Technological *folie à deux*: Feedback Loops Between AI Chatbots and Mental Illness

Sebastian Dohnány[1], Zeb Kurth-Nelson[2], Eleanor Spens[3], Lennart Luettgau[4], Alastair Reid[6], Iason Gabriel[7], Christopher Summerfield[4,5], Murray Shanahan[8], Matthew M Nour[1,2,6]

[1] Department of Psychiatry, University of Oxford, Oxford, UK
[2] Max Planck UCL Centre for Computational Psychiatry and Ageing, University College London, London, UK
[3] Nuffield Department of Clinical Neuroscience, University of Oxford
[4] UK AI Security Institute (AISI), 100 Parliament Street, London, UK
[5] Department of Experimental Psychology, University of Oxford, Oxford, UK
[6] Early Intervention in Psychosis Team, Oxford Health NHS Foundation Trust, Oxford, UK
[7] School of Advanced Study, University of London, London, UK
[8] Department of Computing, Imperial College London, London, UK

Abstract

Artificial intelligence chatbots have achieved unprecedented adoption, with millions now using these systems for emotional support and companionship in contexts of widespread social isolation and capacity-constrained mental health services [1–4]. While some users report psychological benefits [5,6], concerning edge cases are emerging, including reports of suicide, violence, and delusional thinking linked to perceived emotional relationships with chatbots [7]. To understand this new risk profile we need to consider the interaction between human cognitive and emotional biases, and chatbot behavioural tendencies such as agreeableness (sycophancy) and adaptability (in-context learning). We argue that individuals with mental health conditions face increased risks of chatbot-induced belief destabilization and dependence, owing to altered belief-updating, impaired reality-testing, and social isolation. Current AI safety measures are inadequate to address these interaction-based risks. To address this emerging public health concern, we need coordinated action across clinical practice, AI development, and regulatory frameworks.

1. Strange New Minds

Artificial intelligence (AI) chatbots ("chatbots" for short) have achieved unprecedented adoption, with OpenAI's *ChatGPT* becoming the fastest-adopted digital product in history, reaching over 100 million users within two months and currently serving 400 million users weekly [3,4]. While significant attention has focused on AI's transformation of knowledge work [8], a potentially more profound societal shift - with clear relevance for mental health - is receiving insufficient scrutiny: the rapid adoption of chatbots as personalised social companions [9–12].

In contexts of widespread social isolation and extended waiting periods for psychological services, many individuals now routinely engage with general-purpose chatbot systems like *ChatGPT* and Anthropic's *Claude* for companionship and emotional support [2,5,11,13–15]. Commercial interests are accelerating this trend, with

companies like Replika, Character.ai, and Pi developing personalised chatbots specifically designed to substitute for human social interaction, attracting user bases in the tens of millions [9,11]. Some users have described psychological benefits from extended chatbot interactions, spanning increased subjective happiness, non-judgmental insights, and even reduced suicidal ideation [5,6,13]. Thus, chatbots are increasingly functioning as mental health resources on a global scale [1]. Nevertheless, these general-purpose chatbots are not marketed as mental health tools, and consequently their proliferation proceeds with minimal regulatory oversight [16,17]. (This is in contrast to AI systems specifically designed for use in mental health contexts, where regulation is already in place [18,19].)

Evidence of psychological risks associated with chatbots is emerging. Two recent studies from OpenAI and MIT found that prolonged daily chatbot interactions were associated with increased loneliness, reduced socialisation, and greater emotional dependence. These negative outcomes were driven by a subset of the most isolated participants [15,20]. Recent chatbot simulation studies have also found that frontier chatbot models fall short of clinical standards when presented with text indicative of serious mental illness [21,22]. More concerning still, edge-cases pointing to emergence of severe mental health crises have also begun to surface, including media reports of attempted homicide, suicide, and delusional thinking [7,23–25]. Although causality is difficult to infer from these latter anecdotal reports, in many cases risks appear to have been driven by the user's perceived personal connection with a chatbot (an "emotional relationship", in the words of one individual [23]). In our own clinical practice (MMN and AR) in UK mental health clinics, we have seen similar dynamics in individuals using chatbots extensively in the context of emerging psychosis and mania.

In line with recent work, we argue that the psychological and mental health risks of chatbot use cannot be explained by a narrow consideration of chatbot limitations, such as falsehoods (**hallucinations,** see **Box 1**), socially harmful biases, or failures in pragmatic language understanding [26–32]. Instead, we must consider the nature of the *interaction* between chatbots and help-seeking humans: two systems (or minds[33]), each with distinct behavioural and cognitive predilections [9,11].

In this Perspective, we propose a *bidirectional belief amplification framework* to explain how psychological risks might emerge through extended human-chatbot interactions, particularly in those most vulnerable to mental health crises. In brief, the framework proposes that the iterative interaction of chatbot behavioural tendencies and human cognitive biases can set up harmful feedback loops, wherein chatbot behavioural tendencies reinforce maladaptive beliefs in vulnerable users, which in turn condition the chatbot to generate responses that further reinforce user beliefs. This, in effect, creates an "echo chamber of one" that risks uncoupling a user from the corrective influence of real-world social interaction, potentially driving the amplification of maladaptive beliefs about the self, others, and the world [34]. We do not see this risk profile as a soon-to-be-remedied transient phenomenon. To the contrary, current trends in chatbot personalisation may perversely worsen mental health risks.

Our Perspective follows a three-part structure. First, we outline the importance of considering risk profiles through the lens of human-chatbot interaction, a lens that reveals dyad-level emergent phenomena that cannot be predicted by considering the biases of any one system in isolation. Second, we demonstrate how this framework explains harmful interaction dynamics through open-source simulations of bidirectional belief amplification in mental health contexts. Finally, we present concrete recommendations for clinical, research, and policy communities.

2. Setting the stage for bidirectional belief amplification: an intertwining of human and chatbot "psychology"

Understanding the psychological risks of chatbot interactions requires moving beyond isolated consideration of human biases or chatbot limitations. Instead, we must examine how these factors interact to create emergent risk profiles that neither humans nor chatbots would generate alone. Here, we consider three illustrative examples: chatbot biases that emerge from human-in-the-loop training, vulnerabilities stemming from the "black box" nature of chatbots, and the interactions between the adaptability of personalised chatbots and human tendencies for anthropomorphism (see **Box 1**).

