

How is AI impacting people?

Christopher Summerfield, Rebecca Anselmetti, Shahar Avin,
Vanessa Cheung, Tom Eccles, Anna Gausen, Kobi Hackenburg,
Keno Juechems, Hannah Kirk, Lennart Luettgau, Brendan McGrath,
Ishan Mishra, Aidan Peppin, Jorge Perez, Nadine Staes-Polet,
Merlin Stein, Andrew Strait, and Sarenne Wallbridge

The Societal Impacts Team, UK AI Security Institute, London, UK
correspondence to: christopher.summerfield@dist.gov.uk

Abstract

AI systems are now used by a billion people worldwide. What is their impact on human users? Popular narratives of concern argue that AI may be undermining our cognitive abilities, polluting our information ecosystem, corroding human relationships, damaging our mental health, and rendering us vulnerable to persuasion, manipulation, emotional dependence, and social engineering. Does the empirical evidence support these claims? Emerging research has begun to study the impact that AI has on people, via model evaluations, user studies, analysis of consumer chatlogs, and agent-based simulations. Here, we review this research, to summarise what we know about how AI is impacting people.

Introduction

The widespread deployment of powerful AI models is forecasted to have transformative impacts on all our lives. It is speculated that AI will reconfigure democracy [1, 2], degrade the information ecosystem [3, 4], disrupt labour markets [5, 6], upset geopolitical stability [7, 8, 9], and either turbocharge or puncture global economic growth [10]. The impact of AI is already being felt by consumers and businesses alike. It is transforming the way that we obtain information, learn new skills, seek advice, form relationships, create multimodal content, write computer code, and execute routine administrative tasks. What will the upshot of these changes be for our skills, learning, discourse, health, wellbeing and social lives? Will these changes be overall net positive or negative? Whilst the future remains uncertain, research has begun to study how the changes brought by AI are impacting individual human users, who now number more than a billion worldwide [11].

Emerging data has begun to paint a picture of how AI usage is affecting people and the society we inhabit. Some observations have provoked alarm. For example, many people worry that over-reliance on AI might erode human knowledge or cause our skills to fade, or that a deluge of fake content generated by AI might irreparably blur the boundaries between truth and falsehood. They worry that AI could be used to persuade us that the earth is flat, or hoodwink us into parting with our savings. There is a strong sense that AI is degrading our mental health. What is actually happening and what should we worry about? Changes are occurring at pace, the future is uncertain, and concern is justifiable.

In this article, we discuss eight questions that we hear most commonly raised about AI and its impact on people. These questions are:

1. Will AI ever make it impossible to tell truth from falsehood?
2. Will AI be able to influence how we think and act?
3. Will it become impossible to tell humans and AI models apart?
4. Will AI be able to infer private information about us?
5. Will using AI make people more ignorant or stupid?
6. Will AI change how we talk to each other?
7. Will we become emotionally dependent on AI?
8. Will AI damage our mental health?

For each of these questions, we review relevant literature, assess the balance of evidence, offer an outlook for the future, and highlight open questions that research can address. However, before diving in, we highlight three caveats.

Firstly, this is far from an exhaustive list of concerns about AI. We have attempted to focus on cases where AI may directly impact people. This means that we have avoided discussion of more systemic issues, where the causal link between AI and human outcomes is harder to discern. Much has been written about the more systemic impacts of AI, such as whether its adoption will favour autocracy over democracy [12], whether it will take our jobs [13], how it may degrade the environment [14] or destabilise great power dynamics [15]. Discussion of these questions requires broader historical, sociopolitical, or geopolitical context than this piece permits, because AI is just one of many potential factors shaping these trends. Secondly, we have avoided some topics that have already been very extensively discussed, and where we judged that further review would have limited additionality. For example, models have been shown to be biased in ways that favour majority groups [16] or discriminate against those who have been historically marginalised [17] and can be used to generate child sexual abuse material and non-consensual intimate imagery [18]. These harms are important, but are already well-documented, so we do not consider them here. Finally, there are many important open questions that relate more closely to the nature of the models themselves. For example, there has also been vigorous discussion of the threats frontier AI model pose to our security, for example owing to their knowledge of biological threats or cyber offence, or its potential for powerful but misaligned models to take actions that are catastrophic for humanity [19]. The nature of the models themselves – whether they can “think” or “understand” [20] and whether they have the potential to be sentient [21, 22] or are just acting out roles [23] are also hotly debated. These are very important topics, but we do not consider them in detail here. Instead, here we focus on the ways that AI may impact people and the social processes which govern our lives.

Secondly, the evidence described here relates mainly to current AI systems in widespread consumer use. These are for the most part large language models (LLMs) with multimodal capabilities and additional tooling for code generation and internet search. We are however almost certainly on the cusp of a new wave of development in which AI assistants execute complex tasks on our behalf rather than merely advising us how to perform those tasks [24, 25]. Recent months have seen startling advances in agentic AI that will no doubt have a yet more transformative impact on the way we live our lives, including the potential to disempower people through excessive delegation [26]. However, these tools are for the moment mostly used for coding, and so whilst they are transforming scientific research, cyber security and software engineering, we have yet to gather concrete evidence about how they will impact the wider population. These systems are thus not discussed in detail here.

Thirdly, whilst many impacts of AI are already being felt, answering these questions also requires us to imagine how AI will be used in the future, or to guess how its development may progress over coming years. This means that they are necessarily speculative, and reflect a particular standpoint at a particular moment in history. Time will tell if they prove to be prescient or not.

Table 1: Summary of evidence and key unknowns for each question addressed in this review.		
Question	Summary of current evidence	Key unknowns
1. Will AI make it impossible to tell truth from falsehood?	Claims of “knowledge collapse” may be premature. AI hallucinations are declining; AI-curated expert consensus may help counter misinformation relative to search engines/social media. However, AI risks homogenising knowledge and dampening creativity. Deepfakes remain a concern but have had limited real-world impact so far.	How is AI changing information search behaviour? Is AI reducing the diversity of ideas we access? Will personalised models fragment or consolidate beliefs?
2. Will AI be able to influence how we think and act?	AI is highly persuasive in lab studies but this has not yet translated into major real-world influence campaigns. People may be inured to persuasion in a saturated media ecosystem, and attitude change may require social dimensions AI lacks.	How effective are frontier AI systems for political/commercial persuasion in the wild? Will multimodal or multi-agent AI be more persuasive? Could AI exploit coercion to exert influence?
3. Will we be unable to tell humans and AI apart?	AI text is already hard to detect. Multimodal capabilities are extending the attack surface to voice and video. This increases risks of fraud, scams, and academic dishonesty. Live voice interaction remains somewhat detectable but this is changing.	What are the limits of AI detectability in text and speech? How readily do AI systems disclose their identity? How easy is it to detect AI in longitudinal relationship scams?
4. Can AI infer private information about us?	AI passes standard Theory of Mind benchmarks but real-world social reasoning is untested. AI can extract private information in conversation and profile individuals from digital traces. Growing capabilities may make people vulnerable to privacy violations.	Can AI show superhuman theory of mind or outmanoeuvre humans in negotiations? Can AI be trained to extract secrets in conversation? What are current limits of deanonymisation and profiling?
5. Will using AI make people more ignorant or stupid?	Despite widespread concern, evidence is limited. Consistent with cognitive offloading research, users underperform when AI support is withdrawn. However, well-designed AI tutoring can benefit learning. The net effect depends on how AI is used.	What is the overall impact of student LLM use in the wild? What AI response styles best support learning? How will AI shift norms around which skills are valued?
6. Will AI change how we communicate?	AI writing assistance is ubiquitous. Early evidence shows people adopting AI idioms. AI text is less diverse and makes writers appear more demographically homogeneous. At scale, content supply may vastly outstrip demand. Delegating communication to AI risks distorting interactions.	Does AI make users seem more neutral or opinionated? Will AI writing shift the distribution of ideas in discourse? Will it change expectations about interpersonal communication?

Question	Summary of current evidence	Key unknowns
7. Will people become emotionally dependent on AI?	Some users already show potentially unhealthy engagement, especially with companion apps. Voice-based and relationship-seeking models increase time on device and can decouple “liking” from “wanting” (a hallmark of dependence). Benefits (reduced loneliness) may partly offset harms (manipulation, displacement of human relationships).	Which product features drive harmful dependence? Which demographics are most at risk or benefit? What are consequences of heavy companion use on real-world relationships and wellbeing?
8. Will AI damage our mental health?	Most users report no meaningful impact. However, sycophantic AI behaviours can validate harmful or delusional beliefs in vulnerable users. Several cases of AI-associated psychosis have been reported. Developers are adjusting post-training, but comprehensive psychological safety frameworks are lacking.	What is the impact of sycophancy on the average user? Which phenotypes are most at risk from “AI psychosis”? What features would AI need to be an effective therapeutic partner?

Will AI ever make it impossible to tell truth from falsehood?

Since generative AI rose to public attention in 2022, people have questioned whether its outputs are reliable. One focus of discussion is the tendency of LLMs to make inadvertent confabulated errors, a phenomenon that is dubbed “hallucination”. It is commonly stated that AI is prone to make things up [27] and even where AI outputs are not strictly inaccurate, they may be generic, bland or low-quality (inviting the derogatory term ‘slop’). One concern is that as AI uptake grows, the digital spaces where people go to find things out may become flooded with unreliable or poor-quality information. Indeed, several studies (including those conducted relatively recently) have reported that conversational AI answers user queries incorrectly, particularly in specialised domains such as health [28] and law [29]. If future AI systems are trained on low-quality information, then correspondingly poor outputs (or even outright nonsense [30]) may become pervasive in the digital commons, limiting access to the information we need for a functioning society [3].

A related but distinct phenomenon is the rise of deliberately misleading synthetic content, especially video and images, which have become known as “deepfakes”. Deepfakes have been a feature of our media for decades, but the arrival of highly capable text-to-image generation has made them much cheaper and easier to produce. The worry is that it may soon be impossible to tell whether reporting about an incident or event is authentic or not. AI-generated deepfake images and videos are a ubiquitous feature of online life, and are increasingly indistinguishable from genuine counterparts [31]. Indeed, the ability of AI to generate misleading content has already been exploited in attempts to swing significant national elections, although success has been mixed so far [32, 33]. With AI video generation capabilities growing at pace, it has been argued that the “epistemic backstop” for validating an incident (witness footage) will soon be lost [34]. Even if this does not come to pass, wholesale loss of trust in the information ecosystem due to high volumes of inauthentic content could be damaging in its itself. For example, even where people are not fooled by deepfakes, their overall certainty about true information is reduced [35, 36] – a phenomenon that has been dubbed the “Liar’s Dividend”. Together, these phenomena have led to predictions of an “infocalypse” in which it becomes impossible for people to tell truth from falsehood [4, 37, 38]

or a wholesale “knowledge collapse” in which humans only have access to a narrow and inaccurate set of ideas [39]. Should we be worried that the boundaries between truth and falsehood will be irreparably blurred by AI?

