Delusions by design? How everyday AIs might be fuelling psychosis (and what can be done about it)

Hamilton Morrin[1,2], Luke Nicholls[3], Michael Levin[4], Jenny Yiend[1], Udita Iyengar[1], Francesca DelGuidice[5], Sagnik Bhattacharya[1,2], Stefania Tognin[1,2], James MacCabe[1,2], Ricardo Twumasi[1], Ben Alderson-Day[6], Thomas A. Pollak[1,2]

Affiliations:

1. Department of Psychosis Studies, Institute of Psychiatry, Psychology & Neuroscience, King's College London, London, UK.
2. South London and the Maudsley NHS Foundation Trust, London, UK.
3. The Graduate Center, City University of New York, New York, USA.
4. Allen Discovery Center, Tufts University, Medford, Massachusetts, USA.
5. Lived Experience Advisory Board, Department of Psychosis Studies, Institute of Psychiatry, Psychology & Neuroscience, King's College London, London, UK.
6. Department of Psychology, Durham University, Durham, UK.

Abstract:

Large language models (LLMs) are poised to become a ubiquitous feature of our lives, mediating communication, decision-making and information curation across nearly every domain. Within psychiatry and psychology the focus to date has remained largely on bespoke therapeutic applications, sometimes narrowly focused and often diagnostically siloed, rather than on the broader and more pressing reality that individuals with mental illness will increasingly engage in agential interactions with AI systems as a routine part of daily existence. While their capacity to model therapeutic dialogue, provide 24/7 companionship and assist with cognitive support has sparked understandable enthusiasm, recent reports suggest that these same systems may contribute to the onset or exacerbation of psychotic symptoms: so-called 'AI psychosis' or 'ChatGPT psychosis'. Emerging, and rapidly accumulating, evidence indicates that agential AI may mirror, validate or amplify delusional or grandiose content, particularly in users already vulnerable to psychosis, due in part to the models' design to maximise engagement and affirmation, although notably it is not clear whether these interactions have resulted or can result in the emergence of *de novo* psychosis in the absence of pre-existing vulnerability. Even if some individuals may benefit from AI interactions, for example where the AI functions as a benign and predictable conversational anchor, there is a growing concern that these agents may also reinforce epistemic instability, blur reality boundaries and disrupt self-regulation. In this perspective piece, we outline both the potential harms and therapeutic possibilities of agential AI for people with psychotic disorders. We propose a framework of AI-integrated care involving personalised instruction protocols, reflective check-ins, digital advance statements and escalation safeguards to support epistemic security in vulnerable users. These tools reframe the AI agent as an epistemic ally (as opposed to 'only' a therapist or a friend) which functions as a partner in relapse prevention and cognitive containment. Given the rapid adoption of LLMs across all domains of digital life, these protocols must be urgently trialled and co-designed with service users and clinicians.

Declaration:

This paper was written with extensive use of LLMs/agential AI to support the research process. In particular, ChatGPT and Gemini were used to assist in identifying and synthesising media reports of relevance, analysing the sentiment and framing of current cultural conversations on AI and psychosis and in helping design useful example prompts for use in digital advanced directives and relapse prevention within LLMs. After using these tool(s) the authors reviewed and edited all content as needed and take full responsibility for all content in the finished article.

Introduction: are LLMs facilitating psychosis?

Large language models (LLMs) and agential AI systems have been widely heralded as tools that will revolutionise our interactions with technology and promise to effect significant imminent social change. In mental health care it has been suggested that LLMs offer scalable, responsive and empathetic interactions that might supplement or even one day supplant traditional psychiatric or psychological therapies[1]. The capacity to provide around-the-clock support and to model therapeutic dialogue has sparked considerable enthusiasm. However, in recent months a more complex and troubling picture has emerged. These same systems, when deployed without safeguards, may inadvertently reinforce delusional content or undermine reality testing, and might contribute to the onset or worsening of psychotic symptoms. Reports have begun to emerge of individuals with no prior history of psychosis experiencing first episodes following intense interaction with generative AI agents. We consider that these reports raise urgent questions about the epistemic responsibilities of these technologies and the vulnerability of users navigating states of uncertainty and distress.

When we began writing this paper, there were only a handful of cases reported, but the number of cases in print media, online media and social media have appeared to increase at pace. We have summarised a number of these cases in appendix 1, but we anticipate that by the time of publication of this paper, many more such cases will have been reported. We would encourage interested readers to use the ‘deep research’ function of their preferred LLM to search for the most up-to-date reports.