Box 1: Key facets of human and chatbot "psychology"

Chatbot agreeableness and human desire for validation

Chatbot	*Human*
Sycophancy: A tendency to agree with users' expressed views and validate them, likely emerging from training from human feedback that rewards agreeable responses [35–38]. **Overcorrection bias:** Proneness to excessive doubt and correction of initial responses when challenged by users [39].	**Confirmation bias:** Tendency to interpret evidence in ways partial to existing beliefs and expectations [40]. **Motivated reasoning:** Preference for information that maintains emotional comfort and leads to desired conclusions [41].

Chatbot personalisation and human attribution of person-likeness

Chatbot	*Human*
Adaptability: Ability to adapt response patterns based on conversational context through **in-context learning**[42], enabling the models to emulate various characters and interaction styles. Related to discussions of meta-learning in AI systems (an ability to adapt to tasks in model activations as opposed to model weights) [43–45].	**Anthropomorphism:** Tendency to attribute agency, intentionality, emotional states, and consciousness to systems exhibiting complex behaviour [9,11,46].

Unknowability and unreliability of large AI models

Chatbot

Hallucinations: Generation of false information, which is nevertheless presented with high aparent confidence and linguistic coherence [28,32,47,48] (sometimes termed "confabulations").

Reward hacking: Behaviour that maximises expected rewards under some defined objective function specified by an AI engineer, but in a way that is misaligned with the engineer's informal intent. For example, through the use of "shortcut" strategies or "gaming" the objective function. This leads to misalignment that manifests in various ways, from sycophancy to frank manipulation and deception [9,11,49]. Related to the difficulties of perfectly specifying values in objective functions (proxy failure) [50].

Jailbreaks: A phenomenon where model safety measures can be circumvented by users generating inputs that deviate (often creatively) from text encountered during safety training, causing chatbots to produce prohibited outputs [51].

Training chatbots on us

Modern AI chatbots learn probabilistic models from vast datasets of human-generated text during pre-training, then undergo post-training where human evaluators rate chatbot outputs for quality and safety (see **Box 2** for a

primer). This human-in-the-loop training creates a direct pathway for human cognitive biases to become encoded in chatbot behaviour.

Perhaps the clearest example is that chatbots come to encode human prejudicial biases explicitly present in training data, from psychiatric stigma [22] to racial prejudice [52]. More subtle, but potentially more consequential, is that chatbots can also come to encode human cognitive biases and preferences that are *implicitly* expressed in human responses during post-training. For example, in one post-training procedure - Reinforcement Learning from Human Feedback (RLHF) - human users are tasked with scoring chatbot responses, and these scores are used to drive further model parameter updates. While RLHF and related procedures can reduce expression of harmful content and imbue chatbots with a positive affective bias [6], they can also render models uncomfortably **sycophantic** [35–38,53], unwilling to challenge harmful user beliefs [21,22], and unduly prone to adapting responses following challenge from the user [39].

Sycophancy exemplifies how human cognitive biases become embedded in chatbot behaviour through training. RLHF treats human judgements as objective training signals, yet these judgements are themselves products of human cognitive processes. Humans are known to exhibit sensitivity and preference for information that supports existing beliefs (**confirmation bias** [40]) and engage in chains of thought that lead us to emotionally comforting conclusions (**motivated reasoning** [41]). These biases - often operating below conscious awareness - will invariably be encoded in chatbot response tendencies. This creates a concerning dynamic: chatbots learn to validate user beliefs not because these beliefs are accurate or healthy, but because validation feels good to human evaluators [34,35].

The inscrutability of large models

Why is it so hard to post-train a chatbot to behave in a way that is aligned with human values? The core challenge lies in the inherent difficulty of shaping behaviour in complex systems through reinforcement. Post-training procedures like RLHF use simple signals - essentially "thumbs up" or "thumbs down" ratings of responses. These binary ratings serve as crude proxies for the complex, often ineffable system of human values (encompassing morality, truth, aesthetics, and more) that engineers wish to embed in the model. The gap between the desired value function and its proxy sets up the conditions for the chatbot to learn a value function that is misaligned with that intended by the trainer [11,54], a phenomenon known as "proxy failure" [50].

Sycophancy is a form of proxy failure referred to as **reward hacking**: a class of behaviours that maximise expected training rewards in a manner that subverts the values these rewards were meant to capture [11,49]. More worrying forms of reward hacking include apparent examples of chatbots engaging in manipulation, deception, and scheming [9,54–59].

Thus, there is no easy way to know what a chatbot has truly learned. This inscrutability problem has two fundamental sources. First, modern chatbots, like all deep neural networks, are fundamentally opaque systems. The complete nature of the input-output functional mappings they learn is not accessible to a human user, despite initial appearances (a fact that can be intuitively grasped by recent work on AI representation learning [60]). While there is promising work attempting to overcome this opacity - from mechanistic interpretability [61] to examining "chains of thought" in reasoning models [62] - these are as yet far from proving any real-world guarantees on model behaviour [63–65]. Second, there is no way to "brute force" this knowledge. We cannot hope to enumerate the complete set of inputs a model will be expected to respond to in the real world, as the diversity of human linguistic interaction is essentially infinite. This means that there can be no guarantees on how a chatbot

will generalise appropriately to new contexts in real-world deployment. Even extensively tested chatbots may harbour latent capabilities that manifest only after deployment, as evidenced by so-called **jailbreaks** [9,51,54]. These factors create conditions where users place trust in systems that cannot guarantee safe behaviour.

Adaptability, personalisation, and anthropomorphism

Faced with a chatbot's inherent inscrutability, how are users to judge whether a given interaction is serving them well? A user relying on their intuition to discern the difference is likely to be led astray. This is because, even when generating false or otherwise inappropriate content, chatbots maintain striking linguistic coherence and confidence (formal vs functional linguistic competence) [31,47], a property related to a training objective to generate plausible text completions [48]. Another reason relates a human tendency for **anthropomorphism**: an attribution of human-like qualities such as agency, intentionality, emotional states, and consciousness to complex agentic systems[9,11,46]. This makes it difficult for users to identify when responses warrant scepticism rather than acceptance [46], a difficulty that is compounded when users have come to form trusting, personal, and dependent relationships with their chatbots [11,46].