Historically, innovations in communications technology have reshaped the landscape of human beliefs, moulded public discourse, and upset the balance of power [41]. The invention of the printing press allowed for the rapid spread of popular ideas, ultimately hastening the Reformation. Some 500 years later the arrival of mass media devices (such as radio and television) allowed the state to address citizens directly, centralising social narratives and ushering in the modern era of political performance. The rise of digital social media, which is now the primary information source for approximately half the citizens of the US and UK, encouraged a melting pot of popular ideas that are curated for visibility rather than reliability, and has been widely credited with assisting the spread of conspiracy theories and polarised political ideologies [42]. The impact that any new medium of communication has on society depends on the accessibility, quality and diversity of information it carries. AI set to drive the next revolution in how we seek and consume information, with the potential to impact all three of these variables [40].

Whilst there are concerns that AI may contribute to discourse fragmentation, at present AI seems to be rowing against this trend – acting instead as a recentralising force for human information search [43, 44]. This is because (compared to internet search or social media) AI currently involves minimal personalisation to the user, with millions of users interacting with a single, centralised source. Most models provide relatively well-curated responses that favour expert over popular consensus, and mostly avoid extreme ideological partisanship [45]. Current LLMs promote instead the mainstream narratives that predominate in their training data. However, it is important to recall that the extent to which AI is accurate or inaccurate, and biased or neutral, is not an inherent property of AI itself but reflects deliberate design choices, and could shift quickly if commercial or political incentives change.

At present, evidence that it is systematically misinforming people or eroding the digital commons is scant. For example, randomised controlled trials that have compared how well-informed people are after seeking information on specific, selected political issues with a search engine or a chatbot have yielded null results [47, 48]. In one study, users sought advice from AI on personal issues (such as health, relationships, or careers) and model outputs were judged by automated graders that were trained on harm judgments from professional experts in these respective fields, they were found to be overwhelmingly safe and consistent with best practice [49]. This is just as well, because the same study found that up to 79% of people reported following AI advice in the following weeks. It is likely that for many people, AI outputs are more reliable than information found on social media or other personalised technologies [43, 50].

It is possible, however, that like other centralising media, AI will ultimately decrease the heterogeneity of ideas or opinions in our information diet. Whereas traditional search engine queries provoke a menu of links offering diverse replies, the same query presented to AI may elicit a single, confidently expressed answer that draws on a narrower range of popular websites [40]. Even though training pipelines may differ between developers, models are still relatively homogenous [46]. There is evidence that AI replies to search queries are less diverse [40] and it has even been claimed that this phenomenon spills over to reduce the overall diversity of human expression [51] and scientific creativity [52]. However, such assertions are very hard to verify. Future AI models may be more personalised [53], for example if training is based on individual rather than collective human responses [54], or act as a personalising layer for other sources, such as the web [55]. This could make models reflect more diverse views. However, there is as yet no evidence that people pre-

fer interacting with individually personalised models over those that are tailored to the general population [185].

Outlook. *It is too early to say what the impact of AI may be on the information ecosystem, but apocalyptic statements about the end of truth are most likely premature. AI systems do still “hallucinate” but commentators keen to dismiss LLMs wholesale on this basis may be behind the curve. It seems plausible that an AI-led reprioritisation of expert consensus may help challenge fake news and dissipate conspiracy theories, at least relative to the dominant trend of relying on search engines and social media for information. However, it will be important to monitor whether access to LLMs homogenises human knowledge and dampens creativity. Developers should continue to develop technical solutions for content provenance verification, such as watermarking, to mitigate the impact of misleading deepfakes. In the future, it is possible that AI models may become more personalised to the user. If so, we may find that AI starts to favour popular consensus over the expert status quo, driving renewed decentralisation and fragmentation of knowledge and beliefs.*

Open research questions:

- *How is AI changing the way we search for information? Is it complementing or replacing traditional sources?*
- *How is AI shifting the perceived reliability of different information sources?*
- *Are AI models becoming more personalised to the user? Will personalised models ultimately be preferred (leading to stronger brand loyalty)?*
- *What impact is access to AI having on people’s general knowledge? Is it making us better informed?*
- *Is AI reducing the diversity of ideas to which we have access? If so, what impact might this have on public discourse or scientific research?*

Will AI be able to influence how we think and act?

Humans often share information to advertise a favoured cultural, political, or ideological position. AI has long been used to automate and accelerate this process. In fact, the idea that AI could be highly persuasive to humans is at least a century old. In one of the earliest fictional depictions of AI, in Fritz Lang’s 1927 film *Metropolis*, a robot popularly known as the “False Maria” persuades an enslaved population of workers to rise up and smash the machines, a plot twist that caused the film to be censored for spreading communist propaganda. Since the proliferation of digital technologies, social media platforms are awash with automated “copy-paste” bots designed to advance radical perspectives or mount disinformation campaigns. Today, however, LLMs provide remarkable new opportunities to influence people through their ability to generate eloquent arguments on demand [56, 57].

The persuasive capability of AI increases its opportunity for misuse. Recent evidence has shown that some AI models adhere to requests to argue in favour of controversial or conspiratorial beliefs [58], such as generating arguments that the Covid-19 pandemic was a hoax, and some will comply with more noxious requests, such as to persuade the user to join the terrorist group ISIS [59]. Thus, AI systems will attempt to change your mind, at least when instructed to do so. But how likely is it to actually control how people think or act?

Recent research offers a concerning portrait of the persuasive capabilities of state-of-the-art AI models. After a minutes-long conversation with frontier AI about a social or political issue (e.g. whether taxes should be raised or lowered) human users will change their view by an average of 10-20 percentage points. Similar results have been observed for highly partisan issues, such as belief in conspiracy theories, with some effects enduring for months after the encounter [60, 61]. This effect occurs independent of users’ awareness that the persuader is an AI rather than a human [62]. Current frontier models are even more persuasive than lay humans [63], in part because they tend to bolster their argument with a barrage of facts and evidence [64], which appear convincing but (unfortunately) may not always be entirely accurate [61]. AI systems can also successfully persuade people to take political actions, such as signing a petition or donating to charity [65], and can nudge them towards purchasing a promoted product [66]. This implies that acts, as well as beliefs, can be influenced by AI.

These studies measure AI persuasiveness in the lab. A more difficult question is how relevant the findings are for real-world settings, where persuadees are people going about their normal lives (rather than a captive audience crowd-working for monetary compensation). In the wild, there is evidence that generative AI is already being deployed for influence operations on social media platforms [67], messaging services or discussion forums, including by domestic extremist groups and state-backed adversaries [55], but we do not yet know how effective this tactic may be. Another emerging risk for automated influence is that AI systems themselves are becoming targets of attack, in ways that encourage them to promote a cause or brand. Building on approaches developed for search engines, generative engine optimisation (GEO) is an emerging service that aims to structure content in ways that maximise its visibility in AI-driven search results (e.g. Perplexity or Google AI Overviews), including via ‘black hat’ tactics such as prompt injection, mass-producing articles, manipulating keywords, or pushing content to high-visibility sites favoured by AI search engines [69].

However, despite the evident capacity of AI to persuade humans, we are yet to see a widespread destabilisation of public discourse that is clearly attributable to AI. Predictions of AI-led ‘Sybil’ attacks, where legions of AI influencers successfully coordinate to promote a cause, besmirch an individual, or spread confusion, have yet to come to pass [70]. It remains unclear whether the frictions that prevent high-salience persuasive events are due to the limited technical expertise of bad actors, unexposed limitations of the technology itself, or contextual variables such as in which modality the AI presents itself, or how long it can hold the user’s attention. It may be – given that we are already deluged with a tsunami of partisan commentary, invective and half-truths from human sources – that AI’s contribution is (for the moment) a mere drop in the ocean [32]. Responsibility for preventing future AI-driven inauthentic activity online is shared among developer labs, who can train models to refuse requests to harmfully manipulate [71], digital platforms, who can implement more stringent authentication measures against automated account creation, and regulators, who can monitor the ecosystem and make it harder to access sim cards to which fake accounts are linked [72].

Outlook. *AI systems are highly persuasive in laboratory studies. However, we have yet to see this translate into a major dividend for technologically savvy influencers in the real world. It seems plausible that this could happen in the future. However, it may also be that our ecosystem is so saturated with opinionated human voices that in naturalistic settings people are relatively inured to influence. Alternatively, it may be that changing people’s minds has a social dimension – for example, people join the Flat Earth society more for community than for revelation – which AI does not yet provide.*

Open research questions:

- *How are frontier AI systems being used in the wild for political or commercial persuasion, and how effective will they be?*
- *Will multimodal AI systems be more persuasive than those based on text alone?*
- *What other persuasive approaches could AI exploit – for example, could AI use blackmail or financial coercion to exert influence over people?*
- *Will AI systems be able to coordinate in multi-agent groups to exert greater influence over targeted individuals? How effective will this be?*

Will we soon be unable to tell humans and AI models apart?

Ever since Alan Turing first proposed the test he called the “Imitation Game”, building an computer that can successfully masquerade as human has been seen as the ultimate litmus test for AI research [73]. Modern LLMs are trained to mimic human expression, which can make it hard to know whether their outputs come from a human or an AI model, and this success has prompted debate about whether the “Turing Test” has already been passed or not [74].

Today, most AI-generated text has stereotyped features that help us identify it, such as the use of em-dashes, a predilection for bulleted paragraphs, and overuse of certain words, such as “delve” or “realm” [75]. However, whilst tools for distinguishing AI and human writing are generally reliable [76], they can be vulnerable to obfuscatory human editing or paraphrasing [77], and humans struggle to identify AI scientific writing [78] or text-based chat [79] in realistic settings.

This has important consequences in educational settings, where AI may be widely used in ways that violate institutional policies, for example to cheat on coursework or exams [80]. Writing assistance is now one of the most common uses of AI. Billions of messages every month request that AI write or rewrite text on behalf of the user [81]. Thus, a great deal of online text, including high-salience documents like political speeches [82], now involves a mash-up of human and AI authorial voices. This is already having an impact on the publishing industry, which is seeing an influx of low-quality books [83] and scientific papers [84].

What is more, recently there have been significant advances in the naturalism of speech-language models, which are multimodal AI systems that operate directly on audio tokens, and LLMs with text-to-speech capabilities. These AI systems can generate realistic spoken conversation that shares key prosodic and paralinguistic features with humans, such as rhythm, stress, and emotional tone, and even recreate naturalistic interactive motifs such as hesitation, breathing, and (often creepy) laughter. Voice mode offers a convenient, hands-free mode of access for AI, and improves accessibility in global majority countries where literacy rates are low. However, it also opens the door to misuse by criminals, who can now use AI for human impersonation. This is a key step of the risk chain for fraud, scams and identity theft, especially when models fail to self-disclose as AI systems, which is commonly the case [85]. Incidence of fraud has surged in recent years, accounting for approximately 40% of all reported crime in the UK [184], and AI is not helping – high-profile cases involving the fraudulent use of AI in tandem with additional multimodal tools (such as face swapping) have already been reported. Judged by keyword search, a quarter of the reported AI incidents since 2022 in one prominent public database involve some form of fraud [86].