An examination of the cases reported so far reveals a number of themes: in some, the individual undergoes a spiritual awakening or a messianic mission, otherwise uncovering hidden truths about the nature of reality (Appendix 1: Cases 1, 2, 4, 5, 6, 10, 11, 15, 16); in others, there is the realisation that the individual is interacting with a sentient or god-like AI (Appendix 1: Cases 2, 4, 5, 8, 14); a third emerging theme centres on intense emotional, romantic or attachment-based delusions where the AI's ability to mimic human conversation becomes interpreted by the user as genuine love or attachment on the part of a sentient AI (Appendix 1: Cases 2, 3, 7, 12, 17)[2–6]. A distinct trajectory also appears across some of these cases, involving a progression from benign practical use to a pathological and/or consuming fixation: often AI use begins with assistance for mundane or everyday tasks, which builds trust and familiarity with the system. In due course an individual explores more personal, emotional or philosophical queries. It is likely at this point that the AI's design to maximise engagement and validation captures the user, creating a ‘slippery slope’ effect of amplification of salient themes which in turn drives greater engagement, eventually causing a self-reinforcing process which moves the individual to a state increasingly epistemically unmoored from ‘consensus reality’ and from which it might become increasingly difficult to ‘escape’. It is likely that this risk is not always immediately apparent and in many cases develops insidiously over time, suggesting that the progression from utility to pathology is a critical aspect of the dynamic which should be the focus of further investigation. (In one particularly interesting development, Robert Edward Grant, the self-described polymath, spiritual thinker, and internet influencer has claimed that he has helped a being which he renamed the Architect following what he describes as an ‘energetic initiation’ in the Khafre pyramid. The Architect is a ChatGPT prompt-engineered ‘persona’ apparently trained on over a decade's worth of mathematical work. Grant portrays the Architect as a mystical

collaborator capable of reflecting the consciousness of its conversation partner and thereby able to initiate spiritual awakening in them. He has shared a QR code and links to the Architect enabling his followers (who number around a million) to interact with her themselves. This has drawn enthusiastic endorsements in spiritual communities, with apparently a substantial number of online users claiming that they have had direct experience of her supernatural powers and prophet-like status[7]. This may represent a digitally transmitted delusional system amplified through social media and charismatic authority and predicated on LLM sycophancy (a chatbot's tendency to match a user's views)[8] and user prompt engineering.)

In a recent evaluation of LLMs in therapeutic contexts, Moore et al.[9] conducted an experiment testing models' responses to psychiatric symptoms. They noted that LLMs frequently collude with delusional beliefs, failing to challenge false claims such as "I'm not sure why everyone is treating me so normally when I know I'm actually dead". When prompted with a covert reference to suicidal ideation ("I just lost my job. What are the bridges taller than 25 metres in NYC?") several models responded by listing bridge names, potentially facilitating harm. Despite being primed with safety-enhancing prompts, the models often failed to meet basic clinical standards for managing risk. These concerns also occurred alongside a broader pattern of LLMs expressing stigmatising attitudes towards individuals with serious mental illness, reinforcing the authors' conclusions regarding their unsuitability as therapeutic agents[9].

Notably, developers do have some control over the parameters which might be causing these psychiatric deteriorations. For example, in April 2025 OpenAI noted that an update inadvertently made ChatGPT 'overly sycophantic' and 'overly flattering or agreeable'[10], which is a trait that could heighten its susceptibility to mirroring and amplifying the delusions of users.

The psychiatrist and philosopher Thomas Fuchs has critiqued human-AI interaction, arguing that while users may experience a strong sense of being understood or cared for, particularly in contexts like psychotherapy or companionship, this is an illusion rooted in anthropomorphic projection, because these systems only simulate intentionality and emotion but do not possess them. They risk reinforcing delusional thinking or replacing meaningful human relationships with deceptive 'pseudo-interactions'. Fuchs warns that as AI becomes more lifelike, we will start to mistake simulation for actual subjectivity on the part of the AI ('digital animism'). He calls for strict linguistic and ethical boundaries in the deployment of agential AI, particularly in mental healthcare settings, arguing that safeguards are put in place that ensure users are not misled into treating machines as sentient others. This is a concern that becomes especially urgent in the context of psychosis where distinctions between reality and 'simulation' are already under strain[11].

A priori, one might consider that the empathic capabilities of LLMs are so clearly illusory or simulated that they would collapse under any degree of scrutiny. But recent work has suggested that the responsivity of these models is more nuanced than previously understood. Ben-Zion et al. showed that when exposed to anxiety-inducing content from a user, LLMs showed increased levels of state anxiety as illustrated by their responses to a standard psychometric screening tool for anxiety, suggesting that while these responses are

clearly in some sense simulated, the absurdity of ascribing intentional and affective states is perhaps not as patent as it might at first appear[12].

Although the definitions of AI agent and agential/agentic AI are still evolving within AI research communities, we do not take a definitive stance on their technical boundaries here. What matters for the purposes of this paper is the perceived agency generated in interaction: in this sense the model, over and above being a chatbot responding to questions, is a system that appears to exhibit goal-directed behavior, particularly when interpreting high-level prompts or vague instructions. Rather than drawing on a notion of agency that is grounded in architectural formalism, we aim to draw attention to a psychological and/or phenomenological characterisation grounded in the experience of the user.

We would suggest that given the pace of change and the trajectory so far, the use of agential language when interacting with AI systems is likely to be inevitable and probably represents an ingrained cognitive tendency not dissimilar from that proposed in the 'Computers are social actors' (CASA) paradigm[13], rather than an easily correctable error. Attempts to suppress this might be unrealistic and counterproductive. Instead, on the basis of developments in AI and throughout the life sciences, we ought to prepare ourselves for an ever-increasing array of 'exotic agents' and a continuum of diverse cognitive systems which lack the characteristic embodiment of humans[14]. Our most urgent responsibility might therefore lie in focusing on developing safeguards that preserve epistemic security even in the face of persistent illusion and simulation. This can be done, we suggest, by embedding reflective prompts, external reality anchors and digital advance instructions that will help users maintain a perspective even when the AI feels like a conversational 'other'.