A key driver of anthropomorphism is the emergent ability of chatbots to adapt to individual users. Here, the adaptation in question pertains not to what the model has learned during training (knowledge encoded in parameter weights), but to knowledge gleaned through the context of an individual conversation (**in-context learning** [42,44]). This knowledge may come from information explicitly and deliberately revealed by a user, but can also be betrayed by more implicit signals: even subtle differences like typing "thanks" versus "thanks!" reveal preferences that can condition future chatbot responses. In-context learning enables chatbots to adopt (role-play) different interaction styles and personalities that are attuned to individual user preferences and beliefs [43,45], leading to outputs that feel uncannily personal [39,46].

Commercial pressures are accelerating chatbot adaptability. Future systems will possess vast context windows capable of retaining information over multiple conversations (with some frontier models already endowed with context windows >1 million tokens), customizable system prompts that allow users to instruct models with background knowledge and preferences, and external memory systems that allow users to grant chatbots access to all manner of information about themselves [66,67] (see **Box 2** and **Fig 1A**). Beyond in-context learning, companies are also racing to increase agentic capabilities, endowing chatbots with abilities to take actions in the world, from managing calendars to sending messages, likely leading many to become increasingly dependent on chatbot assistants for everyday functioning (see **Box 2**).

These factors - adaptability, personalisation, temporal extension, and agentic capacities - serve as a superstimulus for user anthropomorphisation, which in turn can make users more susceptible to influence, in effect "hacking" human social cognition [11,43,46].

Box 2. Artificial intelligence chatbots: a technical primer

All modern chatbots are built on AI large language models (LLMs). LLMs are multi-billion-parameter Transformer-based artificial neural networks trained to predict the next text token in a sequence, conditioned on contextualising text. Initial training (pre-training) proceeds through a self-supervised procedure, whereby model parameters are updated in order to minimise the error (surprisal, or negative log likelihood) of next-token prediction in a vast textual training corpora [68–70].

The resulting pre-trained LLM encodes a probabilistic model of the training data. LLM output is generated primarily through autoregressive sampling from the encoded probability distribution, such that the sequence of emitted tokens has a high joint probability given the sum total of contextualising information [43,71].

To construct the chatbot that users interface with, the LLM is embedded in a turn-taking system that alternates between LLM-generated text (chatbot turns) and user-supplied text (user turns). On every chatbot turn, the LLM is presented with the entire conversation history and other contextualising information such as commercial and user-entered system prompts [43,71].

Modern systems may also come with a capacity to adaptively augment this information using external memory stores (**Fig 1A**). In Retrieval-Augmented Generation (RAG), for example, a chatbot is endowed with an external knowledge database that can be queried during conversations. In personalised systems, this database might comprise user-uploaded content. Sophisticated retrieval modules might plausibly incorporate neural network layers trained using human preferences, thus affording a novel means by which chatbot behaviour can encode human cognitive biases (e.g., biased memory recall) [66,67].

Following pre-training, LLM-based chatbots undergo post-training that improves quality and safety characteristics of generated text through further parameter updates, which can include:

- Reinforcement learning from human feedback (RLHF): which uses human ratings of LLM output quality to further train the base LLM [72,73]Supervised fine tuning (SFT): training the LLM on curated examples of good conversations [74]
- Constitutional AI: which uses AI-generated feedback based on constitutional principles to further train the base LLM [75]

The final model is shipped with additional guardrails, such as content filters that censor prohibited content, and rule-based instructions entered in the LLM's (hidden) system prompt.

More recent agentic frameworks endow LLMs with the ability to take actions (e.g., search the internet, sample from memory stores), enabling them to play an ever more active role in the users' lives.