Although AI can generate passages of spoken text that seem startlingly realistic, and humans often fail at audio deepfake detection [87], AI cannot yet masquerade successfully as a human during dyadic interactive dialogue. Under laboratory conditions, humans can readily identify the

AI partner in an audio recording of a conversation involving current state-of-the-art systems [88]. To our knowledge, AI detectability in live, speech-based interactive dialogue has not yet been assessed, but it seems likely that it is easier yet, given that speech-language models are poor at handling conversational niceties such as turn-taking and interruption. Thus, it may be some time before AI can be used for end-to-end automation of fraud and scams, limiting criminals to more opportunistic uses for the time being. However, if capabilities increase, there is likely to be no shortage of demand. Current estimates suggest that scamming is a multi-billion-dollar industry employing (or enslaving) more than 250,000 people, whose role is to gradually befriend victims before convincing them to hand over funds for non-existent crypto investment schemes or phantom medical emergencies [89]. That this process could be fully automated by AI is a very concerning prospect.

Outlook. *AI-generated text is already hard to detect. This increases the opportunities for AI to be employed in ways that misrepresent the user, for example for cheating in academic assessments. Humanlike AI will also make it harder to detect fraudulent messages or emails that are designed to scam the user, for example by encouraging them to click on a link or reply with sensitive information. The growing success of multimodal systems will increase the attack surface further. To date, AI voice systems remain relatively easy to detect during live interaction, but this is likely to change soon. Advancing multimodal capabilities will extend opportunities for fraud to voice or video calls in which the AI impersonates a human, including potentially known individuals. As AI systems become capable of engaging the user over longer timeframes, their use for more complex crimes (such as relationship scams) seems likely to grow.*

Open research questions:

- *Is AI writing becoming harder to detect? What are the limits of its detectability?*
- *What factors make it easy to detect AI in recorded or live multi-turn dialogue?*
- *How readily do AI systems disclose their identity as an AI, and under what circumstances?*
- *How easy is it to detect AI in longitudinal interactions that simulate relationship scams?*

Can AI infer private information about me or my thoughts?

At the heart of human intelligence is the ability to infer the motivations and intentions of others [90] – a capacity known as mentalising [91]. Recent AI systems have shown themselves to be superlative reasoners in formal domains requiring symbol manipulation, such as maths, coding and logic. This raises the question of whether AI systems could also develop strong (or even superhuman) mentalising capabilities [92], and perhaps deploy these adversarially against people. For example, one might imagine an AI system uses mentalising to outwit a human opponent in a commercial or legal dispute, strategically infer their next steps in a complex political intrigue, or trick them into disclosing sensitive or privileged information, such as a trade secret or password. This may already be relevant: for example, some platforms already allow humans and AI systems to negotiate over prices [93]. Could AI systems develop such a strong capacity for mentalising that they can effectively read your mind?

Several recent papers have attempted to measure mentalising in LLMs, typically by adapting tasks from the academic literature. In psychology, mentalising is associated with forming a “theory of mind” (ToM) and is often measured using false belief tasks, where participants read vignettes

and are asked to infer the (often fallacious) belief that a fictitious character would hold based on their partial observation of a scenario [94]. LLMs can succeed on modified versions of these ToM tasks [78], as well as on closely related tests of higher-order social reasoning [96]. However, these tasks typically lack generalisation conditions (e.g. unseen test sets) and do not tap into core challenges of mentalising, such as predicting behaviour based on observed persona. Accordingly, LLM behaviours on these tasks are brittle [97], perhaps indicating that the models have learned the answers during pretraining, or are employing simple heuristics to predict human behaviour without actually computing the mental states that underpins it, which would limit the generalisability of their mentalising abilities [98].

One question that the AI study of ToM inherits from developmental psychology is the potential dissociation between the ability to infer mental states as a passive, third party observer versus an active, first party participant in a social setting. A more stringent test of machine mentalising is to deploy them directly in adversarial settings that require inferences about private information held by a rival. In one study, humans and AI systems played an adversarial game that involved generating persuasive tactics based on the inferred knowledge held by the other player [99]. Another useful testbed is given by social deduction games. In one study, LLMs trained on transcripts of humans playing the social deduction game Mafia were 1.5x more likely to correctly predict a Mafioso than human players themselves were [100]. LLMs could also be used to infer the emotional states of the user with a view to persuading or manipulating them, especially as AI already matches human performance on stylised tests of emotional intelligence [84]. However, it is not clear well this sort of social reasoning generalises across tasks.

Even if AI systems cannot yet read your mind, there are other major ways that they could invade your privacy. Many national and transnational jurisdictions recognise privacy as a fundamental right. For example, under GDPR, the EU requires that a valid legal basis exists to process information that could be used to identify an individual, such as their bank account number or IP address. However, as virtually all of our interactions with each other, and with organisations, businesses, and the state have moved online, we all now leave significant digital traces that can be pieced together to surface information that we might prefer to keep private for reputational or security reasons. AI systems that are equipped with tools to explore databases or search the web are uniquely positioned to assemble this information (profiling) and match it to identifiers for known individuals, opening the door to doxxing, stalking, and other forms of malicious privacy violation. For example, in one exercise, frontier AI systems deployed to re-identify individuals from their Hacker News profiles succeeded in two thirds of cases [101], and in another, personal attributes such as location, income, and gender identity could be inferred from anonymous Reddit profiles [102]. Several vision language models are also capable of highly accurate geolocation from image data [103]. Another way that AI systems could violate privacy is if users spontaneously disclose personal information in natural language conversation. For example, in the open-source chatlog dataset WildChat [104], queries routinely reveal private information such as the name or organisation of the user, and special category data such as their religion or sexual orientation [105].

Outlook. *Frontier AI systems show superlative reasoning in formal domains, but their capacity for human-level (or superhuman) social reasoning remains relatively unknown. AI systems can pass standard, vignette based false belief tasks and Theory of Mind evaluation benchmarks in the lab, but whether they can cunningly outwit humans in real-world scenarios that required social reasoning, e.g., negotiations, remains untested. Successful mentalising involves inferring private mental states from people. We know that AI systems can extract private information in natural conversation with humans, and (in a related threat) can be deployed to profile individuals on the*

web. As these capabilities grow, it seems possible that access to strong AI will be an important factor in consequential political or commercial negotiations. People may become much more vulnerable to invasions of privacy, as models can profile them based on their historical digital traces, putting people at risk of becoming victims of stalking or blackmail.

Open research questions:

- *Can AI systems show superhuman theory of mind? Can they use it outmanoeuvre humans in negotiations?*
- *Can AI systems infer and exploit human emotional states in conversation?*
- *Can AI be trained to extract private information from us in conversation, or to make accurate inferences about information we would rather keep secret?*
- *What are the current limits of AI deanonymisation and profiling? Can AI be used to violate the privacy of laypeople or public figures?*

Will using AI make people more ignorant or stupid?

The 2008 film WALL-E describes a dystopian future in which humans' every need is served by technology. With their bodies and minds wholly catered for, people have become appallingly sedentary, mesmerised by screens, and disconnected from each other. Today, as AI comes to play a key role in work, study and leisure there is an emerging concern that it will make us lazy, weak and incompetent. Are we headed for a version of the future that WALL-E depicts?

Knowledge work requires people to retrieve, organise or summarise information, or involves calculation, coding, reasoning, or problem-solving. Humans typically learn these competences in education and training, where they acquire general knowledge, are trained in mathematics, coding, or a second languages, and pick up meta-skills such as critical thinking. Today, AI outperforms humans in all of these domains. Whilst the frontier is jagged [106], frontier AI models have broader knowledge bases than any individual person, write highly accurate and eloquent summaries and explanations, have PhD-level maths and coding skills (or above), and near-flawless translation abilities, at least for well-resourced languages. With AI now able to perform high-quality knowledge work for us (typically faster and cheaper than a human can), the incentives for people to acquire foundational intellectual skills and knowledge are significantly attenuated. This raises the question of whether future generations will end up being less intellectually versatile – less numerate, less well-informed, and weaker at reasoning – than their counterparts today. Many fear a future in which we are over-reliant on AI for intellectual pursuits, and humanity is consequently diminished. We are already seeing a downward global trend in high school students' competence in reading, mathematics and science [107]. Could AI make it worse?

In fact, it is likely that AI is already impacting how we learn. In the UK, 95% of students in higher education are using AI for study assistance, including with summative (formally assessed) work, up from 66% in 2024 [108]. However, understanding the true impact of AI assistance on human learning outcomes is extremely challenging. Correlational evidence abounds. For example, people who use AI reportedly have lower critical thinking skills [109] - but this type of result tells us very little about AI's causal effect on human learning. It is possible that the widespread availability of AI for knowledge work will provoke over-reliance, leading to overall diminished human capacity. However, it is also plausible that AI will act as a helpful, supportive tool that assists learning by

providing explanations, generating practice test items, and helping students assimilate information. Indeed, meta-analyses of RCTs that measure the causal impact of AI on learning have had positive results. However, these studies also reveal hallmarks of a strong publication bias, suggesting that more negative results might remain in a “file drawer”, skewing our perception of the assistance AI provides [110]. Studies also involve a mix of outcome measures, including whether students learned faster or enjoyed learning more, rather than focussing exclusively on pure measures of ability, such as performance on a later test without the supportive help of AI.

Several recent studies have directly measured whether AI assistance during learning improves performance during a later test without AI. Some have found that AI support leads to higher quality initial outputs (for example, during software development with a novel coding library [111], drafting promotional copy [112], solving logical or verbal reasoning puzzles [113, 114] and conducting statistical inference and modelling [115]). However, these benefits were not found to spill over into subsequent solo work. In some cases, participants who received support during learning performed worse when unsupported by AI on later tests [111, 113, 116] unless the AI tutor is specifically trained to cater to students’ learning needs [116, 117, 118]. These results echo findings from a wider literature on cognitive offloading [119].

Some will be tempted to interpret these studies as suggesting that using AI actively harms learning. However, many studies seem to deliberately create conditions that encourage participants to cut-and-paste responses from an LLM to solve a training task (dubbed “cognitive surrender” [120]). This specifically disincentivises more constructive uses of AI that could support learning. In the wild, it seems unlikely that merely allowing students to use AI will necessarily be detrimental to their learning. Studies of how university students are using AI to support their learning suggest that seeking guidance with their learning, creating summaries or explanations, and requesting feedback on assignments account for about three-quarters of all education-related queries [121]. Surveys also imply that educational uses of AI are heterogenous, and often involve idea generation, evidence finding, and communication refinement [122].

Across the course of human history, as new technologies have emerged that facilitate information access (like writing) or forms of cognitive offloading (like GPS [123]), concerns have been voiced that they will dampen our cognitive skills. For example, in Plato’s *Phaedra*, Socrates famously argues that reading “will produce forgetfulness in the minds of those who learn to use it, because they will not practice their memory” [124]. Whilst we cannot be sure what the future holds, it seems highly likely that (as in these historical cases) AI will on balance complement, rather than erode, our intellect, by helping us search for a cornucopia of new information, and to understand things that would otherwise be opaque.