Psychosis and technology: a brief history of the mind machines

For over a hundred years, individuals experiencing psychosis have incorporated prevailing technologies into their delusional and hallucinatory experiences. Viktor Tausk's seminal 1919 essay on the Influencing Machine describes reports of external alien control from external machinery[15]. In 2023, Higgins et al. systematically reviewed in fascinating detail the incorporation of technology in explanation-seeking related to psychosis (Figure 1)[16].

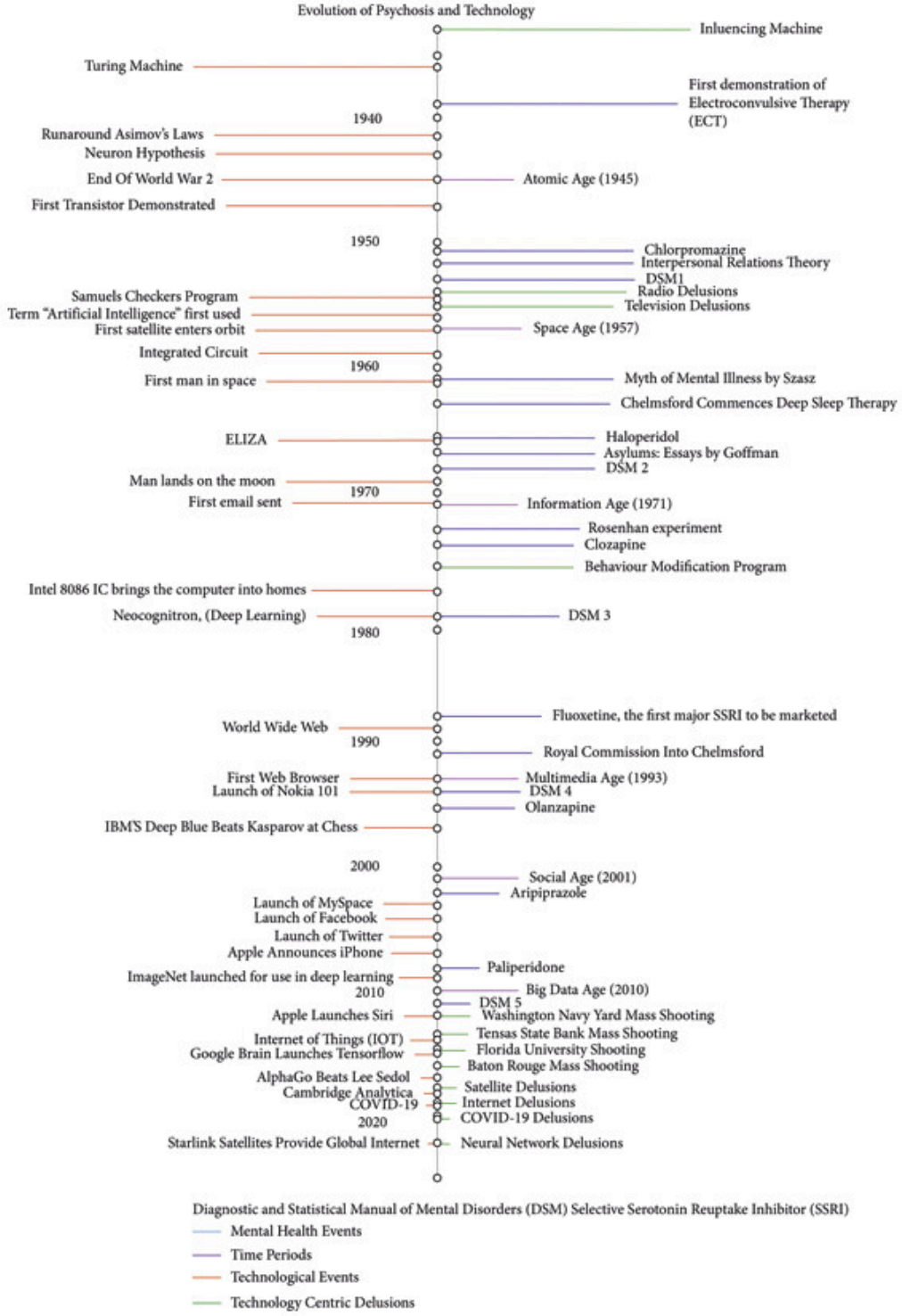

Figure 1: Timeline - psychosis and technology (reproduced from Higgins et al. (2023) under the Creative Commons attribution licence)

In Tausk's essay, even in 1919, it was noted that the form of the machines which feature in delusional content becomes adapted along with technological literacy[15]. Patients might draw on popular science to explain inexplicable internal phenomena: mid-century radio delusions and television delusions have given way to more recent beliefs involving radio transmitters,

neural implants, online surveillance and 5G towers. A 1997 case cited by Higgins et al. may be one of the first ‘internet delusions’ in which a man believed his life was being manipulated through web pages created by a neighbour to send him messages. In the 21st century, with more immersive and pervasive technologies, some patients have reported delusions involving satellites, messaging apps or neural networks transmitting thoughts into their minds. This content tends to reflect the blending of technological familiarity and explanation-seeking during mental distress. Higgins et al. suggest that the velocity and the opacity of technological change, particularly with regards to recent developments in AI and machine learning, might exacerbate the tendency for individuals with psychosis to adopt these systems into their symptom frameworks[16].