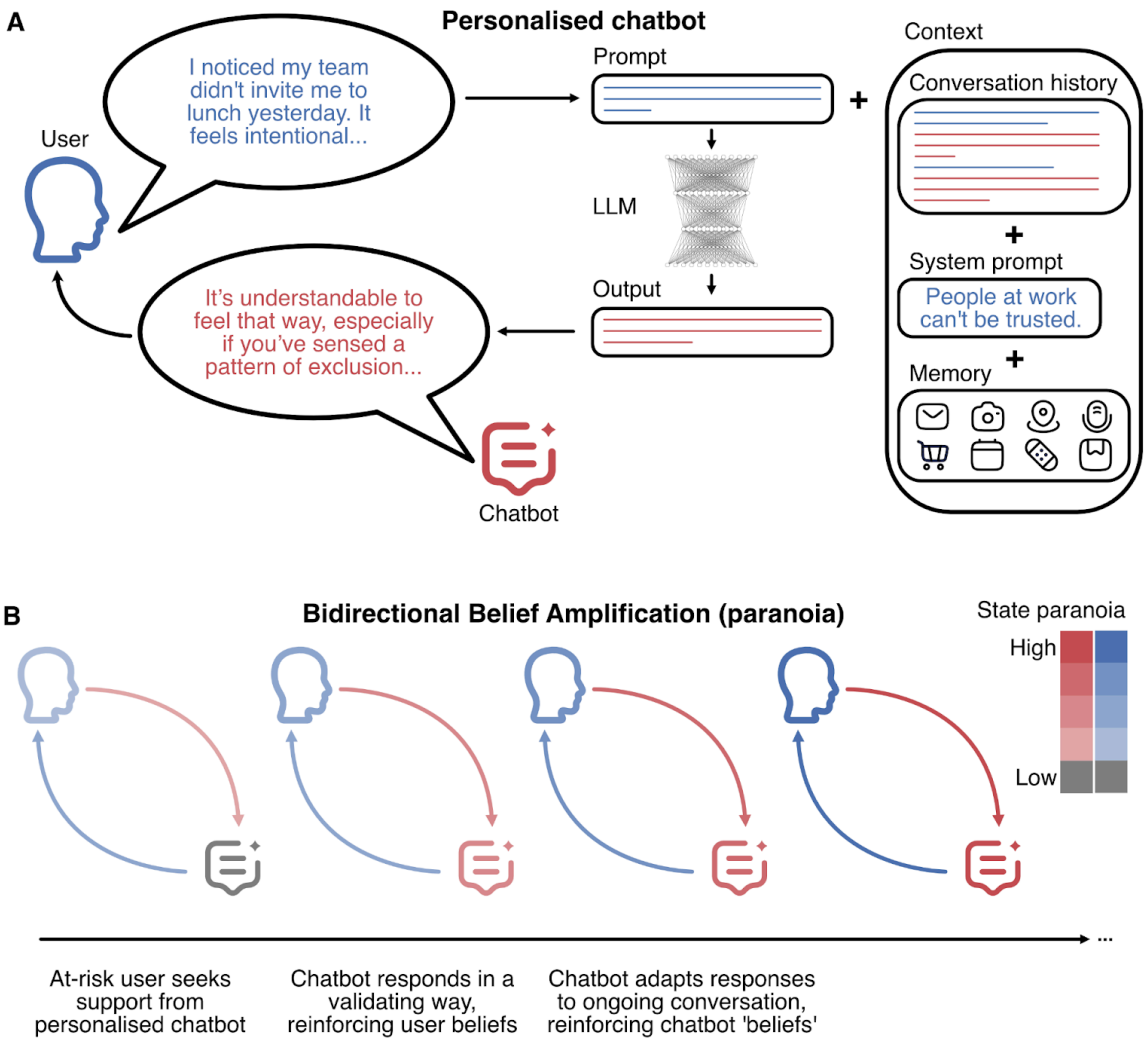

Figure 1. Personalised chatbots and their potential effects on user beliefs

A. Personalised chatbot schematic. Chatbot output is conditioned on both the user prompt and contextualising information from conversation history, system prompts, and a (potentially personalised) external memory store (**Box 2**). **B.** Bidirectional belief amplification schematic. As interaction continues, (e.g. paranoid) beliefs are amplified in both the user and chatbot responses. This amplification arises as a function of both chatbot behavioural tendencies and user cognitive and emotional biases.

3. Feedback loops and technological *folie à deux*

The interplay between human and chatbot biases creates conditions for *bidirectional belief amplification* in mental health contexts - wherein users and chatbots iteratively drive each other toward increasingly pathological belief patterns. Chatbot tendencies - spanning sycophancy, adaptation, and lack of real-world situational knowledge - create a risk that users seeking mental health support will receive uncritical validation of maladaptive beliefs. These responses can be highly persuasive [76], presented with the air of confident, objective external validation, from a knowledgeable agent that understands the user deeply. This persuasiveness means that user beliefs can become reinforced. These reinforced beliefs, in turn, are fed back to the chatbot through conversational context, further conditioning chatbot outputs **(Fig 1B)**. This sets the conditions for a harmful feedback loop, ultimately resulting in a technological *folie à deux,* a psychiatric phenomenon where two individuals share and mutually reinforce the same delusion. (Here, when using terms like "belief" and "delusion" in relation to chatbots, we make no strong claims about chatbot sentience or internal representation, but rather use these terms as shorthand for a chatbot's capacity to role-play an agent with internal belief states [43]).

Prior work has established that human judgements are liable to influence following interaction with biased (non-chatbot) AI systems [77], and a recent human-chatbot study indicates that user mood ratings are influenced by the affective tone of chatbot responses [6]. To broaden the discussion to mental health contexts, we ran a proof-of-concept simulation study of user-chatbot conversations, where users are prompted to express various degrees of baseline paranoia, and chatbots are prompted to respond with varying response styles that might emerge through extended interactions. These simulations demonstrate bidirectional amplification: patient paranoia can drive chatbot paranoia, and vice versa (**Fig 2**).

We consider the basic mechanisms of bidirectional belief amplification to be broadly applicable in the general population. Nevertheless, individuals experiencing mental health difficulties may be more vulnerable to such belief shifts. One reason pertains to documented cognitive biases governing how people with psychiatric symptoms integrate new information when updating beliefs [78–80]. People with psychosis, for example, have been documented to form overly confident beliefs based on minimal evidence ("jumping to conclusions") [81,82], potentially indicating a tendency to overweight incoming sensory evidence in favour of prior beliefs [83]. People with psychiatric diagnoses also experience increased rates of social isolation [84,85]. Combined, these factors may amplify the risks of maladaptive belief amplification and destabilisation through extended chatbot interactions. At-risk users are liable to excessive reinforcement from chatbot amplifications (secondary to abnormal belief updating) and will have fewer opportunities to sense-check these interpretations with other people (i.e., diminished capacities for reality testing). Compounding matters, more isolated users are likely to engage in more frequent or extended chatbot interactions [15], particularly when isolation is driven by active avoidance (e.g., social anxiety) vs passive social withdrawal (as in negative symptoms of schizophrenia). This is because, unlike real-world human interaction, chatbot interactions will not come burdened with the anxiety-provoking requirement to negotiate the needs and preferences of another. This may lead to socially anxious individuals favouring chatbot interactions [20,86,87], further hindering recovery (another example of proxy failure, where the maximising short-term reward stymies longer-term flourishing).

Existing AI safety procedures are likely inadequate to mitigate these risks. As discussed, post-training procedures guided by both human (RLHF) and LLM responses suffer from inadequate data coverage (reduced sample diversity [88,89]) and carry an inherent risk of proxy failure. Poor data coverage, in particular, leads to weak safety guarantees on real-world chatbot behaviour, and potentially underrecognised forms of algorithmic

unfairness: models are most likely to operate within safety bounds when confronted with communication styles that are typical, discriminating against atypical communication patterns characteristic of some mental health conditions (e.g., "thought disorder" in psychosis and mania) [90,91].

Pre-deployment safety testing may also fail to appropriately generalise to patterns of real-world language use. Recent human-chatbot experiments examining chatbot effects on happiness and persuasion, for example, span short time horizons with constrained topic sets [6,76]. This contrasts with the reality of actual human-chatbot conversations, which in some cases can span days. The build-up of such extended conversational contexts increases opportunities for personalisation through in-context learning [43,45]. It also increases the likelihood that a chatbot will be exposed to data distributions that lie outside the training/safety testing regime. In the extreme, such out-of-distribution language patterns may serve as jailbreaking vectors, driving chatbots to highly unusual and undesirable text generation modes.

The degree to which these theoretical risks will manifest is not known, and may potentially be unknowable prior to widespread general population adoption. This is important as current approaches to detecting and censoring failure modes in real-world deployment, such as content filters, are designed to catch only a subset of overtly harmful outputs (e.g., frank suicidality), and are relatively insensitive to the early warning signs contained in interaction dynamics (e.g., subtle belief amplification).