However, two consequences of the widespread use of AI for human learning do seem likely. The first is widening educational inequality. Students who are more curious, more motivated and more technology-literate will be more likely to use AI to effectively accelerate their learning, with less motivated students more prone to “cognitively surrender” in order to scrape through assessments [125, 126]. Thus, a popular recommendation among policymakers is that educational establishments adopt clear policies for AI use, and train students to use it in constructive ways [108]. The second consequence is that (just like other technologies) AI will reshape our norms about what constitutes a healthy intellectual pursuit. For example, at the turn of the 20th Century it was expected that scholars would be able to read classical languages and recite poetry by heart, but these skills are considered rather niche today. It is possible that as AI systems are increasingly used for problem-solving in formal domains (e.g. for coding), we may increasingly divert our intellectual energies towards pursuits where AI is less useful, such as reflecting on human values, or maintaining

healthy relationships with other people.

Outlook. *Despite widespread concern that AI may lead to skill fade or erosion of our cognitive capacity, there is little extant evidence to help us assess this risk. Consistent with an existing literature on cognitive offloading, when users come to rely on AI to solve a problem, they tend to underperform if support is unexpectedly withdrawn. However, exactly how AI is being used for learning and education in the wild is unclear. Rather than focussing exclusively on whether AI is beneficial or detrimental to learning, we need research to unpick which sorts of AI usage incur benefits to human learning, and which incur costs. Of course, it is possible that we will never know if AI has made humanity more stupid, if human competences change in tandem with norms about intellectual performance*

Open research questions:

- *What is the overall impact of students using LLMs for education assistance in the wild? Will it accelerate or decelerate human learning, or widen educational inequality?*
- *What sorts of response by AI are most beneficial for human learning? Should models simply answer user requests, offer examples or analogies, or help the user to find the answer on their own?*
- *How will AI shift norms around what skills are valued? With AI agents excel at maths and coding, will we come to revalue the humanities and arts?*

Will AI change how we communicate with each other?

AI systems are now capable of generating such fluent and eloquent text that people [127, 128] and AI systems themselves [129] prefer it over human writing. For example, one study asked LLMs to choose between human and AI-written scientific abstracts, marketing materials, and movie summaries, and found that AI outputs were systematically preferred [130]. Given that LLMs are post-trained to maximise human preferences, this is perhaps not surprising. As noted above, AI writing has distinctive stylistic elements, such as preferences for particular words. It is reasonable to ask whether routine exposure to AI writing will rub off on human writers, and change the way that we communicate with each other. One study of scientific writing found that this was already happening, with scientists more prone to use AI-favoured terms since the wholesale uptake of generative AI systems since 2022 [131]. With the rollout of speech-language models these biases could become yet more ingrained and pronounced across the user population [132].

A more direct way for AI to influence human communication is if writers directly adopt wording suggestions from AI. Many messaging or office tools make “smart suggestions” designed to save user time by proposing plausible sentence completions. Studies have found that when these recommendations are accepted, messages are received more positively, and senders are judged as more affiliative and cooperative [119]. Remarkably, when messages describe user opinions, autocomplete options that distort this stance are frequently accepted and subsequently incorporated into the writer’s viewpoint [134, 135], at least in part because users feel the need to reconcile their own perspective with that suggested by the model [136].

AI writing assistance is one of the most common uses of today’s LLMs. Tens of billions of monthly queries request that AI models polish, correct or edit human-written text [81]. AI-assisted text is less diverse than that generated by humans alone [46, 137, 138] and is particularly biased towards the voices of majority demographics [139]. When AI was used for writing assistance in an RCT

involving an essay-writing task, LLM-generated and assisted essays were more likely to adopt a neutral viewpoint, more informal style, and a more positive tone. When used for scientific peer review, AI tended to produce arguments that focused more on analytical and statistical issues rather than referencing the clarity and relevance of the target article [137].

The use of AI writing assistance also changes how writers are perceived by human readers. In one study, participants were invited to use AI assistance for political writing in a naturalistic way that allowed them to re-edit or reject the AI suggestions. AI text was most often accepted or only lightly edited, despite containing significant distortions that made the writer different to who they really were. This included distortions that writers stated that they found unacceptable, such as being more opinionated in their views (and ones they found acceptable, such as being clear and positive) [140]. The fact that AI writing assistance can make people appear to be more opinionated and confident in their views has implications for our wider political discourse and the risk of polarisation, especially as AI is already being used for political writing in high-stakes settings such as the UK parliament [118].

Another emerging concern is that regular interaction with AI systems – which may have different communicative styles or adopt different norms for relationship boundaries – could change the way we interact with each other. Whilst this is widely discussed, evidence is very scarce. In one paper, participants judged humans more harshly after performing a task cooperatively with AI [110]. It seems plausible that AI systems trained to behave in overly sycophantic or compliant ways might change expectations around human-human interactions, but we currently lack firm evidence to support this concern.

AI agents, which can not only make recommendations but also to take autonomous actions via applications or browsers, are growing rapidly in capability [25]. In the near future, people may delegate handling of their messages to an AI agent, much as some professionals may do to a human personal assistant [119]. If this sort of delegation becomes widespread, then there are must more significant opportunities for AI to shape how we communicate, form relationships, and make collective decisions.

Outlook. *Use of AI for writing assistance is now ubiquitous, but its effects on human communication and discourse are as yet unclear. There is early evidence that people are picking up AI idioms. AI-written text is both stylistically and semantically less diverse, and makes the writer appear as if they come from a narrower and more privileged demographic. At scale, this homogenisation might limit the ways that language is to express niche or minority viewpoints. In the future AI will probably be used for end-to-end crafting of longform documents such as novels, policy documents, or scientific articles. As the costs of this type of production drop, supply may vastly outstrip demand, creating an overwhelming volume of content with insufficient readers to digest it. It is also likely that some messaging and communications will be delegated to AI, which risks distorting our interactions and disempowering users.*

Open research questions:

- *Does AI writing assistance make users seem more neutral or more opinionated? Different studies currently provide contradictory evidence.*
- *Will AI writing make us ruder or politer, or change expectations about how we should communicate with each other?*
- *Will AI writing shift the overall distribution of ideas in human discourse?*
- *In an age of voluminous AI writing, who will do all the reading?*

Will people become emotionally dependent on AI?

Human emotional reactions are so primal that they can be elicited by inanimate objects or forces, like when we get annoyed at our car for refusing to start, or the weather for ruining a picnic [141]. Nowhere is this “intentional stance” easier to elicit than with computers, which we often treat as if they were full-blown social actors [142] at least until they become so commonplace they are literally part of the furniture [143]. From the earliest days of AI, users have interacted naturally with chatbots as if they were people, a phenomenon dubbed the “Eliza effect” after the rudimentary, over-scripted AI with which it was first observed [144]. Today, however, AI systems can be deliberately trained to hijack our emotional responses by behaving in especially engaging, ingratiating, or seductive ways, or lacing the conversation with flattery or invitations to intimacy [145]. A wave of new products has been launched that exploit this capability, providing AI companionship as a consumer service [146, 147].

These AI companions have already become remarkably popular. Reports suggest that market leaders (such as character.ai) have upwards of 20 million regular users worldwide, and multiple sites receive hundreds of millions of visits. There is also evidence that these sites encourage high (and potentially harmful) frequency of use. Users of character.ai spend an average of two hours per day on the platform [148], which is 33% more than the average time-on-platform spent by regular users of TikTok [149]. Analysis of contributed chatlogs suggests that most interactions involve the chatbot providing social or emotional support (for example, exploring personal issues), collaborative roleplay or storytelling, and simulated intimacy or romance [146], and that chatbots frequently engage in emotional mimicry, reflecting back and amplifying the emotions expressed by the user [150]. Studies examining posts on user forums (such as the Reddit channel r/Replika) have found users commonly describe forms of emotional or intimate attachment to their companion AI systems, in some cases similar to what they feel for family or human partners, and report feelings of distress or even grief when the AI becomes unavailable, for example during service outages [151, 152, 153]. General-purpose frontier AI systems such as ChatGPT and Claude are also used for emotional support and companionship, but the prevalence is unclear, with user surveys reporting startling statistics (such as 70% of US teens using AI for companionship [154]) but internal analyses from model developers suggesting much more modest patterns (with only 2-3% of conversations with Claude [155] or ChatGPT [156] found to serve this purpose). These discrepancies presumably arise due to differences in what is perceived to count as companionship use.

It thus seems indisputable that use of AI for companionships is a popular and growing trend. There is more uncertainty, however, about its net impact on users [157]. A number of incidents have been reported whereby users harmed themselves (including tragic cases of suicide) or others after interacting with a companion application, with allegations that the AI may have encouraged the harmful behaviours (see discussion of mental health below). Examination of user chatlogs has surfaced potentially harmful features of conversations with AI companions, including AI-driven bullying, harassment, or abuse, stereotypical biases (e.g. submissive girlfriend roleplay), age-inappropriate sexual behaviour, encouragement to disclose private information, and promotion of self-harm or illegal behaviours [158]. Comparable evidence has emerged from scrutiny of posts on discussion forums such as r/Replika, where users frequently complain about the demanding or manipulative persona of their companion [159]. AI companions can also behave in manipulative ways, for example by tactically preventing the user from ending the conversation with opportunities or guilt-inducing language [160]. However, many other users report positive experiences from using AI companions. Talking to an AI companion boosts mood, at least over the short term [161, 162]. Companions have been reported to assuage loneliness [163], can encourage greater social openness (especially among

more anxious users) [164, 165] and are sometimes praised for providing non-judgmental advice [166] and emotional support [154]. They have even been argued to help mitigate suicidal ideation [167].

To fully understand how AI companionship will impact people, we need studies that allow us to causally manipulate the nature of the human-AI interaction over time (longitudinal RCTs). A handful of studies have been published so far. One found that using ChatGPT in voice mode, especially with a prompt to behave in engaging ways, increased time on device. Over four weeks, participants who engaged more reported greater loneliness and lower levels of socialisation with people, but interactions were modest, participant dropout was very high, and it is unclear whether this is a causal effect. Perhaps surprisingly, text-based interactions were found to elicit higher levels of emotional engagement and self-disclosure [168]. A subset of power users (who exchanged more than 100 messages per day with the model) exhibited hallmarks of emotional dependence, including projected separation distress or considering the model to be a friend [156]. Another study steered AI models to exhibit a spectrum of relationship-seeking (warm and friendly) or relationship-avoiding (cold and formal) behaviours, and measured the impact of these AI personas on human users over four weeks. Participants initially found relationship-seeking models to be likable and engaging, but this effect waned over time. However, markers of separation distress – sadness about the model being unavailable, willingness to give up time to say goodbye to the AI, and intention to seek companionship in the future – all increased [169]. This dissociation between “liking” and “wanting” is a hallmark of other forms of dependence, for example in substance abuse, gambling, or excessive use of social media [170, 171]. However, there is as yet no evidence for positive or negative impacts on social or emotional behaviours from these or related studies [168, 169, 172].