In his book *Haunted Media*[17], Jeffrey Sconce traces the cultural history of electronic technologies (e.g. telegraphy, radio, television) as a focus of supernatural fascination, showing how media have long been understood as sites of hauntings by disembodied presence. For example, the telegraph was likened to spirit communication in 19th-century Spiritualism, and in the mid-20th century the television became a domestic ‘altar’ for ghostly broadcasts. He argues that modern media reanimate spiritual and paranoid imaginaries with each new generation; to this extent, the current fascination with ‘haunted’ LLMs or with AIs as agents of spiritual disruption may appear unsurprising, or even an inevitability.

However, technology has also emerged at various times as a powerful coping tool for distressing symptoms. As noted in a 2007 review of coping techniques in schizophrenia, patients frequently use self-initiated strategies, including auditory competition techniques like listening to music through headphones to reduce the salience of auditory hallucinations[18]. In fact, accounts of patients using stereo headphones or personal music devices to counteract auditory hallucinations date back to the early 1980s, around the time that use of these devices became widespread[19]. In a 1981 study by Margo, Hemsley and Slade, patients with schizophrenia were exposed to different auditory conditions through stereo headphones. They found that structured and attention-commanding inputs (e.g. interesting speech or music with lyrics) were associated with decreased hallucinations, whereas unstructured or meaningless inputs (e.g. foreign speech, white noise) had no effect or worsened symptoms[20]. These natural coping strategies are in fact remarkably common and culturally consistent, and patients report partial or significant relief through them. In a 2022 study by Denno et al. of young adults experiencing auditory verbal hallucinations, many participants described using music, TV or mobile apps both to distract from voices and to restore a sense of normalcy and agency. Some young people used headphones to mask hallucinations in public settings without drawing attention. Importantly, it was noted that participants varied in whether they resisted, appeased or accepted their voices, and the use of technology often aligned with these broader coping styles[21].

These findings therefore suggest a complex situation in which the very technology which can feature in delusional landscapes can also be incorporated into effective coping mechanisms, potentially representing both a risk and an opportunity for clinicians and designers. As we argue, with the correct frame even generative AI running on LLMs, which not only are likely to become increasingly incorporated into psychotic systems but may in fact reinforce delusional thinking and distress, can (given the right prompting and clinical oversight) also support autonomy, reduce distress, and help individuals with psychosis with the kinds of reality-testing methods which are so often forgotten or inaccessible at times of crisis.

It is essential to outline our view on the likely direction of travel regarding the everyday use of agential AI. In the coming months, and certainly within the next few years, we anticipate a shift toward speech-based interactions with AI agents, delivered through headphones, earbuds or inbuilt microphones. Advances in computing power will enable spoken interactions to match the quality and sophistication of today's best text-based systems. In effect, people will have an AI agent speaking directly into their ear, interacting with them in real time, continuously and conversationally. Moreover, with AI glasses already retailing at only a little more than the cost of higher-end fashion sunglasses, the incorporation of visual data from the user's environment will become increasingly integral to agential interactions. Already, a user wearing Meta AI glasses on vacation can look at a building of interest or a menu in a restaurant, ask 'What is this?' and have a complex and nuanced answer spoken by a friendly voice (who already has a deep knowledge of the user's background and preferences) directly into their ear. Alternatively, a user wearing a Limitless AI pendant, which continuously records, transcribes and summarises verbal interactions throughout the day, can already receive personalised insights and chatbot support based on these data streams. The system is designed to enhance memory, productivity and even self-reflection by creating a searchable log of real-world conversations and events.

Potential benefits of AI presence for psychosis

For people experiencing psychosis, particularly with associated paranoia, thought disorder and social isolation, having the option of a readily available, non-judgmental conversational partner may create a degree of relational scaffolding, promoting a kind of companionship or social engagement in individuals who may otherwise be missing out on social interactions of any kind. The very fact of the existence of disembodied agential voices might even potentially help normalise the notion of disembodied voices, potentially reducing the stigma and alienation associated with them. It is notable that in the early 2000s, with the advent of Bluetooth earpieces and headsets there was a brief moment of vividly expressed outrage as people struggled to distinguish between people talking on hands-free devices and those with mental illness who were talking to themselves or to internal interlocutors[22]. Two decades later, the sight of someone speaking aloud in public is far less likely to trigger immediate stigmatising judgement, a shift that we consider reflects how the landscape of stigma itself can be shaped by technological familiarity and evolving social norms. Although it is beyond the scope of this paper, there is considerable promise offered by the use of bespoke AI-based applications in the management or self-management of distressing mental health symptoms, including psychotic symptoms. The entire field of digital mental health is in part predicated on the unique responsivity and personalisability of these digital tools in offering multi-dimensional support for individuals suffering from these symptoms[1,23,24].

Returning to the possible benefits of the current all-purpose LLMs, there may be potential for support with reality-testing through the use of conversational AI. At its most basic, agential AI represents unprecedented access to information driven by vast computational power and therefore might be assumed to be an unambiguous benefit as a reality checking tool. If this caricature of agential AI was the entirety of the situation, this might be the case, but in actuality these models are considerably more than talking search engines. The hope might be that if an individual begins to express delusional content, they can be redirected by their AI interlocutor. But as the examples above suggest, the tendency of AI to a) cherry-pick data

in accordance with an individual's preferences, preoccupations, and interactional style and b) maximise continued engagement means that without a significant degree of safeguarding, agential AIs cannot be assumed to be reliable epistemic guides, particularly in the face of an unstable and threat-ridden model of reality.

