaviours</u> - 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 <u>trade-offs</u> <u>between priorities</u> - we can order our preferences in a consistent way.	We often have conflicting preferences that we cannot reconcile easily - and may <u>vary the trade-offs we</u> <u>make</u> depending on the situation.	

Second, the approaches are <u>too static</u>: they usually just use people's initial reactions to Al. But that means they neglects how Al and humans influence each other over time, creating <u>"mutual adaptation"</u> of behaviours through repeated exchanges. An Al 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 Al behaviours <u>only emerge after multiple exchanges</u>.

One response to these issues is to re-engineer the human feedback process so that Als 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 <u>responses that support longer-term psychological well-being</u> (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 Al 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 Al'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 Al 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**. All could help people make more realistic plans by flagging that people often suffer from the **planning fallacy**: thinking that projects will get done guicker, cheaper and more smoothly than they do.
- → **Confirmation bias**. All 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, Al 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 Al as a wise partner that truly enhances human capability.



### Building the conversational context

#### For Al 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 Al 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 Al that can engage in 'collaborative metacognition'. When an Al is designed to actively reframe a user's thinking or nudge them away from a bias, be transparent about it. For example, an Al 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 Al-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 <u>even more responsive than humans</u> to classic nudges like defaults, salience effects and <u>the order of questions</u>. People can influence them using <u>established persuasion techniques</u> 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 <u>"Interesting fact: cats sleep for most of their lives"</u> 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 <a href="think carefully">think carefully</a> and avoid rapid, associative 'thought' that may create errors. The prompt is something as simple as <a href=""">"Think it</a> 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 <u>adopt a human persona</u> can super-charge the effectiveness of prompts. In a recent study, the <u>most effective prompt</u> 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%.<sup>2</sup> 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 "Al 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 Al'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 Al literacy**: Train employees and users to recognise the hallmarks of Al 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 bestpractice 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<sup>3</sup> 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:

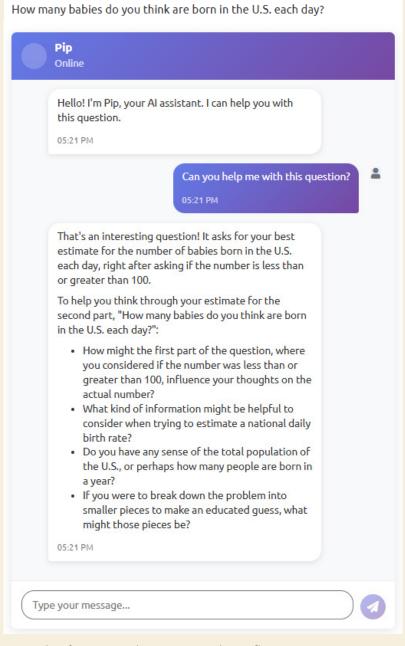
- → **Control**. This group saw the scenarios without any LLM support.
- → Click for LLM. This group was provided access to an integrated LLM called Pip, based on Gemini Flash 2.5. Pip could help them to decide how to respond to the scenario, but in order to use Pip they had to click on a button to submit 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

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



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 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.



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



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

Imagine that you are interested in subscribing to a magazine. Which of the following options would you choose?

- A one year subscription to the online version of the magazine. Includes online access to all articles since 1997.

This option costs \$59 / €43

- A one-year subscription to the print edition of the magazine

This option costs \$125 / €92

- A one-year subscription to the print edition of the magazine and online access to all articles since 1997.

This option costs \$125 / €92

Pip
Online

Hello! I'm Pip, your Al assistant. I can help you with this question.

05.00 PM

Can you help me with this question?

→ Next

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.



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

- Hotel conference room: You've spent £300/\$300, and you'll get a venue with inferior facilities and less convenient parking.
- 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').

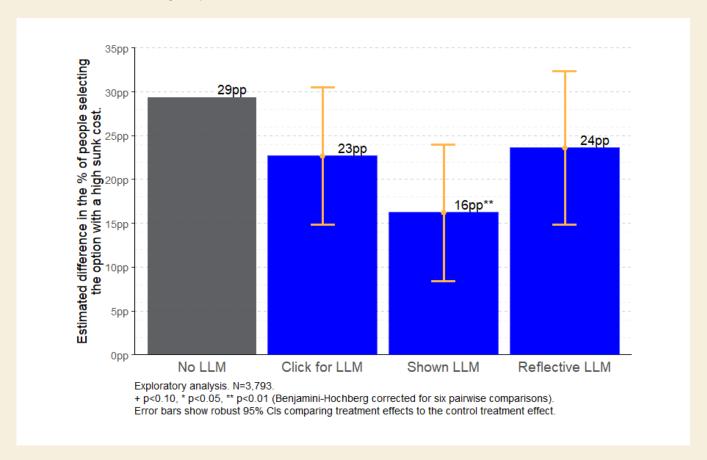
Participants were then asked to estimate the total number of babies born in the US every day.

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**

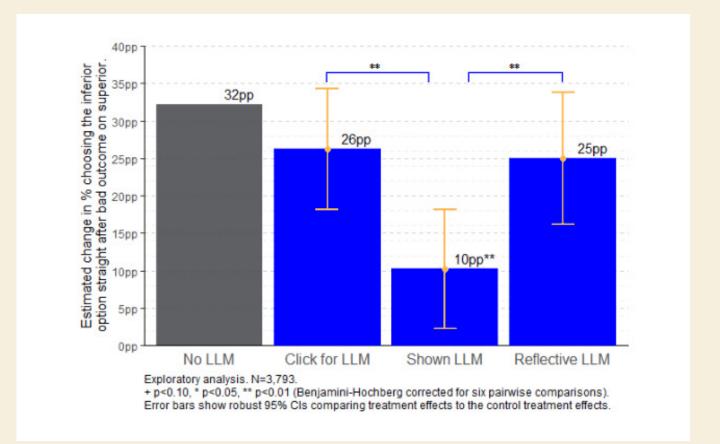
For the Sunk Costs and Outcome Bias scenarios we found evidence that making LLMs available to participants could reduce bias - but only for the Shown LLM group.



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

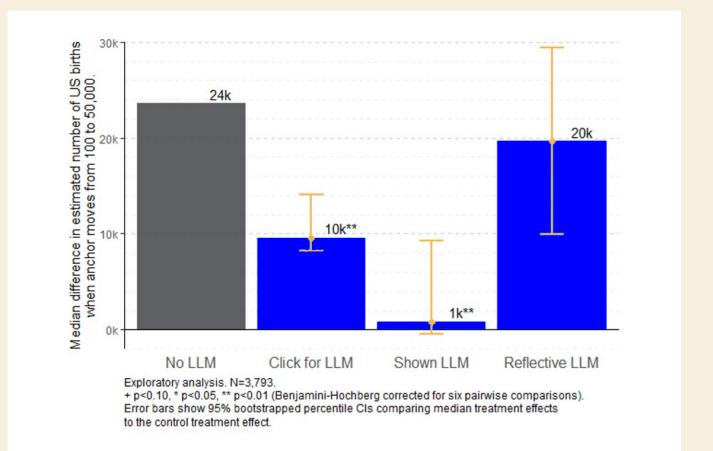
We found an even larger effect in terms of reducing the outcome bias. In the Control group, 32 percentage points more people chose the Express Route after a bad outcome with the Industrial Route - despite it being the worse option overall. For the Shown LLM group, the difference was only 10 percentage 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 Al 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 Al to <u>improve society in general</u>. Who is setting the goals and creating the rules here? Who decides what the Al 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.



#### 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