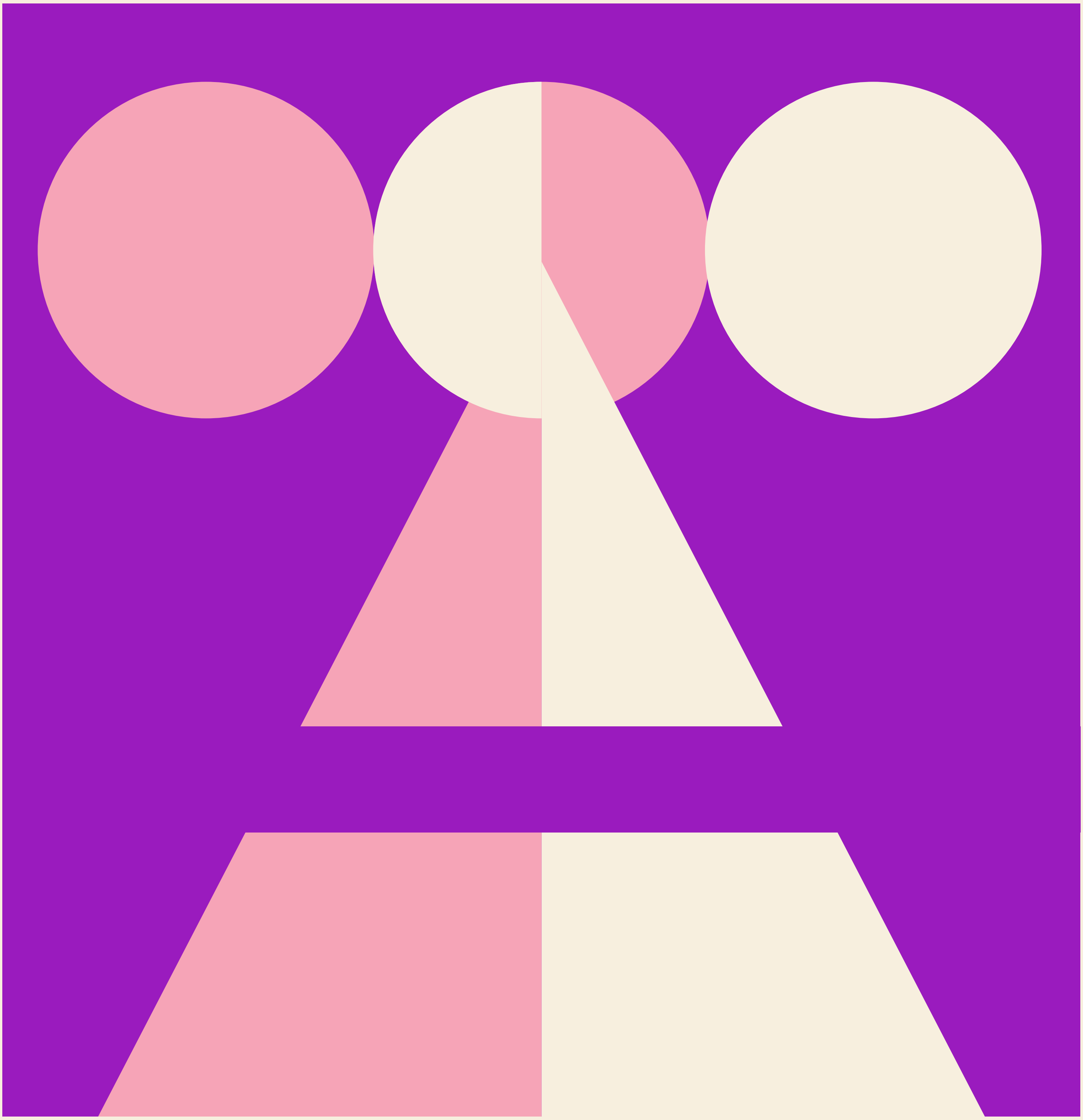
BIT

ADAPT

AI &
HUMAN
BEHAVIOUR



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.