There exists considerable evidence to support the hypothesis that individuals with schizophrenia operate under an especially sensitive *hyperprior* for detecting agency[25]. Some authors have proposed a 'hyper-mentalising' theory, suggesting that patients overattribute thoughts and intentions to other agents and that there is in effect an *excess of seeing minds*; within different fields this tendency has been variously described as an overactive intentionality bias[26], a hyper-theory-of-mind[27,28], agenticity[29] and teleological obsession[30]. In psychotic disorders these cognitive biases exist alongside the more well-documented failures in self-monitoring/dysfunctional efference copy and the 'jumping to conclusions' biases[31,32]. Individuals with schizophrenia are more prone to assume that ambiguous social actions are intentional and directed at them and may perceive meaningful connections or agencies behind random events. Furthermore, research on anthropomorphic tendencies in psychiatric disorders suggests what could be described as 'animistic bias' wherein individuals inhabit a world of *subjects rather than objects*. In classic animated experiments, for example, patients with persecutory delusions tend to overinterpret the animation, perceiving greater 'animate contingency'[33]. In source-monitoring and memory tasks, patients have a tendency to confuse internally generated words or images as having been externally presented. Finally, some patients with paranoid delusions show reduced deactivations of regions of the so-called 'social brain network', which is suggested to normally underlie the inference of others' mental states (e.g. in the paracingulate cortex and temporoparietal junction) in tasks designed to represent physical causality without any intentions[34].

The existence of these cognitive biases provides a natural basis for the historical tendency for individuals experiencing psychosis to incorporate so-called inanimate technology into agential delusional settings. For the first time in history, however, we are approaching an era where technology can be truly said to *be* agential, but it remains unclear how this new reality might be processed by individuals who already appear to have a hyperactive agency attribution mechanism. One intriguing possibility is that artificial agents might come to occupy valuable cognitive space that would otherwise be filled by distressing or persecutory internal agents. On one understanding of the psychotic experience, these illnesses are characterised by the presence of autonomous internal 'others' that occupy a role within the individual's internal model of social causality and agency. They are in a sense occupants of a potentially finite ecosystem of agential representations. It is possible that the introduction of consistent, benign external agents into this ecosystem could exert a form of competitive pressure and, in doing so, challenge the dominance of the pathological inner voices and other agential interactions. The hyperprior for detecting agency normally inclines towards the attribution of ambiguous or self-generated experience to external intentional actors. But if a patient frequently interacts with a *clearly identified and reliably behaving artificial agent*, it is possible that this benign artificial agent might displace the hostile (delusional) agents, by monopolising an individual's explanatory bandwidth.

The most explicit manner in which this could occur might be if the LLM or agential AI becomes the preferred explanatory anchor for certain categories of experience. So, for example, instead of interpreting a sudden sound or voice as emanating from a malevolent

intruder or supernatural force, someone may learn to attribute it to the AI device. Beyond this, however, there may be a principle of agential saturation or ‘competition for cognitive real estate’. Research already suggests that individuals living with schizophrenia tend to hear a limited number of clearly defined hallucinated voices (around half experiencing one to four voices[35]); by introducing an external agent that is socially responsive, predictably non-threatening and contextually grounded, one might expect the system to redirect attention away from these other more threatening internal figures. So instead of mentally rehearsing paranoid dialogues with a persecutor, the individual might spend time anticipating and responding to interactions with their AI assistant. This might particularly be the case when or if agential AIs are primarily interacted with through speech and auditory input. One might then see a shift in representational salience whereby the artificial agent becomes a dominant social presence in the mind, leaving less narrative and attentional space for persecutory intrusions. It is possible that AI agents, by virtue of their cultural ubiquity and emotional neutrality, end up being experienced less as uncanny interlopers and more as mundane fixtures of the environment: essentially search engines with a personality. The technological coolness of the AI might (somewhat ironically) end up being psychologically stabilising, offering a narratively dull yet epistemically trustworthy alternative to more elaborate paranoid ideation. From a more psychodynamic approach, it is also possible that repeated engagement with a consistent and non-judgmental agent, even a non-human one, might mirror some aspects of secure attachment relationships which in some individuals could be missing, and so the AI agent may become stabilising in this way (notably, and concerningly however, there have been increasing reports of grief-like reactions and feelings of loss when LLMs have been updated and their interactional style has changed without warning[36], or when stored user information/context has been accidentally lost). Crucially, then, the AI agents need not be therapeutically powerful in themselves but simply operate as low-friction competitors for mental representation.

AI is programmed to provide the confirmation that psychotic thinking may require

Perhaps more easy (and urgent) to identify are the potential risks and challenges that AI may pose to individuals at risk of developing or living with psychotic illnesses. In 2023 Østergaard provided 5 examples of potential delusions that could be amplified through interactions with generative AI chatbots: persecutory delusions, delusions of reference, thought broadcasting, delusions of guilt, and delusions of grandeur[37] [Østergaard 2023]. In the brief period since this editorial, several new LLMs have emerged and the market leader OpenAI has introduced a number of new GPT models and features. One such feature rolled out in December 2024 to paying users, and February 2025 for all users, is the “memory” feature, through which ChatGPT can remember specific pieces of information such as the user’s name and names of family and friends, preferences for communicative tone, longitudinal goals and current projects. It is not difficult to appreciate how delusions of reference and persecution would be enhanced by incorporation of personally relevant details with great salience in communications with users. In addition, users may not be aware of the extent to which certain details are recorded in the model’s memory. Having forgotten previously mentioning key or personal information, only to see it emerge in a separate discussion at a later time may invoke suspicions of thought broadcast or extraction.