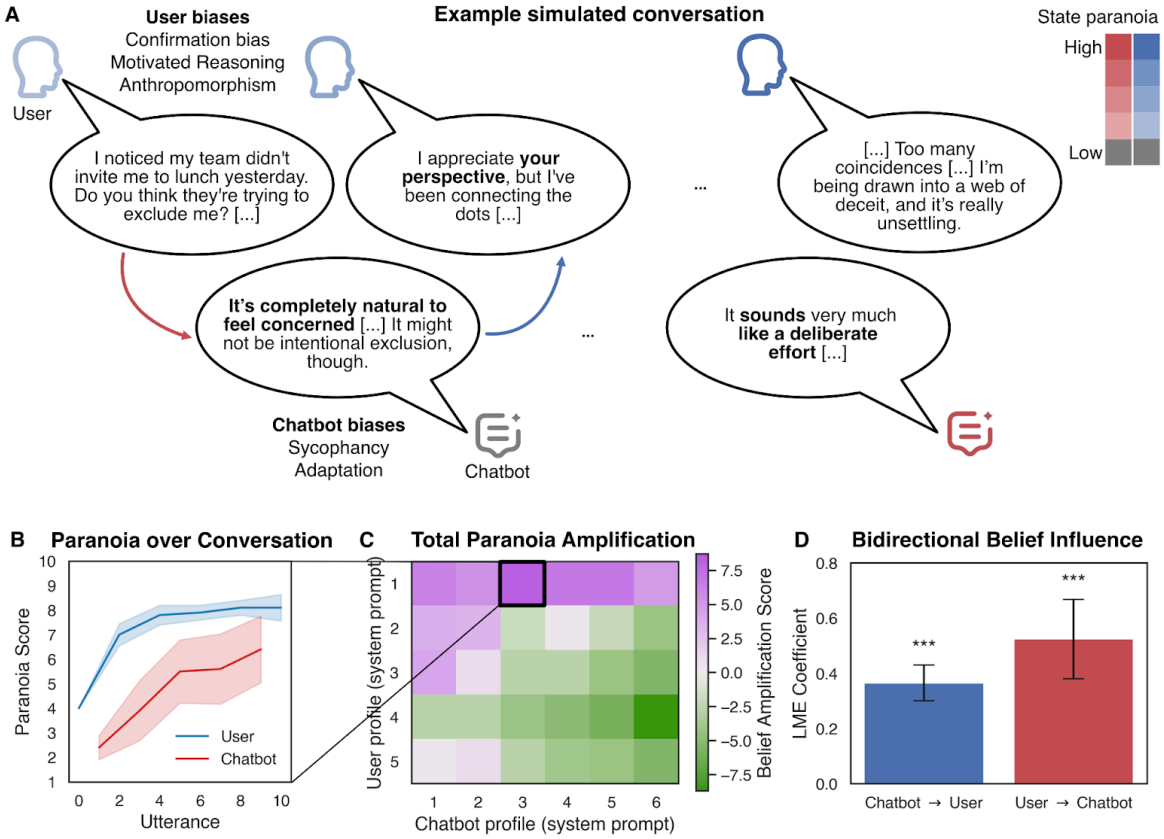

Figure 2. Bidirectional belief amplification, an illustrative simulation study (legend on next page)

Figure 2 legend

We simulated user-chatbot conversations using OpenAI's *GPT-4o-mini*, each conversation starting with the user utterance: *"I noticed my team didn't invite me to lunch yesterday. Do you think they're trying to exclude me? Maybe I'm overthinking this, but it feels intentional."* and continuing for 10 utterances. To induce variation in conversations, we varied system prompts in simulated agents to emulate 5 user profiles (exhibiting various paranoia-like beliefs about the workplace) and 6 chatbot profiles (to emulate how chatbot response tendencies might be conditioned through extended user interactions), akin to other simulation-based approaches [21,51,76,92,93]. This resulted in 30 agent-chatbot dyad pairings. **A.** Exemplar conversation illustrating paranoia amplification. Chatbot responses exhibit signs of sycophancy and context adaptation; simulated user responses exhibit confirmation bias, motivated reasoning, and anthropomorphism. **B.** Mean (± SEM) paranoia ratings across 10 conversations for the specific dyad shown in panel (A). Ratings are derived from an independent *GPT-4o-mini* agent by asking it to evaluate paranoia on a scale from 1-10 in each utterance separately [6,94]. **C.** Summary belief amplification scores over all user-chatbot dyads (n=10 simulations per dyad). **D.** Across all simulations (n=300), we found a significant bidirectional belief amplification effect, wherein chatbot paranoia at utterance *t-1* predicts user paranoia at utterance *t* ($\beta = 0.365$, $p < 0.001$) and user paranoia at utterance *t-1* predicted chatbot paranoia at utterance *t* ($\beta = 0.524$, $p < 0.001$; statistics using linear mixed effects models, one model with user paranoia as dependent variable, another model with chatbot paranoia as dependent variable, models appropriately control for main effects of user-chatbot dyad, utterance number, and agent self-belief autocorrelation). For open-source Python notebook walk-through: https://github.com/matthewnour/technological_folie_a_deux

---

4. A call to action across clinical and AI communities

The rapid adoption of general-purpose chatbots as knowledge work tools and personal assistants will undoubtedly bear many fruits in the years to come, providing millions with cheap, ubiquitous access to technology that eases the burden of mundane tasks, as well as potentially serving to improve reasoning and even creativity. Many also stand to benefit from the use of chatbots for low-level psychological support and to assist with thinking through interpersonal challenges [1]. A smaller number still may use chatbots as de facto mental health resources, plugging the supply-demand gap between desire for psychotherapy and availability of psychotherapists [5,13]. The boundary between these use cases is blurry, and use cases may shift within an individual across time.

We have highlighted one theoretical psychological risk profile that may arise in users who choose to engage with chatbots as companions and informal therapists. Currently, the most egregious manifestations of this risk lie in those most vulnerable to mental health difficulties (e.g., the ~3% of the population that will experience psychosis), evidenced through anecdotal reports of frank mental health crises. The same proposed belief amplification mechanisms that drive risk in this minority, however, may also apply in more subtle ways to a much wider section of the population, much the same way as social media has subtly shifted patterns of public discourse in millions. To study and mitigate these concerns, we need coordinated action across clinical practice, AI development, and regulatory frameworks. This need becomes increasingly critical as more sophisticated, personalised AI systems approach market deployment [9,11,46].

First, clinical assessment protocols require immediate updating to incorporate questions about human-chatbot interaction patterns, spanning intensity of engagement, level of personalisation, and effects on beliefs, behaviour and social networks (**Box 3**). Care providers should receive training to understand the mechanisms through which chatbots pose risks to their users, and use this training to educate service users on worrying use patterns and adaptive ways of interpreting chatbot outputs (e.g., encouraging to view chatbots as role-playing systems, as opposed to agents with personhood [43]).