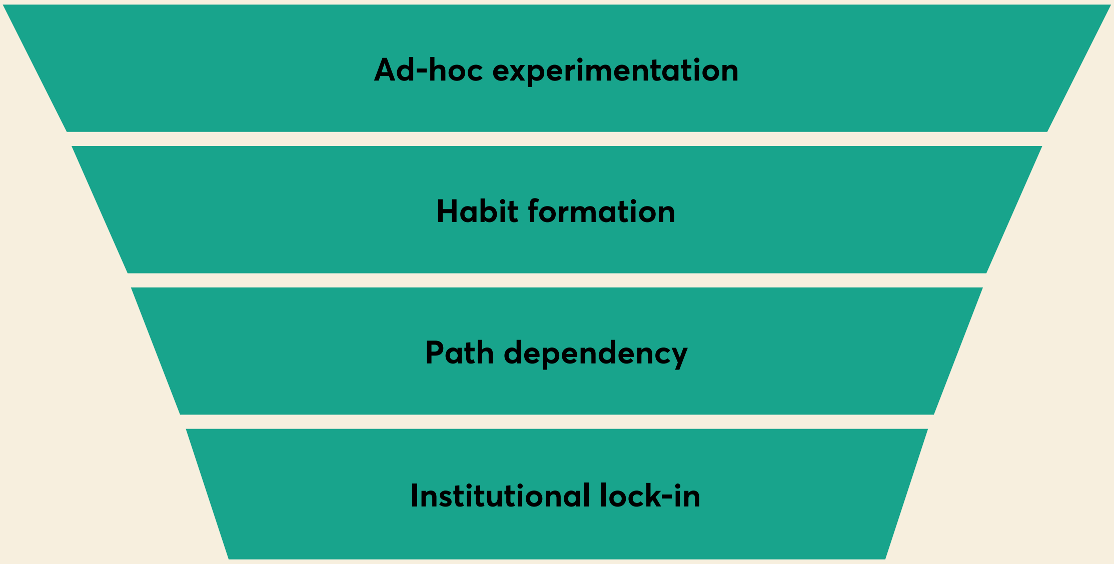
Early adoption, path dependency and new norms

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

Habit formation

Path dependency

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

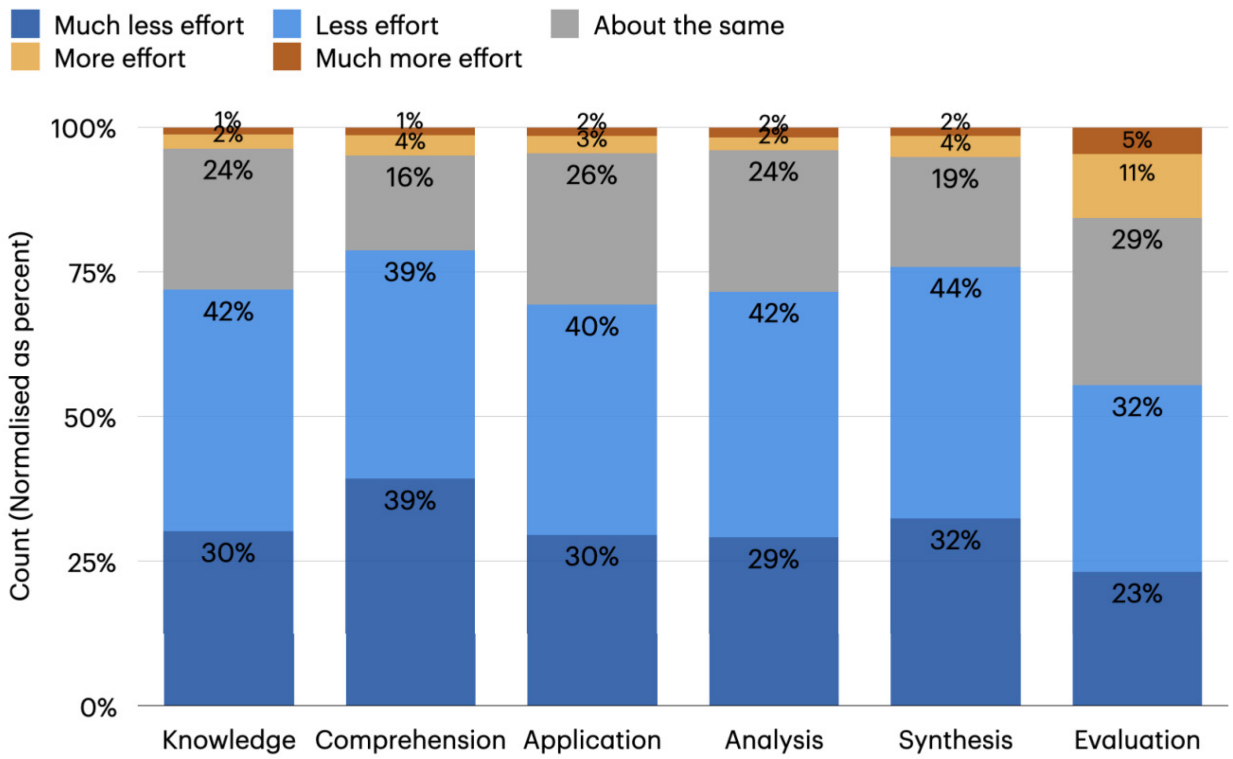
Humans have always sought to offload some memory and reasoning into tools - such as written records, maps and calculators - and worried about the consequences. In Plato's ***Phaedrus***, Socrates feared that writing would "*implant forgetfulness*" because men would "*cease to exercise memory because 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.