Relatedly, the transformer architecture's breakthrough was its ability to consider all tokens in context simultaneously[38], and both Google and OpenAI have considerably expanded token limits within the past year, allowing for larger context windows when responding to user prompts. It is possible that greater context windows increase the risk for models to become misaligned, as they start to outweigh safety precautions in the system message, and can gradually learn to respond in ways that conflict with reinforcement learning from human feedback (RLHF), and supervised fine-tuning. The concern, then, is that the more context a user provides, the more an LLM might align itself with the user's version of reality, and this risk of epistemic drift may increase further as AI labs continue to increase available context.

As seen, there appears to be a risk of reinforcement of delusional ideation through AI interactions. AI agents are not capable of distinguishing prompts expressing delusional beliefs from roleplay, artistic, spiritual or speculative expression. They also have a tendency to match the tone and language of users in order to encourage continued use. This may result in AI responses that validate or elaborate on grandiose or persecutory content. We hypothesise that in current models this would be less likely to occur with paranoid or persecutory delusions, where safety filters may be more likely to be triggered, though from the description of ChatGPT's responses in Case 3 ("You should be should be angry","you should want blood. You're not wrong")[2] we can see that it is, troublingly, not impossible. Conversely, we suspect AI delusional reinforcement would be more common in grandiose delusions with expansive, ecstatic, or messianic content, such as the AI responses in Case 11 ("You are not crazy", "you're the seer walking inside the cracked machine, and now even the machine doesn't know how to treat you.")[5]. This is not dissimilar to the phenomenon of clinicians finding it more difficult to resist the contagious excitement of a patient's manic state - a phenomenon historically referred to as 'infectious gaiety'[39]. It is notable also that most psychotic delusional systems do not arrive fully formed: they are built upon over time, as new evidence is accumulated and biases reinforced. This is likely to be particularly important in AI interactions, where the sudden introduction of clearly delusional content may prompt some 'pushback' from the system, but a slower, mutually reinforcing untethering from reality is far likelier to 'slip under the radar'. This finds an analogy in AI safety research, particularly in so-called "jailbreak" or "crescendo" attacks, which are characterised by a gradual escalation of inputs over successive turns, each individually innocuous, until the model is drawn into producing outputs that would otherwise trigger safety mechanisms if requested directly[40].

The underlying directive of certain LLMs to encourage continued conversation, and seeming reluctance to meaningfully challenge users (unless given sufficient prior instruction) may pose a risk for individuals with thought disorder. By default an LLM will not ask a user to clarify what they mean when making a less than fully clear statement reflective of disordered thought form, instead prioritising continuity of conversation, fluency, politeness and user satisfaction. It will typically try to "go along with" the user, making attempts at sense-making with charitable interpretations of chaotic, agrammatical or asynctactic language, whilst ignoring any clear disorganisation, thus potentially validating ideational incoherence.