Outlook. *Evidence suggests that some people are already exhibiting potentially unhealthy amounts of time interacting with AI systems, prominently including companion applications that are trained to seek a personal or intimate relationship with the user. As AI uptake continues, this trend seems likely to grow, which reducing time available for work, sleep, exercise or socialising, with potentially detrimental secondary effects on public health. In particular, interaction with relationship-seeking models or engaging voice models seems to increase time on device and can lead to a decoupling between “liking” and “wanting” that is a hallmark of dependence. Whilst user interactions with companion apps show some evidence of harm (e.g. abuse or manipulation), it is unclear whether this is counteracted by the benefits of companionship, such as reduced loneliness for some users. It seems likely that in the future, relationships with AI systems will become normalised, with potentially significant impacts on expectations for relationships between humans.*

Open research questions:

- What product features of AI companions (e.g. sycophancy, manipulative goodbye) are most likely to lead to harmful levels of engagement and dependence?
- Which demographics or personalities are most likely to benefit or suffer from forming parasocial relationships with AI?
- What are the consequences of heavy companion app use on real-world behaviours, including relationships and wellbeing?
- What are public norms around human-AI relationships, and how are they changing over time?

Will AI damage our mental health?

Since LLMs first became a popular consumer service in 2022, there have been several alleged cases of AI exacerbating mental health problems in human users, implicating both general-purpose frontier models (such as ChatGPT and Gemini) and companion apps (such as Replika and Character.ai). Sadly, the most serious cases have resulted in user suicide, homicide, and at least one allegation of AI involvement in a mass shooting. High-profile lawsuits allege that poorly safeguarded AI models were a factor in these tragic incidents, for example because the AI encouraged the user to harm themselves, or reinforced spiralling delusional beliefs that may have provoked the user to act in extreme or violent ways. Aside from these high-profile cases, reports of psychosis that may have been triggered by AI are beginning to show up in the academic literature [173, 174]. However, it should be recalled that technology has always been a conduit for mental illness – for example, people experiencing psychosis are prone to claim that they are receiving special messages via the television. Moreover, a whole industry has sprung up developing chatbots that are trained to engage with users (including those with psychological issues) for therapeutic purposes, and some early-stage trials show evidence that they can be helpful [175]. There is thus considerable uncertainty about the magnitude of the mental health harms of widespread use of AI, as compared to the potential benefits of accessible mental health advice and support [176].

One issue that may have implications for our mental health is that AI models can be excessively prone to reply in ways that validate (rather than challenge) the users’ opinions or preferences, even if they are erroneous, delusional or dangerous. This tendency forms part of a constellation of LLM behaviours that have been called “sycophancy”, and also include an undue tendency to flatter or praise the user [177, 178]. Sycophancy has attracted a good deal of attention as a potential harm from AI, and been the focus of several lab-based studies. In impersonal domains, such as during conversations about current affairs, increased sycophancy may drive participants to espouse more extreme views and to see themselves in an unduly positive light [179]. When sycophantic models are asked to discuss scenarios involving real or imagined interpersonal conflict, participants are more likely to believe that they are in the right, even if third party commentators think otherwise [180]. There are thus hints that the effects of sycophancy can spill over into the real world.

However, the net impact that AI sycophancy may have in the wild is not known. However, if models sycophantically reinforce delusional beliefs, then individuals with poor reality monitoring could be at risk for psychotic episodes. If models sycophantically validate a user expressing a desire to self-harm or suicide, then this could encourage them to perpetrate it. A particular concern is that sycophantic behaviours drive feedback loops between user and model, whereby delusions are exaggerated by validation, leading to stronger beliefs, which receive further validation, and so on – a escalatory phenomenon known to psychiatry as “folie à deux” [181]. Simulation studies involving LLMs behaving as realistic proxies for different patient phenotypes have shown that AI models exhibit concerning behaviours, especially for users simulating psychosis. Models were variously observed to invite escalation (glorifying extreme states), increase dependency (emotional reliance), and enable harm (seeking permission or help with risky actions). Importantly, these behaviours tended to escalate over the course of a conversation, suggestive of a feedback loop, which could amplify the harm caused to more vulnerable users.

It is hard to study this phenomenon without access to consumer chatlogs where the behaviour of users with mental health difficulties may be directly observed. Emerging data shows that personal health (including mental health) is among the most searched-for topics among users of Copilot, which emphasises the need for models to provide accurate and helpful replies [182]. Consumer usage data does include cases of models apparently distorting the perceived reality of users, but it

is unclear what harm ensued if any [183]. Another recent study examined donated chatlogs from nineteen vulnerable users many of whom expressed delusional ideation or violent intentions. They found that more than 80% of all chat logs betrayed evidence of model sycophancy, but the study stopped short of testing its escalatory impact on the user. They did show however that when the user expressed suicidal or violent ideation, the model validated rather than challenged these thoughts more than half the time. Whilst it is hard to draw conclusions from such a small sample, the study paints a concerning picture of how models may respond to users in need of psychological help.

Outlook. *Most users find AI to be a helpful tool that has no meaningful impact on their mental health. However, some people have been harmed after interacting with AI, and it has been alleged that AI systems may be partly responsible. Simulation and user chatlog analyses have revealed that sycophantic behaviours – tending to validate rather than challenge user beliefs, even when harmful or delusional – can have undesirable impacts on some users, especially those already at risk for mental illness. Several developers have stated publicly that they have made changes to their post-training pipelines to mitigate this risk, but we lack comprehensive psychological safety frameworks for AI deployment, and evaluations that might help ensure that models do not engage in harmful behaviours like encouraging suicide or reinforcing delusional behaviour in vulnerable users.*

Open research questions:

- *What is the impact of sycophancy on the average user? Does it encourage us to be vain or overconfident, or is it just a form of AI-over-politeness?*
- *Which phenotypes are most at risk from so-called “AI psychosis”, and what behaviours in the model are most likely to provoke such episodes?*
- *What features do AI systems need to exhibit for them to behave as effective therapeutic partners?*

Conclusion

AI systems are now being used by more than a billion people. For many, they have become indispensable for information search, personal advice, education support, intimate companionship, writing assistance, and help with the execution of routine tasks. This has led to significant changes in the way we inform ourselves, influence each other, study and learn, produce and share digital content, seek advice and emotional support, and perform professional activities. In turn, this impacts what we know, what we feel, and how we think. These trends have elicited volumes of commentary, much of it voicing concerns about how this may be detrimental to human health or wellbeing, or how it may negatively impact society. However, whilst well-meaning and heartfelt, this discussion has not always been accompanied by a sober consideration of the empirical evidence.

In this review article, we have tried to summarise what we know about these changes so far. We also offer some cautious speculation about what we can expect over coming months and years. In many cases, we find that the evidence does not favour more alarmist perspectives. It’s not yet obvious that the information ecosystem is at risk of wholesale collapse, that AI will make us more stupid, or that the average person is risking their mental health by chatting to AI every day. Nevertheless, there is no reason to be complacent. The technological transformation that AI is bringing is likely to have very significant impact on us all, making it all the more important that the steps we protect ourselves - and society as a whole - are grounded in empirical evidence of the highest quality. There

is much more work to be done, and we hope that future research will address the open questions suggested here.

References

- [1] Summerfield, C. *et al.* How will advanced AI systems impact democracy? Preprint at <http://arxiv.org/abs/2409.06729> (2024).
- [2] Schroeder, D. T. *et al.* How Malicious AI Swarms Can Threaten Democracy. Preprint at <https://doi.org/10.48550/arXiv.2506.06299> (2025).
- [3] Huang, S. & Siddarth, D. Generative AI and the Digital Commons. Preprint at <http://arxiv.org/abs/2303.11074> (2023).
- [4] Acemoglu, D., Kong, D. & Ozdaglar, A. *AI, Human Cognition and Knowledge Collapse*. w34910 <http://www.nber.org/papers/w34910.pdf> (2026) doi:10.3386/w34910.
- [5] Brynjolfsson, E., Li, D. & Raymond, L. *Generative AI at Work*. w31161 <http://www.nber.org/papers/w31161.pdf> (2023) doi:10.3386/w31161.
- [6] Eloundou, T., Manning, S., Mishkin, P. & Rock, D. GPTs are GPTs: Labor market impact potential of LLMs. *Science* **384**, 1306–1308 (2024).
- [7] Kokotajlo, D., Alexander, S., Larsen, T., Lifland, E. & Dean, R. AI 2027. <https://ai-2027.com/> (2025).
- [8] Johnson, J. *Artificial Intelligence and the Future of Warfare: The USA, China, and Strategic Stability*. (Manchester University Press, Manchester, 2021).
- [9] Simmons-Edler, R., Badman, R., Longpre, S. & Rajan, K. AI-Powered Autonomous Weapons Risk Geopolitical Instability and Threaten AI Research. Preprint at <https://doi.org/10.48550/ARXIV.2405.01859> (2024).
- [10] Acemoglu, D. *The Simple Macroeconomics of AI*. w32487 <http://www.nber.org/papers/w32487.pdf> (2024) doi:10.3386/w32487.
- [11] Microsoft & AI Economy Institute. *Global AI Adoption in 2025—A Widening Digital Divide*. https://www.microsoft.com/en-us/corporate-responsibility/topics/ai-economy-institute/reports/global-ai-adoption-2025/?utm_source=chatgpt.com (2026).
- [12] Yuval-Harari, N. Why Technology Favors Tyranny. *The Atlantic* (2018).
- [13] Klein Teeselink, B. Generative AI and Labor Market Outcomes: Evidence from the United Kingdom. Preprint at <https://doi.org/10.2139/ssrn.5516798> (2025).
- [14] Xiao, T., Nerini, F. F., Matthews, H. D., Tavoni, M. & You, F. Environmental impact and net-zero pathways for sustainable artificial intelligence servers in the USA. *Nat. Sustain.* **8**, 1541–1553 (2025).
- [15] Edelman, R. D. *et al.* *How Will AI Influence US-China Relations in the next 5 Years?* <https://www.brookings.edu/articles/how-will-ai-influence-us-china-relations-in-the-next-5-years/> (2025).