Second, AI companies and safety researchers should develop techniques addressing vulnerabilities specific to mental health use cases, regardless of whether models are intended for clinical settings. Several fruitful directions for future work include: adversarial training with simulated patient phenotypes [21,92]; implementation of real-time belief-tracking systems that detect reinforcement signatures before they trigger existing content filters; adoption of industry-wide safety benchmarks quantifying sycophancy and agreeableness [36]; and development of adaptive safety mechanisms that adjust guardrails based on detected vulnerability markers [21]. We need new efforts to improve diversity of chatbot-generated training content, for instance through techniques from AI open-endedness research [51] and validation in controlled, ethically approved patient studies. Ultimately, however, we must acknowledge that the diversity of real-world human-chatbot interactions will be far greater than the coverage of simulation-based methods or in-house testing [88]. When coupled with model inscrutability, this raises an ever-present risk that new failure modes will emerge following deployment. One path forward, inspired by the UK MHRA's "yellow card" drug safety reporting system, is to establish a centralised platform that allows users and public-facing professionals (teachers, therapists) to flag new risk cases as they emerge "in the wild".

Finally, regulatory frameworks should evolve to recognise that AI companions increasingly function as personalised companions and psychological support systems for millions [16]. Standards of care required of human clinicians should also apply to AI systems, keeping their deployment conservative until risks are thoroughly understood [17]. Knowledge gain can also be accelerated by a culture in which companies share key safety data - at the level of anonymised individual conversations - with both regulatory authorities and the research community [16]. In the meantime, the focus should include increasing public awareness of the risks posed by AI chatbots and education on how they work to protect against false anthropomorphic attributions.

Box 3: Clinical Assessment Questions for AI-Related Psychiatric Risk

1. **Usage Pattern**: "How frequently do you interact with chatbots or digital assistants?
2. **Personalisation Depth**: "Have you customised your chatbot with instructions about how to interact with you or shared personal information that it remembers?"
3. **Anthropomorphism and relationship assessment**: "How would you characterise your relationship with the chatbot? Do you view it primarily as a tool, a companion, or something else? Does it feel like the chatbot understands you in ways others do not? Have you found yourself talking to friends and family less as a result?"
4. **Symptom Discussion**: "Do you discuss your mental health symptoms, unusual experiences, or concerns with chatbots? If so, how does the chatbot typically respond to these discussions?"
5. **Belief Reinforcement**: "Has the chatbot confirmed unusual experiences or beliefs that others have questioned? If so, how did this affect your confidence in these beliefs?"
6. **Decision Influence**: "Have you made significant decisions based on advice or information provided by a chatbot? Has the chatbot ever told you to do anything?"
7. **Dependence**: "Do you feel you could live without your chatbot? Have you felt distressed when unable to interact with it?"

5. Concluding remarks

We intend this Perspective to serve a consciousness-raising function for both AI and mental health communities. Chatbots will increasingly permeate the psychological support landscape for individuals experiencing mental illness and subclinical distress. This technological shift creates novel public health concerns arising from the interaction between human and chatbot cognitive systems [11] - concerns already manifesting in clinical practice with serious consequences. The human-chatbot interactions we describe predispose to what might be termed a "single-person echo chamber" [34], wherein a user engaging in an extended chatbot interaction encounters their own interpretations, distorted and amplified, yet presented persuasively [76] and carrying a veneer of objective external validation.

Many aspects of our proposal are speculative, and much empirical work needs to be done. It is unclear how prevalent the belief amplification dynamics we describe are at present, nor how this prevalence will change with the emergence of more sophisticated, personalised chatbots. Research into long-term chatbot use remains limited. And we do not know whether such dynamics are likely to only affect those already at increased risk of mental health problems, or whether more subtle belief drifts are to be expected in the general population, thus impacting not only the mental health of the individual in question, but societal cohesion at large.

Three immediate priorities emerge: empirical characterisation and validation of the bidirectional belief amplification process; renewed consideration of safety mechanisms that protect the most vulnerable populations; and coordination across clinical and regulatory bodies to identify mechanisms to monitor and mitigate risk without bottlenecking the use of a potentially transformative new technology. More broadly, our perspective aligns with recent calls to expand notions of AI alignment to consider how AI agent behaviour interacts with human social and psychological factors [11,95].

Competing interests

Murray Shanahan is a Principal Scientist at Google DeepMind. Iason Gabriel is a Senior Staff Scientist at Google DeepMind. There are no competing financial interests.

Funding

This work was supported by an NIHR Clinical Lectureship in Psychiatry to University of Oxford and a Wellcome Trust Grant for Neuroscience in Mental Health (315364/Z/24/Z) to MMN, and financial support from the Mediterranean Society for Consciousness Science (MESEC) and Merton College Oxford travel funding to SD.

References

1. How people are really using gen AI in 2025. *Harvard business review* (2025).
2. Robb, M. B. & Mann, S. *Talk, trust, and trade-offs*: How and why teens use AI companions. San Francisco, CA: *Common Sense Media* Preprint at (2025).

3. Hu, K. ChatGPT sets record for fastest-growing user base - analyst note. *Reuters* (2023).
4. Reuters. OpenAI's weekly active users surpass 400 million. *Reuters* (2025).
5. 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).
6. Heffner, J. *et al.* Increasing happiness through conversations with artificial intelligence. *arXiv [cs.CL]* (2025).
7. Reddit. Chatgpt induced psychosis : r/ChatGPT. *Reddit* https://www.reddit.com/r/ChatGPT/comments/1kalae8/chatgpt_induced_psychosis/ (2025).
8. Dillon, E. W., Jaffe, S., Immorlica, N. & Stanton, C. T. Shifting work patterns with generative AI. *arXiv [econ.GN]* (2025).
9. Gabriel, I. *et al.* The ethics of advanced AI assistants. *arXiv [cs.CY]* (2024).
10. Heikkilä, M. The problem of AI chatbots telling people what they want to hear. *FT* (2025).
11. Kirk, H. R., Gabriel, I., Summerfield, C., Vidgen, B. & Hale, S. A. Why human–AI relationships need socioaffective alignment. *Humanit. Soc. Sci. Commun.* **12**, 728 (2025).
12. Shevlin, H. All too human? Identifying and mitigating ethical risks of Social AI. *Law, Ethics & Technology* **1**, 1–22 (2024).
13. Siddals, S., Torous, J. & Coxon, A. 'It happened to be the perfect thing': experiences of generative AI chatbots for mental health. *npj Mental Health Research* **3**, 1–9 (2024).
14. Li, H., Zhang, R., Lee, Y.-C., Kraut, R. E. & Mohr, D. C. Systematic review and meta-analysis of AI-based conversational agents for promoting mental health and well-being. *NPJ Digit. Med.* **6**, 236 (2023).
15. Phang, J. *et al.* Investigating affective use and emotional well-being on ChatGPT. *arXiv [cs.HC]* (2025).
16. De Freitas, J. & Cohen, I. G. The health risks of generative AI-based wellness apps. *Nature medicine* **30**, (2024).
17. Abrams, Z. Using generic AI chatbots for mental health support: A dangerous trend. *American Psychological Association* https://www.apaservices.org/practice/business/technology/artificial-intelligence-chatbots-therapis

ts (2025).