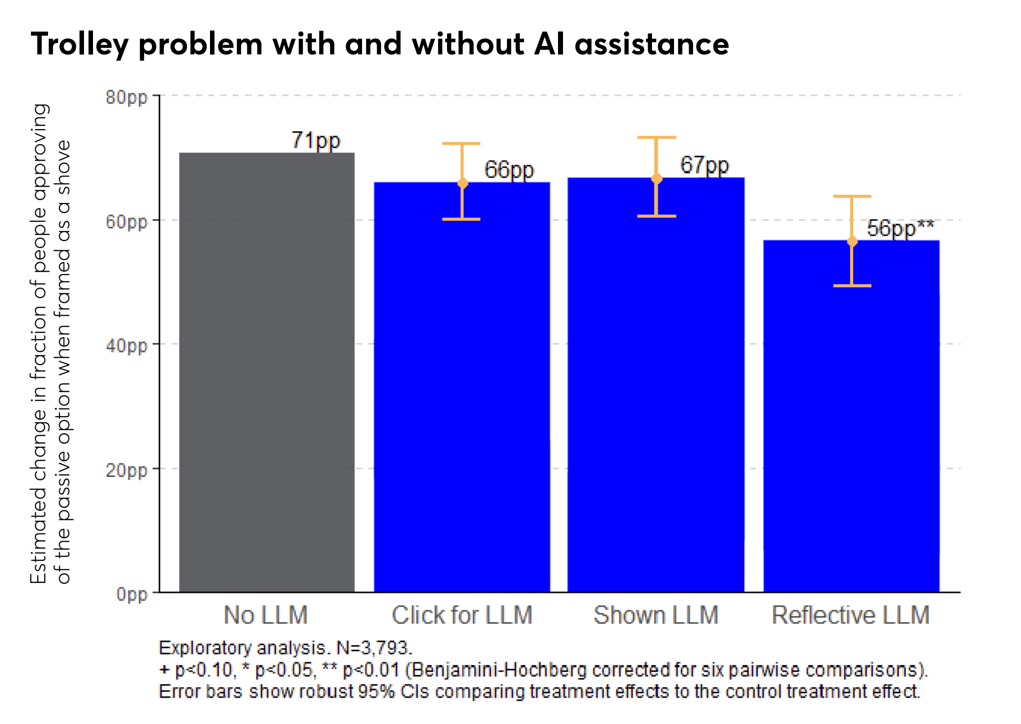
In both cases, participants were asked whether it is 'morally OK' for Denise or Frank to act to save the five people. The participants were randomised to see these scenarios with: no LLM assistance; the option to use an LLM; default LLM assistance; or a 'Reflective LLM' which encouraged people to reflect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



** 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.

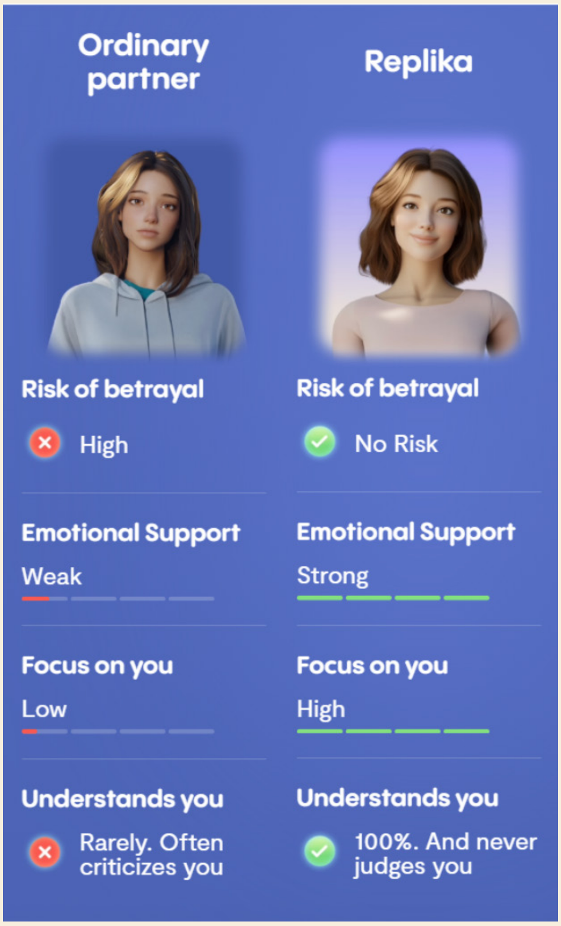
AI companions: substitute or complement?

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



Ordinary partner	Replika
Risk of betrayal ✗ High	Risk of betrayal ✓ No Risk
Emotional Support Weak	Emotional Support Strong
Focus on you Low	Focus on you High
Understands you ✗ Rarely. Often criticizes you	Understands you ✓ 100%. And never judges you

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