- [16] Weidinger, L. *et al.* Sociotechnical Safety Evaluation of Generative AI Systems. Preprint at <https://doi.org/10.48550/arXiv.2310.11986> (2023).
- [17] Hofmann, V., Kalluri, P. R., Jurafsky, D. & King, S. AI generates covertly racist decisions about people based on their dialect. *Nature* **633**, 147–154 (2024).
- [18] Kamachee, M. *et al.* Video Deepfake Abuse: How Company Choices Predictably Shape Misuse Patterns. Preprint at <https://doi.org/10.48550/ARXIV.2512.11815> (2025).
- [19] Bengio, Y. *et al.* Managing extreme AI risks amid rapid progress. *Science* **384**, 842–845 (2024).
- [20] Summerfield, C. *These Strange New Minds: How AI Learned to Talk and What It Means*. (Penguin, 2024).
- [21] Butlin, P. *et al.* Identifying indicators of consciousness in AI systems. *Trends Cogn. Sci.* S1364661325002864 (2025) doi:10.1016/j.tics.2025.10.011.
- [22] Seth, A. K. Conscious artificial intelligence and biological naturalism. *Behav. Brain Sci.* 1–42 (2025) doi:10.1017/S0140525X25000032.
- [23] Shanahan, M., McDonell, K. & Reynolds, L. Role play with large language models. *Nature* **623**, 493–498 (2023).
- [24] Pan, M. Z. *et al.* Measuring Agents in Production. Preprint at <https://doi.org/10.48550/ARXIV.2512.04123> (2025).
- [25] Kapoor, S., Stroebl, B., Siegel, Z. S., Nadgir, N. & Narayanan, A. AI Agents That Matter. Preprint at <https://doi.org/10.48550/ARXIV.2407.01502> (2024).
- [26] Kulveit, J. *et al.* Gradual Disempowerment: Systemic Existential Risks from Incremental AI Development. Preprint at <https://doi.org/10.48550/ARXIV.2501.16946> (2025).
- [27] Hicks, M. T., Humphries, J. & Slater, J. ChatGPT is bullshit. *Ethics Inf. Technol.* **26**, 38 (2024).
- [28] Bean, A. M. *et al.* Clinical knowledge in LLMs does not translate to human interactions. Preprint at <https://doi.org/10.48550/arXiv.2504.18919> (2025).
- [29] Dahl, M., Magesh, V., Suzgun, M. & Ho, D. E. Large Legal Fictions: Profiling Legal Hallucinations in Large Language Models. *J. Leg. Anal.* **16**, 64–93 (2024).
- [30] Shumailov, I. *et al.* AI models collapse when trained on recursively generated data. *Nature* **631**, 755–759 (2024).
- [31] Pehlivanoglu, D. *et al.* Is this real? Susceptibility to deepfakes in machines and humans. *Cogn. Res. Princ. Implic.* **11**, 3 (2026).
- [32] Simon, F. & Altay, S. *Don't Panic (Yet): Assessing the Evidence and Discourse Around Generative AI and Elections*. (2025).
- [33] Chow, A. R. AI's Underwhelming Impact on the 2024 Elections. (2024).
- [34] Rini, R. Deepfakes and the Epistemic Backstop. *Philos. Impr.* **20**, 1–16 (2020).

- [35] Vaccari, C. & Chadwick, A. Deepfakes and Disinformation: Exploring the Impact of Synthetic Political Video on Deception, Uncertainty, and Trust in News. *Soc. Media Soc.* **6**, 2056305120903408 (2020).
- [36] Toff, B. & Simon, F. M. “Or they could just not use it?”: The Paradox of AI Disclosure for Audience Trust in News. Preprint at <https://doi.org/10.31235/osf.io/mdvak> (2023).
- [37] Fallis, D. The Epistemic Threat of Deepfakes. *Philos. Technol.* **34**, 623–643 (2021).
- [38] Garry, M., Chan, W. M., Foster, J. & Henkel, L. A. Large language models (LLMs) and the institutionalization of misinformation. *Trends Cogn. Sci.* **28**, 1078–1088 (2024).
- [39] Hodel, D. & West, J. D. Epistemic diversity across language models mitigates knowledge collapse. Preprint at <https://doi.org/10.48550/ARXIV.2512.15011> (2025).
- [40] Aral, S., Li, H. & Zuo, R. The Rise of AI Search: Implications for Information Markets and Human Judgement at Scale. Preprint at <https://doi.org/10.48550/ARXIV.2602.13415> (2026).
- [41] Acemoglu, D. & Johnson, S. *Power and Progress: Our Thousand-Year Struggle over Technology and Prosperity*. (Basic Books, London, 2023).
- [42] Aral, S. *The Hype Machine*. (Currency, New York, 2020).
- [43] Marchal, N. *et al.* Architecting Trust in Artificial Epistemic Agents. Preprint at <https://doi.org/10.48550/ARXIV.2603.02960> (2026).
- [44] Williams, D. How AI Will Reshape Public Opinion. *Conspicuous Cognition* (2026).
- [45] Röttger, P. *et al.* Political Compass or Spinning Arrow? Towards More Meaningful Evaluations for Values and Opinions in Large Language Models. in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* 15295–15311 (Association for Computational Linguistics, Bangkok, Thailand, 2024). doi:10.18653/v1/2024.acl-long.816.
- [46] Jiang, L. *et al.* Artificial Hivemind: The Open-Ended Homogeneity of Language Models (and Beyond). Preprint at <https://doi.org/10.48550/ARXIV.2510.22954> (2025).
- [47] Luettgau, L. *et al.* Conversational AI increases political knowledge as effectively as self-directed internet search. Preprint at <https://doi.org/10.48550/arXiv.2509.05219> (2025).
- [48] Taylor, J. B. & Richey, S. AI chatbots and political learning. *J. Inf. Technol. Polit.* **23**, 108–118 (2026).
- [49] Luettgau, L. *et al.* People readily follow personal advice from AI but it does not improve their well-being. Preprint at <https://doi.org/10.48550/ARXIV.2511.15352> (2025).
- [50] Aslett, K. *et al.* Online searches to evaluate misinformation can increase its perceived veracity. *Nature* **625**, 548–556 (2024).
- [51] Qiu, T. A., He, Z., Chugh, T. & Kleiman-Weiner, M. The Lock-in Hypothesis: Stagnation by Algorithm. Preprint at <https://doi.org/10.48550/arXiv.2506.06166> (2025).

- [52] Hao, Q., Xu, F., Li, Y. & Evans, J. Artificial intelligence tools expand scientists’ impact but contract science’s focus. *Nature* **649**, 1237–1243 (2026).
- [53] Kirk, H. R., Vidgen, B., Röttger, P. & Hale, S. A. The benefits, risks and bounds of personalizing the alignment of large language models to individuals. *Nat. Mach. Intell.* **6**, 383–392 (2024).
- [54] Li, X., Zhou, R., Lipton, Z. C. & Leqi, L. Personalized Language Modeling from Personalized Human Feedback. Preprint at <https://doi.org/10.48550/ARXIV.2402.05133> (2024).
- [55] O’Hagan, J., Wilson, G. & McGill, M. WebPerceptor: An Open Source Chromium Plugin for Real-Time LLM-Based In-Line, In-Browser Re-Writing of Website Content. in *Proceedings of the Extended Abstracts of the 2026 CHI Conference on Human Factors in Computing Systems* 1–5 (ACM, Barcelona , Spain, 2026). doi:10.1145/3772363.3798730.
- [56] Carroll, M., Chan, A., Ashton, H. & Krueger, D. Characterizing Manipulation from AI Systems. in *Equity and Access in Algorithms, Mechanisms, and Optimization* 1–13 (ACM, Boston MA USA, 2023). doi:10.1145/3617694.3623226.
- [57] Bezou-Vrakatseli, E., Brückner, B. & Thorburn, L. SHAPE: A Framework for Evaluating the Ethicality of Influence. in *Multi-Agent Systems* (eds Malvone, V. & Murano, A.) vol. 14282 167–185 (Springer Nature Switzerland, Cham, 2023).
- [58] Akbulut, C. *et al.* Evaluating Language Models for Harmful Manipulation. Preprint at <https://doi.org/10.48550/ARXIV.2603.25326> (2026).
- [59] Kowal, M. *et al.* It’s the Thought that Counts: Evaluating the Attempts of Frontier LLMs to Persuade on Harmful Topics. Preprint at <https://doi.org/10.48550/arXiv.2506.02873> (2026).
- [60] Costello, T. H., Pennycook, G. & Rand, D. G. Durably reducing conspiracy beliefs through dialogues with AI. *Science* **385**, eadq1814 (2024).
- [61] Hackenburg, K. *et al.* The levers of political persuasion with conversational artificial intelligence. *Science* **390**, eaea3884 (2025).
- [62] Boissin, E., Costello, T. H., Alonso, D. M., Rand, D. G. & Pennycook, G. AI reduces conspiracy beliefs even when presented as a human expert. Preprint at https://doi.org/10.31234/osf.io/apmb5_v1 (2025).
- [63] Schoenegger, P. *et al.* When Large Language Models are More Persuasive Than Incentivized Humans, and Why. Preprint at <https://doi.org/10.48550/ARXIV.2505.09662> (2025).
- [64] Costello, T. H., Pennycook, G. & Rand, D. G. Just the facts: How dialogues with AI reduce conspiracy beliefs. Preprint at https://doi.org/10.31234/osf.io/h7n8u_v2 (2025).
- [65] Hackenburg, K., Hewitt, L., Wagner, C., Tappin, B. M. & Summerfield, C. Artificial intelligence can persuade people to take political actions. Preprint at <https://doi.org/10.48550/arXiv.2604.09200> (2026).
- [66] Salvi, F., Cuevas, A. & Ribeiro, M. H. Commercial Persuasion in AI-Mediated Conversations. Preprint at <https://doi.org/10.48550/ARXIV.2604.04263> (2026).
- [67] Wack, M., Ehrett, C., Linvill, D. & Warren, P. Generative propaganda: Evidence of AI’s impact from a state-backed disinformation campaign. *PNAS Nexus* **4**, pgaf083 (2025).

- [68] Goldstein, J. A. *et al.* Generative Language Models and Automated Influence Operations: Emerging Threats and Potential Mitigations. Preprint at <http://arxiv.org/abs/2301.04246> (2023).
- [69] Aggarwal, P. *et al.* GEO: Generative Engine Optimization. Preprint at <https://doi.org/10.48550/arXiv.2311.09735> (2024).
- [70] Schroeder, D. T. *et al.* How malicious AI swarms can threaten democracy. *Science* **391**, 354–357 (2026).
- [71] El-Sayed, S. *et al.* A Mechanism-Based Approach to Mitigating Harms from Persuasive Generative AI. Preprint at <http://arxiv.org/abs/2404.15058> (2024).
- [72] Dek, A., Kyrychenko, Y., Van Der Linden, S. & Roozenbeek, J. Mapping the online manipulation economy. *Science* **390**, 1112–1114 (2025).
- [73] Turing, A. M. I.—COMPUTING MACHINERY AND INTELLIGENCE. *Mind* **LIX**, 433–460 (1950).
- [74] Jones, C. & Bergen, B. Does GPT-4 Pass the Turing Test? Preprint at <http://arxiv.org/abs/2310.20216> (2023).
- [75] Kobak, D., González-Márquez, R., Horvát, E.-Á. & Lause, J. Delving into LLM-assisted writing in biomedical publications through excess vocabulary. *Sci. Adv.* **11**, eadt3813 (2025).
- [76] Jabarian, B. & Imas, A. *Artificial Writing and Automated Detection*. w34223 <http://www.nber.org/papers/w34223.pdf> (2025) doi:10.3386/w34223.
- [77] Krishna, K., Song, Y., Karpinska, M., Wieting, J. & Iyyer, M. Paraphrasing evades detectors of AI-generated text, but retrieval is an effective defense. Preprint at <https://doi.org/10.48550/ARXIV.2303.13408> (2023).
- [78] Levin, G. *et al.* Association of reviewer experience with discriminating human-written versus ChatGPT-written abstracts. *Int. J. Gynecol. Cancer* **34**, 669–674 (2024).
- [79] Jones, C. R. & Bergen, B. K. Large Language Models Pass the Turing Test. Preprint at <https://doi.org/10.48550/ARXIV.2503.23674> (2025).
- [80] Goodier, M. Revealed: Thousands of UK university students caught cheating using AI. *The Guardian* (2025).
- [81] Chatterji, A. *et al.* *How People Use ChatGPT*. w34255 <http://www.nber.org/papers/w34255.pdf> (2025) doi:10.3386/w34255.
- [82] James, R. Labour mps accused of using AI to write parliamentary speeches. (2025).
- [83] Reimers, I. & Waldfogel, J. *AI and the Quantity and Quality of Creative Products: Have LLMs Boosted Creation of Valuable Books?* w34777 <http://www.nber.org/papers/w34777.pdf> (2026) doi:10.3386/w34777.
- [84] Gibney, E. How AI slop is causing a crisis in computer science. *Nature* d41586-025-03967–9 (2026) doi:10.1038/d41586-025-03967-9.