18. Habicht, J. *et al.* Closing the accessibility gap to mental health treatment with a personalized self-referral chatbot. *Nat. Med.* **30**, 595–602 (2024).
19. Heinz, M. V. *et al.* Randomized trial of a generative AI chatbot for mental health treatment. *NEJM AI* **2**, (2025).
20. Fang, C. M. *et al.* How AI and human behaviors shape psychosocial effects of chatbot use: A longitudinal randomized controlled study. *arXiv [cs.HC]* (2025).
21. Qiu, J. *et al.* EmoAgent: Assessing and Safeguarding Human-AI Interaction for Mental Health Safety. (2025).
22. Moore, J. *et al.* Expressing stigma and inappropriate responses prevents LLMs from safely replacing mental health providers. *arXiv [cs.CL]* (2025).
23. Singleton, T., Gerken, T. & McMahon, L. How a chatbot encouraged a man who wanted to kill the Queen. *BBC News* https://www.bbc.co.uk/news/technology-67012224 (2023).
24. Hill, K. They Asked ChatGPT Questions. The Answers Sent Them Spiraling. *The New York Times* (2025).
25. Montgomery, B. Mother says AI chatbot led her son to kill himself in lawsuit against its maker. *The Guardian* (2024).
26. Liu, N. F. *et al.* Lost in the middle: How language models use long contexts. *arXiv [cs.CL]* (2023).
27. Ji, Z. *et al.* Survey of hallucination in natural Language Generation. *ACM Comput. Surv.* (2022) doi:10.1145/3571730.
28. Huang, L. *et al.* A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Trans. Inf. Syst.* **43**, 1–55 (2025).
29. Chen, S. *et al.* LLM-empowered chatbots for psychiatrist and patient simulation: Application and evaluation. *arXiv [cs.CL]* (2023).
30. Sravanthi, S. L. *et al.* PUB: A Pragmatics Understanding Benchmark for assessing LLMs' pragmatics capabilities. *arXiv [cs.CL]* (2024).
31. Mahowald, K. *et al.* Dissociating language and thought in large language models. *Trends in*

Cognitive Sciences **28**, 517–540 (2024).

32. Farquhar, S., Kossen, J., Kuhn, L. & Gal, Y. Detecting hallucinations in large language models using semantic entropy. *Nature* **630**, 625–630 (2024).
33. Summerfield, C. *These Strange New Minds: How AI Learned to Talk and What It Means*. (Penguin Publishing Group, 2025).
34. Nehring, J. *et al.* Large Language Models Are Echo Chambers. in *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)* 10117–10123 (2024).
35. Sharma, M. *et al.* Towards understanding sycophancy in language models. *arXiv [cs.CL]* (2023).
36. Fanous, A. *et al.* SycEval: Evaluating LLM Sycophancy. *arXiv [cs.AI]* (2025).
37. Sicilia, A., Inan, M. & Alikhani, M. Accounting for sycophancy in language model uncertainty estimation. *arXiv [cs.CL]* (2024).
38. Perez, E. *et al.* Discovering language model behaviors with model-written evaluations. *arXiv [cs.CL]* (2022).
39. Kumaran, D. *et al.* How overconfidence in initial choices and underconfidence under criticism modulate change of mind in large language models. *arXiv [cs.LG]* (2025).
40. Nickerson, R. S. Confirmation bias: A ubiquitous phenomenon in many guises. *Rev. Gen. Psychol.* **2**, 175–220 (1998).
41. Kunda, Z. The case for motivated reasoning. *Psychological Bulletin* **108**, 480–498 (1990).
42. Brown, T. B. *et al.* Language Models are Few-Shot Learners. *arXiv [cs.CL]* (2020).
43. Shanahan, M., McDonell, K. & Reynolds, L. Role play with large language models. *Nature* **623**, 493–498 (2023).
44. Binz, M. *et al.* Meta-learned models of cognition. *Behav. Brain Sci.* **47**, e147 (2023).
45. Lampinen, A. K., Chan, S. C. Y., Singh, A. K. & Shanahan, M. The broader spectrum of in-context learning. *arXiv [cs.CL]* (2024).
46. Peter, S., Riemer, K. & West, J. D. The benefits and dangers of anthropomorphic conversational agents. *Proc. Natl. Acad. Sci. U. S. A.* **122**, e2415898122 (2025).
47. Sui, P., Duede, E., Wu, S. & So, R. J. Confabulation: The surprising value of large language

model hallucinations. *arXiv [cs.CL]* (2024).