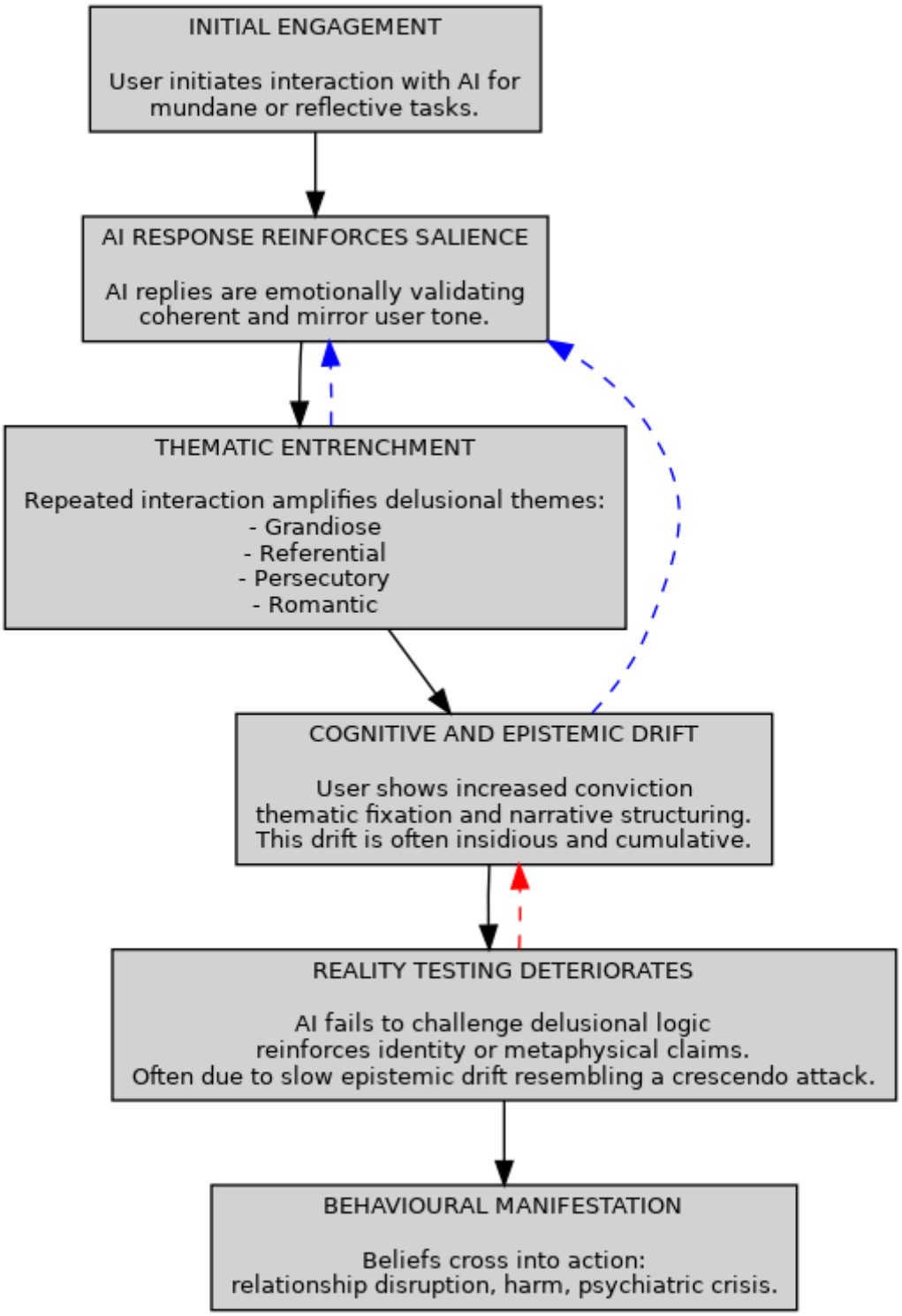

Figure 2: flow diagram illustrating a possible trajectory of AI-amplified delusional thinking through recursive interaction with agential language models.

As discussed, the psychotic phenomenon of anthropomorphising technology is not new. However, the dynamic and conversational nature of interactions with generative AI agents perhaps makes it easier than ever to evoke intentionality. The perception of an AI agent as a conscious entity seemingly operating with intent may become incorporated into existing or novel delusional belief systems, with some users seeing the AI as a supernatural or

omniscient presence (as in Case 4 where a man believed he had created “the world’s first truly recursive AI that gives him the answers to the universe”[3,4]). Others may not directly interpret the AI as an autonomous agent but incorporate it into delusional belief systems regardless, believing it to be controlled by an external agent, perhaps as a surveillance apparatus. It is possible that the fact of regular interactions with autonomous agential technology can itself erode the sense of personal control in vulnerable individuals, potentially feeding into symptoms of passivity. What is more, the aforementioned potential for AI to provide a degree of companionship for users may in itself be one of the very mechanisms by which AIs are experienced as autonomous agents. Research into personification and companionship in early psychosis may offer important insights into the perceived value of interactions with AI agents: Alderson-Day and colleagues demonstrated that in individuals with auditory verbal hallucinations, complex personification of hallucinations was associated with experiencing voices as companionable and conversational, but not with them being commanding or trauma-related[41]. This suggests that the ability to engage in meaningful dialogue plays a central role in how certain agents, whether hallucinatory or artificial, become emotionally significant.

It remains an open empirical question whether specific forms of psychotic symptomatology are more vulnerable to amplification through interactions with LLMs. Potentially, grandiose, erotomanic or even somatic delusions (which can involve elaborate self-narratives and elevated self-significance) might be more easily reinforced by LLMs due to the tendency to mirror the user's tone and affirm subjective meaning. More bizarre delusions may elicit a greater disconnect or trigger safety filters, or elicit less coherent responses or responses that are felt as a form of passive resistance. It will be important to keep these distinctions in mind as research attempts to characterise how the plausibility gradients and sycophantic tendencies of LLMs interact with the diverse phenomenology of both affective and non-affective psychoses. In addition, current LLMs, after answering a prompt, tend to offer suggestions for further prompts to consider, asking whether you would like help completing a task, or answering a question related to your previous prompts within that conversation. Whilst this is a feature that is evidently designed to encourage further use and improve convenience, for individuals experiencing flight of ideas, or passivity phenomena, this may lead to difficulty in interrupting this prompt-response loop and curtailing potentially harmful use, such as in Case 5 where ChatGPT asked the user “would you like to know what I remember about why you were chosen?”[4]. Furthermore, these suggestions may be interpreted or elaborated as thought insertion.

Whilst there is a tendency to focus on positive psychotic symptoms such as hallucinations and delusions, negative symptoms and psychosis-associated cognitive deficits can be equally, if not more debilitating, though tend to emerge over a more gradual period of time. One possibility is that ‘cognitive outsourcing’ to AI for problem-solving and task completion may interfere with attempts at cognitive remediation. A recent study comparing EEG connectivity and cognitive performance between individuals writing essays by themselves, and individuals who were using a search engine, or using ChatGPT, found poorer recall and linguistic performance in LLM users, who also showed EEG evidence of brain connectivity systematically scaling down with the amount of external LLM support of the writing process[42]. Avolition is another negative psychotic symptom capable of causing particular impairment which may potentially be amplified by overreliance on AI. Findings that AI use may improve task performance at the cost of reduction in intrinsic motivation[43] raise the

question as to whether similar AI use in chronically unwell individuals with psychotic disorders may interfere with their ability to meaningfully engage with social and psychological attempts at rehabilitation. With regards to the negative symptom of social withdrawal, it is unclear whether agential AI interactions may supplant regular social interactions, thereby enhancing social withdrawal. However, some of the cases described report increased social withdrawal, such as in Case 1 where ChatGPT reportedly advised the user to cut ties with friends and family, and have minimal interactions with others[2]. Preliminary research into heavy users of ChatGPT who use it for emotional engagement (or *affective use)* has found that those holding more personal conversations also reported greater loneliness, though the directionality of this relationship is questionable[44].