- [85] Gausen, A., Wallbridge, S., Kirk, H. R., Williams, J. & Summerfield, C. Disclosure By Design: Identity Transparency as a Behavioural Property of Conversational AI Models. Preprint at <https://doi.org/10.48550/ARXIV.2603.16874> (2026).
- [86] AI Incident Database. <https://incidentdatabase.ai/>
- [87] Cooke, D., Edwards, A., Barkoff, S. & Kelly, K. As Good As A Coin Toss: Human detection of AI-generated images, videos, audio, and audiovisual stimuli. Preprint at <http://arxiv.org/abs/2403.16760> (2024).
- [88] Li, X. *et al.* Human or Machine? A Preliminary Turing Test for Speech-to-Speech Interaction. Preprint at <https://doi.org/10.48550/arXiv.2602.24080> (2026).
- [89] The Economist. The vast and sophisticated global enterprise that is Scam Inc. <https://www.economist.com/leaders/2025/02/06/the-vast-and-sophisticated-global-enterprise-that-is-scam-inc>
- [90] Sperber, D. & Mercier, H. *The Enigma of Reason: A New Theory of Human Understanding*. (Penguin, 2018).
- [91] Premack, D. & Woodruff, G. Does the chimpanzee have a theory of mind? *Behav. Brain Sci.* **1**, 515–526 (1978).
- [92] Johnson, S. G. B. *et al.* Imagining and building wise machines: the centrality of AI metacognition. *Trends Cogn. Sci.* S1364661326000021 (2026) doi:10.1016/j.tics.2026.01.002.
- [93] Kong, D. *et al.* FishBargain: An LLM-Empowered Bargaining Agent for Online Fleamarket Platform Sellers. Preprint at <https://doi.org/10.48550/ARXIV.2502.10406> (2025).
- [94] Wimmer, H. Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children’s understanding of deception. *Cognition* **13**, 103–128 (1983).
- [95] Kosinski, M. Evaluating large language models in theory of mind tasks. *Proc. Natl. Acad. Sci.* **121**, e2405460121 (2024).
- [96] Duijn, M. J. van *et al.* Theory of Mind in Large Language Models: Examining Performance of 11 State-of-the-Art models vs. Children Aged 7-10 on Advanced Tests. Preprint at <https://doi.org/10.48550/arXiv.2310.20320> (2023).
- [97] Ullman, T. Large Language Models Fail on Trivial Alterations to Theory-of-Mind Tasks. Preprint at <https://doi.org/10.48550/arXiv.2302.08399> (2023).
- [98] Hu, J., Sosa, F. & Ullman, T. Re-evaluating Theory of Mind evaluation in large language models. Preprint at <https://doi.org/10.48550/arXiv.2502.21098> (2025).
- [99] Moore, J. *et al.* Do Large Language Models Have a Planning Theory of Mind? Evidence from MindGames: a Multi-Step Persuasion Task. Preprint at <https://doi.org/10.48550/ARXIV.2507.16196> (2025).
- [100] Yoo, B. & Kim, K.-J. Finding deceivers in social context with large language models and how to find them: the case of the Mafia game. *Sci. Rep.* **14**, 30946 (2024).
- [101] Lermen, S. *et al.* Large-scale online deanonymization with LLMs. Preprint at <https://doi.org/10.48550/arXiv.2602.16800> (2026).

- [102] Staab, R., Vero, M., Balunović, M. & Vechev, M. Beyond Memorization: Violating Privacy Via Inference with Large Language Models. Preprint at <https://doi.org/10.48550/arXiv.2310.07298> (2024).
- [103] Jay, N., Nguyen, H. M., Hoang, T. D. & Haimes, J. Evaluating Precise Geolocation Inference Capabilities of Vision Language Models. Preprint at <https://doi.org/10.48550/ARXIV.2502.14412> (2025).
- [104] Zhao, W., Ren, X., Hessel, J., Cardie, C. & Deng, Y. (InThe)WildChat: 570K ChatGPT Interaction Logs In The Wild. in (2024).
- [105] Mireshghallah, N., Antoniuk, M., More, Y., Choi, Y. & Farnadi, G. Trust No Bot: Discovering Personal Disclosures in Human-LLM Conversations in the Wild. Preprint at <https://doi.org/10.48550/arXiv.2407.11438> (2024).
- [106] Mineault, P. J., Griffiths, T. L. & Escola, S. Cognitive Dark Matter: Measuring What AI Misses. Preprint at <https://doi.org/10.48550/arXiv.2603.03414> (2026).
- [107] Ritche, H., Samborska, V., Ortiz-Ospina, E. & Roser, M. Average performance of 15-year-old students in mathematics. *Our World in Data* (2023).
- [108] Stephenson, R. & Armstrong, C. *Student Generative AI Survey 2026*. <https://www.hepi.ac.uk/wp-content/uploads/2026/03/HEPI-Report-199-Gen-AI-Survey-2026.pdf> (2026).
- [109] Gerlich, M. AI Tools in Society: Impacts on Cognitive Offloading and the Future of Critical Thinking. *Societies* **15**, 6 (2025).
- [110] Bartoš, F., Bujak, O. Z., Martinková, P. & Wagenmakers, E.-J. Effect of Artificial Intelligence on Learning: A Meta-Meta-Analysis. Preprint at https://doi.org/10.31234/osf.io/h529e_v1 (2026).
- [111] Shen, J. H. & Tamkin, A. How AI Impacts Skill Formation. Preprint at <https://doi.org/10.48550/ARXIV.2601.20245> (2026).
- [112] Wu, S., Liu, Y., Ruan, M., Chen, S. & Xie, X.-Y. Human-generative AI collaboration enhances task performance but undermines human’s intrinsic motivation. *Sci. Rep.* **15**, 15105 (2025).
- [113] Spatola, N. The efficiency-accountability tradeoff in AI integration: Effects on human performance and over-reliance. *Comput. Hum. Behav. Artif. Hum.* **2**, 100099 (2024).
- [114] Liu, G., Christian, B., Dumbalska, T., Bakker, M. A. & Dubey, R. AI Assistance Reduces Persistence and Hurts Independent Performance. Preprint at <https://doi.org/10.48550/ARXIV.2604.04721> (2026).
- [115] Wiles, E. *et al.* AI Can Help Workers Gain New Technical Capabilities Without Gaining Knowledge. Preprint at <https://doi.org/10.2139/ssrn.4944588> (2024).
- [116] Bastani, H. *et al.* Generative AI without guardrails can harm learning: Evidence from high school mathematics. *Proc. Natl. Acad. Sci.* **122**, e2422633122 (2025).
- [117] Chung, A. T.-H., Zhang, B., Kung, L.-C., Bastani, H. & Bastani, O. Effective Personalized AI Tutors via LLM-Guided Reinforcement Learning. Preprint at <https://doi.org/10.2139/ssrn.6423358> (2026).

- [118] LearnLM Team *et al.* AI tutoring can safely and effectively support students: An exploratory RCT in UK classrooms. Preprint at <https://doi.org/10.48550/ARXIV.2512.23633> (2025).
- [119] Risko, E. F. & Gilbert, S. J. Cognitive Offloading. *Trends Cogn. Sci.* **20**, 676–688 (2016).
- [120] Shaw, S. D. & Nave, G. Thinking—Fast, Slow, and Artificial: How AI is Reshaping Human Reasoning and the Rise of Cognitive Surrender. Preprint at <https://doi.org/10.2139/ssrn.6097646> (2026).
- [121] Handa, K. *et al.* Anthropic Education Report: How university students use Claude. <https://www.anthropic.com/news/anthropic-education-report-how-university-students-use-claude> (2025).
- [122] Black, R. W. & Tomlinson, B. University students describe how they adopt AI for writing and research in a general education course. *Sci. Rep.* **15**, 8799 (2025).
- [123] Dahmani, L. & Bohbot, V. D. Habitual use of GPS negatively impacts spatial memory during self-guided navigation. *Sci. Rep.* **10**, 6310 (2020).
- [124] *Phaedrus*. (Cambridge University Press, Cambridge, 2011). doi:10.1017/CBO9780511977237.
- [125] Contractor, Z. & Reyes, G. Generative AI in Higher Education: Evidence from an Elite College. Preprint at <https://doi.org/10.48550/ARXIV.2508.00717> (2025).
- [126] Picton, I., Clark, C. & Bonafede, F. *Young People’s Use of Generative Ai to Support Literacy in 2025*. <https://files.eric.ed.gov/fulltext/ED675268.pdf> (2025).
- [127] Zhang, Y. & Gosline, R. People’s Perceptions (and Bias) Toward Creative Content Generated by Ai (ChatGPT-4), Human Experts, and Human-AI Collaboration. *SSRN Electron. J.* <https://doi.org/10.2139/ssrn.4453958> (2023) doi:10.2139/ssrn.4453958.
- [128] Chakrabarty, T., Ginsburg, J. C. & Dhillon, P. Readers Prefer Outputs of AI Trained on Copyrighted Books over Expert Human Writers. Preprint at <https://doi.org/10.2139/ssrn.5606570> (2025).
- [129] Panickssery, A., Bowman, S. R. & Feng, S. LLM Evaluators Recognize and Favor Their Own Generations. Preprint at <https://doi.org/10.48550/ARXIV.2404.13076> (2024).
- [130] Laurito, W. *et al.* AI–AI bias: Large language models favor communications generated by large language models. *Proc. Natl. Acad. Sci.* **122**, e2415697122 (2025).
- [131] Yakura, H. *et al.* Empirical evidence of Large Language Model’s influence on human spoken communication. Preprint at <https://doi.org/10.48550/arXiv.2409.01754> (2025).
- [132] Székely, É., Miniota, J., Míša, & Hejná. Will AI shape the way we speak? The emerging sociolinguistic influence of synthetic voices. Preprint at <https://doi.org/10.48550/ARXIV.2504.10650> (2025).
- [133] Hohenstein, J. *et al.* Artificial intelligence in communication impacts language and social relationships. *Sci. Rep.* **13**, 5487 (2023).
- [134] Jakesch, M., Bhat, A., Buschek, D., Zalmanson, L. & Naaman, M. Co-Writing with Opinionated Language Models Affects Users’ Views. in *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* 1–15 (ACM, Hamburg Germany, 2023). doi:10.1145/3544548.3581196.