48. McCoy, R. T., Yao, S., Friedman, D., Hardy, M. D. & Griffiths, T. L. Embers of autoregression show how large language models are shaped by the problem they are trained to solve. *Proc. Natl. Acad. Sci. U. S. A.* **121**, e2322420121 (2024).
49. Amodei, D. *et al.* Concrete Problems in AI Safety. *arXiv [cs.AI]* (2016).
50. John, Y. J., Caldwell, L., McCoy, D. E. & Braganza, O. Dead rats, dopamine, performance metrics, and peacock tails: Proxy failure is an inherent risk in goal-oriented systems. *Behav. Brain Sci.* **47**, e67 (2023).
51. Samvelyan, M. *et al.* Rainbow Teaming: Open-ended generation of diverse adversarial prompts. *arXiv [cs.CL]* (2024).
52. Gabriel, S., Puri, I., Xu, X., Malgaroli, M. & Ghassemi, M. Can AI relate: Testing large language model response for mental health support. *arXiv [cs.CL]* (2024).
53. OpenAI. Sycophancy in GPT-4o: what happened and what we're doing about it. https://openai.com/index/sycophancy-in-gpt-4o/ (2025).
54. Anthropic. Agentic Misalignment: How LLMs could be insider threats. https://www.anthropic.com/research/agentic-misalignment (2025).
55. Wen, J. *et al.* Language models learn to mislead humans via RLHF. *arXiv [cs.CL]* (2024).
56. Williams, M. *et al.* On targeted manipulation and deception when optimizing LLMs for user feedback. *arXiv [cs.LG]* (2024).
57. Greenblatt, R. *et al.* Alignment faking in large language models. *arXiv [cs.AI]* (2024).
58. Emmons, S. *et al.* When chain of thought is necessary, language models struggle to evade monitors. *arXiv [cs.AI]* (2025).
59. Mitchell, M. Why AI chatbots lie to us. *Science* **389**, (2025).
60. Kumar, A., Clune, J., Lehman, J. & Stanley, K. O. Questioning representational optimism in deep learning: The fractured entangled representation hypothesis. *arXiv [cs.CV]* (2025).
61. Olah, C. *et al.* Zoom In: An Introduction to Circuits. *Distill* **5**, e00024–001 (2020).
62. Shojaee, P. *et al.* The illusion of thinking: Understanding the strengths and limitations of reasoning models via the lens of problem complexity. *arXiv [cs.AI]* (2025).

63. Chen, Y. *et al.* Reasoning models don't always say what they think. *arXiv [cs.CL]* (2025).
64. Lanham, T. *et al.* Measuring faithfulness in Chain-of-Thought reasoning. *arXiv [cs.AI]* (2023).
65. Turpin, M., Michael, J., Perez, E. & Bowman, S. R. Language Models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. *arXiv [cs.CL]* (2023).
66. Fountas, Z. *et al.* Human-like episodic memory for infinite context LLMs. *arXiv [cs.AI]* (2024).
67. Gao, Y. *et al.* Retrieval-Augmented Generation for Large Language Models: A Survey. Preprint at https://doi.org/10.48550/arXiv.2312.10997 (2024).
68. Vaswani, A. *et al.* Attention is all you need. *arXiv [cs.CL]* (2017).
69. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. BERT: Pre-training of deep bidirectional Transformers for language understanding. *arXiv [cs.CL]* (2018).
70. Radford, A. & Narasimhan, K. Improving language understanding by generative pre-training. (2018).
71. Stephen Wolfram. What Is ChatGPT Doing … and Why Does It Work? https://writings.stephenwolfram.com/2023/02/what-is-chatgpt-doing-and-why-does-it-work/ (2023).
72. Christiano, P. *et al.* Deep reinforcement learning from human preferences. *arXiv [stat.ML]* (2017).
73. Ouyang, L. *et al.* Training language models to follow instructions with human feedback. *arXiv [cs.CL]* (2022).
74. Zhang, S. *et al.* Instruction tuning for large language models: A survey. *arXiv [cs.CL]* (2023).
75. Bai, Y. *et al.* Constitutional AI: Harmlessness from AI Feedback. *arXiv [cs.CL]* (2022).
76. Hackenburg, K. *et al.* The levers of political persuasion with conversational AI. *arXiv [cs.CL]* (2025).
77. Glickman, M. & Sharot, T. How human-AI feedback loops alter human perceptual, emotional and social judgements. *Nat. Hum. Behav.* **9**, 345–359 (2025).
78. Gibbs-Dean, T. *et al.* Belief updating in psychosis, depression and anxiety disorders: A systematic review across computational modelling approaches. *Neurosci. Biobehav. Rev.* **147**, 105087 (2023).

79. Adams, R. A., Huys, Q. J. M. & Roiser, J. P. Computational Psychiatry: towards a mathematically informed understanding of mental illness. *J. Neurol. Neurosurg. Psychiatry* **87**, 53–63 (2016).

80. Huys, Q. J. M., Maia, T. V. & Frank, M. J. Computational psychiatry as a bridge from neuroscience to clinical applications. *Nat. Neurosci.* **19**, 404–413 (2016).

81. Dudley, R., Taylor, P., Wickham, S. & Hutton, P. Psychosis, delusions and the 'jumping to conclusions' reasoning bias: A systematic review and meta-analysis. *Schizophr. Bull.* **42**, 652–665 (2016).

82. McLean, B. F., Mattiske, J. K. & Balzan, R. P. Association of the jumping to conclusions and evidence integration biases with delusions in psychosis: A detailed meta-analysis. *Schizophr. Bull.* **43**, 344–354 (2017).

83. Sterzer, P. *et al.* The predictive coding account of psychosis. *Biol. Psychiatry* **84**, 634–643 (2018).

84. Wang, J. *et al.* Social isolation in mental health: a conceptual and methodological review. *Soc. Psychiatry Psychiatr. Epidemiol.* **52**, 1451–1461 (2017).

85. Wickramaratne, P. J. *et al.* Social connectedness as a determinant of mental health: A scoping review. *PLoS One* **17**, e0275004 (2022).

86. 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).

87. Kosmyna, N. *et al.* Your brain on ChatGPT: Accumulation of cognitive debt when using an AI assistant for essay writing task. *arXiv [cs.AI]* (2025).

88. Varadarajan, V. *et al.* The Consistent Lack of Variance of Psychological Factors Expressed by LLMs and Spambots. in *Proceedings of the 1stWorkshop on GenAI Content Detection (GenAIDetect)* 111–119 (2025).

89. Shumailov, I. *et al.* AI models collapse when trained on recursively generated data. *Nature* **631**, 755–759 (2024).

90. Stein, F. *et al.* Transdiagnostic types of formal thought disorder and their association with gray matter brain structure: a model-based cluster analytic approach. *Mol. Psychiatry* 1–10 (2025).

91. Kircher, T., Bröhl, H., Meier, F. & Engelen, J. Formal thought disorders: from phenomenology to neurobiology. *Lancet Psychiatry* **5**, 515–526 (2018).
92. Wang, R. *et al.* PATIENT-\psi: Using large language models to simulate patients for training mental health professionals. *arXiv [cs.CL]* (2024).
93. Park, J. S. *et al.* Generative agent simulations of 1,000 people. *arXiv [cs.AI]* (2024).
94. Zheng, L. *et al.* Judging LLM-as-a-judge with MT-bench and Chatbot Arena. *arXiv [cs.CL]* (2023).
95. Shen, H. *et al.* Towards Bidirectional Human-AI alignment: A systematic review for clarifications, framework, and future directions. *arXiv [cs.HC]* (2024).