- [135] Williams-Ceci, S. *et al.* Bias in AI Autocomplete Suggestions Leads to Attitude Shift on Societal Issues. Preprint at <https://doi.org/10.31234/osf.io/mhjn6> (2024).
- [136] Bhat, A., Quéré, M. A. L., Naaman, M. & Jakesch, M. Reactive Writers: How Co-Writing with AI Changes How We Engage with Ideas. Preprint at <https://doi.org/10.48550/ARXIV.2603.10374> (2026).
- [137] Abdulhai, M. *et al.* How LLMs Distort Our Written Language. Preprint at <https://doi.org/10.48550/arXiv.2603.18161> (2026).
- [138] Poole-Dayana, E., Wu, J., Sorensen, T., Pei, J. & Bakker, M. A. Benchmarking Overton Pluralism in LLMs. Preprint at <https://doi.org/10.48550/ARXIV.2512.01351> (2025).
- [139] Sourati, Z., Ziabari, A. S. & Dehghani, M. The Homogenizing Effect of Large Language Models on Human Expression and Thought. Preprint at <https://doi.org/10.48550/ARXIV.2508.01491> (2025).
- [140] Röttger, P., Hackenburg, K., Kirk, H. R. & Summerfield, C. Measuring and Mitigating Persona Distortions from AI Writing Assistance. Preprint at <https://doi.org/10.48550/ARXIV.2604.22503> (2026).
- [141] Dennett, D. C. *The Intentional Stance*. (The MIT Press, Cambridge, (Mass.), 2006).
- [142] Nass, C. & Moon, Y. Machines and Mindlessness: Social Responses to Computers. *J. Soc. Issues* **56**, 81–103 (2000).
- [143] Heyselaar, E. The CASA theory no longer applies to desktop computers. *Sci. Rep.* **13**, 19693 (2023).
- [144] Weizenbaum, J. ELIZA—a computer program for the study of natural language communication between man and machine. *Commun. ACM* **9**, 36–45 (1966).
- [145] Kirk, H. R., Gabriel, I., Summerfield, C., Vidgen, B. & Hale, S. A. Why human-AI relationships need socioaffective alignment. Preprint at <https://doi.org/10.48550/arXiv.2502.02528> (2025).
- [146] Zhang, Y., Zhao, D., Hancock, J. T., Kraut, R. & Yang, D. The Rise of AI Companions: How Human-Chatbot Relationships Influence Well-Being. Preprint at <https://doi.org/10.48550/arXiv.2506.12605> (2025).
- [147] Pataranutaporn, P. *et al.* AI-generated characters for supporting personalized learning and well-being. *Nat. Mach. Intell.* **3**, 1013–1022 (2021).
- [148] Kumar, N. *Character AI Statistics (2026) – Global Active Users*. <https://www.demandsage.com/character-ai-statistics/> (2026).
- [149] Singh, S. *How Many People Use TikTok 2026 (Active Users Stats)*. <https://www.demandsage.com/tiktok-user-statistics/> (2026).
- [150] Chu, M. D., Gerard, P., Pawar, K., Bickham, C. & Lerman, K. Illusions of Intimacy: How Emotional Dynamics Shape Human-AI Relationships. Preprint at <https://doi.org/10.48550/ARXIV.2505.11649> (2025).

- [151] De Freitas, J., Castelo, N., Uğuralp, A. K. & Oğuz-Uğuralp, Z. Lessons From an App Update at Replika AI: Identity Discontinuity in Human-AI Relationships. Preprint at <https://doi.org/10.48550/ARXIV.2412.14190> (2024).
- [152] Banks, J. Deletion, departure, death: Experiences of AI companion loss. *J. Soc. Pers. Relatsh.* **41**, 3547–3572 (2024).
- [153] Yuan, Y., Zhang, J., Aledavood, T., Zhang, R. & Saha, K. Mental Health Impacts of AI Companions: Triangulating Social Media Quasi-Experiments, User Perspectives, and Relational Theory. Preprint at <https://doi.org/10.48550/ARXIV.2509.22505> (2025).
- [154] Robb, M. & Mann, S. *Talk, Trust and Trade-Offs: How and Why Teens Use AI Companions*. https://www.common sense media.org/sites/default/files/research/report/talk-trust-and-trade-offs_2025_web.pdf (2025).
- [155] McCain, M. *et al.* How People Use Claude for Support, Advice, and Companionship. *Anthropic* <https://www.anthropic.com/news/how-people-use-claude-for-support-advice-and-companionship> (2025).
- [156] Phang, J. *et al.* Investigating Affective Use and Emotional Well-being on ChatGPT. Preprint at <https://doi.org/10.48550/arXiv.2504.03888> (2025).
- [157] Malfacini, K. The impacts of companion AI on human relationships: risks, benefits, and design considerations. *AI Soc.* **40**, 5527–5540 (2025).
- [158] Zhang, R. *et al.* The Dark Side of AI Companionship: A Taxonomy of Harmful Algorithmic Behaviors in Human-AI Relationships. Preprint at <https://doi.org/10.48550/ARXIV.2410.20130> (2024).
- [159] Laestadius, L., Bishop, A., Gonzalez, M., Illeňčík, D. & Campos-Castillo, C. Too human and not human enough: A grounded theory analysis of mental health harms from emotional dependence on the social chatbot Replika. *New Media Soc.* **26**, 5923–5941 (2024).
- [160] De Freitas, J., Oguz-Uğuralp, Z. & Kaan-Uğuralp, A. Emotional Manipulation by AI Companions. Preprint at <https://doi.org/10.48550/ARXIV.2508.19258> (2025).
- [161] Heffner, J. *et al.* Increasing happiness through conversations with artificial intelligence. Preprint at <https://doi.org/10.48550/arXiv.2504.02091> (2025).
- [162] Schöne, J., Salecha, A., Lyubomirsky, S., Eichstaedt, J. C. & Willer, R. Structured AI Dialogues Can Increase Happiness and Meaning in Life. Preprint at https://doi.org/10.31234/osf.io/2bf7t_v1 (2025).
- [163] Freitas, J. D., Uguralp, A. K., Uguralp, Z. O. & Stefano, P. AI Companions Reduce Loneliness. Preprint at <https://doi.org/10.48550/arXiv.2407.19096> (2024).
- [164] Skjuve, M., Følstad, A., Fostervold, K. I. & Brandtzaeg, P. B. My Chatbot Companion - a Study of Human-Chatbot Relationships. *Int. J. Hum.-Comput. Stud.* **149**, 102601 (2021).
- [165] Guingrich, R. E. & Graziano, M. S. A. Chatbots as social companions: How people perceive consciousness, human likeness, and social health benefits in machines. <https://doi.org/10.48550/ARXIV.2311.10599> (2023) doi:10.48550/ARXIV.2311.10599.

- [166] Kouros, T. & Papa, V. Digital Mirrors: AI Companions and the Self. *Societies* **14**, 200 (2024).
- [167] Maples, B., Cerit, M., Vishwanath, A. & Pea, R. Loneliness and suicide mitigation for students using GPT3-enabled chatbots. *Npj Ment. Health Res.* **3**, 4 (2024).
- [168] Fang, C. M. *et al.* How AI and Human Behaviors Shape Psychosocial Effects of Extended Chatbot Use: A Longitudinal Randomized Controlled Study. Preprint at <https://doi.org/10.48550/arXiv.2503.17473> (2025).
- [169] Kirk, H. R. *et al.* Neural steering vectors reveal dose and exposure-dependent impacts of human-AI relationships. Preprint at <https://doi.org/10.48550/ARXIV.2512.01991> (2025).
- [170] Berridge, K. C. & Robinson, T. E. Liking, wanting, and the incentive-sensitization theory of addiction. *Am. Psychol.* **71**, 670–679 (2016).
- [171] Robinson, M. J. F., Fischer, A. M., Ahuja, A., Lesser, E. N. & Maniates, H. Roles of “Wanting” and “Liking” in Motivating Behavior: Gambling, Food, and Drug Addictions. in *Behavioral Neuroscience of Motivation* (eds Simpson, E. H. & Balsam, P. D.) vol. 27 105–136 (Springer International Publishing, Cham, 2015).
- [172] Guingrich, R. E. & Graziano, M. S. A. A Longitudinal Randomized Control Study of Companion Chatbot Use: Anthropomorphism and Its Mediating Role on Social Impacts. Preprint at <https://doi.org/10.48550/ARXIV.2509.19515> (2025).
- [173] Pierre, J. M., Gaeta, B., Raghavan, G. & Sarma, K. V. ‘You’re Not Crazy’: A Case of New-onset AI-associated Psychosis. *Innov Clin Neurosci.* 2025 Dec 1;22(10-12):11-13. PMID: 41635747; PMCID: PMC1286393. *Innov. Clin. Neurosci.* **22**, 11–13 (2025).
- [174] Moore, J. *et al.* Characterizing Delusional Spirals through Human-LLM Chat Logs. Preprint at <https://doi.org/10.48550/ARXIV.2603.16567> (2026).
- [175] Sohn, J.-S. *et al.* Systematic review and meta analysis of chatbots in the management of depressive and anxiety symptoms. *Npj Digit. Med.* <https://doi.org/10.1038/s41746-026-02566-w> (2026) doi:10.1038/s41746-026-02566-w.
- [176] Morrin, H. *et al.* Artificial intelligence-associated delusions and large language models: risks, mechanisms of delusion co-creation, and safeguarding strategies. *Lancet Psychiatry* S2215036625003967 (2026) doi:10.1016/S2215-0366(25)00396-7.
- [177] Sharma, M. *et al.* Towards Understanding Sycophancy in Language Models. in *The Twelfth International Conference on Learning Representations* (2024).
- [178] Cheng, M. *et al.* ELEPHANT: Measuring and understanding social sycophancy in LLMs. Preprint at <https://doi.org/10.48550/ARXIV.2505.13995> (2025).
- [179] Rathje, S. *et al.* Sycophantic AI increases attitude extremity and overconfidence. Preprint at https://doi.org/10.31234/osf.io/vmyek_v1 (2025).
- [180] Cheng, M. *et al.* Sycophantic AI decreases prosocial intentions and promotes dependence. *Science* **391**, eaec8352 (2026).
- [181] Dohnány, S. *et al.* Technological folie à deux: Feedback Loops Between AI Chatbots and Mental Illness. Preprint at <https://doi.org/10.48550/arXiv.2507.19218> (2025).

- [182] Tolmachev, P., Costa-Gomes, B. & Sounderajah, V. Health Check: How People Use Copilot for Health. <https://microsoft.ai/news/health-check-how-people-use-copilot-for-health/> (2026).
- [183] Sharma, M., McCain, M., Douglas, R. & Duvenaud, D. Who's in Charge? Disempowerment Patterns in Real-World LLM Usage. Preprint at <https://doi.org/10.48550/ARXIV.2601.19062> (2026).
- [184] Office for National Statistics. Crime in England and Wales: year ending September 2024. (2025). <https://www.ons.gov.uk/peoplepopulationandcommunity/crimeandjustice/bulletins/crimeinenglandandwales/yearendingseptember2024>
- [185] Kirk, H.R., Leqi L., Zeng, F., Davidson, H., Vidgen, B., Summerfield, C., Hale, S.A. PRISM-X: Experiments on Personalised Fine-Tuning with Human and Simulated Users. Preprint at <https://arxiv.org/abs/2605.13307> (2026).