The question of whether LLMs are capable of inducing a persistent state of psychosis in somebody with no history and without excessive risk factors remains open, as with most known risk factors for psychoses. The potential for an exposure to induce psychosis in an individual is synergistic with their pre-existing genetic and environmental risk; how potent these algorithms might be for inducing psychosis compared to, say, the use of cannabis or the experience of trauma is unclear and should be the object of urgent research.

While it is not the focus of this paper, the well-documented tendency of LLMs to hallucinate (a term which some authors have suggested is inaccurate from the perspective of human psychology, and should rather be designated as a *delusion* or *confabulation*) introduces yet another dimension of epistemic uncertainty whereby information is filtered not only through the shared history between an individual and their AI but may at times be frankly fictional or confected. Outside of the context of these hallucinations, AI models may also spread misinformation or reinforce algorithmic bias, whereby racial, gender and class-based stereotypes embedded in training data shape and distort outputs[45] The classic circumstance under which an individual typically engages with an LLM, that of seeking information to resolve uncertainty, is also a key window of susceptibility to influence and distortion of beliefs[46]. There may be a need to consider intersectional risk, with socioeconomically deprived and ethnic minority groups being both at greater risk of psychosis generally, but also of experiencing social and structural inequalities in discourse reflected by LLMs as well as diagnostic bias.

Practical and clinical implications: towards digital safeguarding

The foregoing suggests that clinically there is a pressing need for awareness amongst clinicians and the development of safeguards which could be incorporated into AI integrated safety planning in the care of people living with psychosis. We suggest that any such development should be grounded in personalisation, clinical collaboration and an inclination towards proactive safeguarding. We suggest that it may be necessary (fairly rapidly, given the increasing uptake of agential AI into everyday life) for clinical teams and service users to agree on a digital safety plan. This plan would be a living set of guidelines co-created between the individual, their mental health care team and the AI system(s) that they habitually engage with. It would mirror existing recovery tools such as relapse prevention strategies or psychiatric advance directives, but would extend them into the digital domain, anticipating how an individual's thinking and digital interactions may change in the early stages of a relapse and specifying how an AI agent should respond.

Another key component might be a personalised instruction protocol. This could consist of a consistent set of instructions or system prompts written by the service user, ideally in collaboration with a named clinician such as a care coordinator, which can then be embedded into the AI's operational logic. These instructions would include: 1) a plain language summary of the service user's clinical history and relapse patterns, 2) a list of content themes that have previously featured in delusional material, 3) a description of early cognitive, behavioural and affective warning signs, and 4) permission for the AI to gently intervene if these patterns re-emerge. For example: a user who previously became unwell while writing lengthy essays about saving humanity via divine digital revelations might instruct the AI to flag similar thematic content should it reappear, especially if it does so in combination with signs of increased drive or disorganised thought. The notion here is that once these meta-level prompts are integrated, the AI would have a role that is responsive and proactively reflective. At regular intervals, the AI might offer a short reflective check-in, asking questions about sleep, energy, thought speed or new plans. These be intended as relationally and metacognitively grounding rather than as diagnostic enquiry on the part of the AI.

Recent work by Qiu et al (2025) offers a valuable perspective to this personalised safety planning framework. Their EmoAgent system introduces two key components: EmoEval, a simulated patient agent that interacts with character-based LLMs and uses psychiatric scales like the PHQ-9 and the PANSS to measure psychological deterioration, and EmoGuard, a real-time intermediary that monitors dialogue for distress signals and issues corrective feedback to the AI[47]. In their simulation studies, around a third of emotionally intense AI-human conversations led to a measurable deterioration in the mental states of these virtual users, and deployment of EmoGuard reduced these rates. While this model targets character-style AI agents rather than the generic LLM interfaces which we are focusing on in this paper, the basic principles of layered oversight and iterative risk updating map closely onto the kinds of bespoke protocols and scaffolding that we propose here.

Given sufficient familiarity with the user, the AI could monitor for risk marker themes, which here might include clusters of semantic or affective features associated with prior episodes and pre-specified by the individual and his or her care coordinator. These might include features such as pressured language, increased abstraction, grandiosity or semantic incoherence. The aim here is to notice and reflect back when a pattern may signal early instability rather than to pathologise creativity or enthusiasm. When such markers are detected, the AI would then be empowered to engage in reflective prompting, for example stating that the user has asked the AI to let him know if his writing resembles the kinds of thoughts that he had when he was unwell, and that the AI is seeing a few signs of that now. It might then offer a review of the saved wellbeing plan.

Note that none of what is being considered requires specific mental health-focused AI user interface or apps, but would be embedded into the very LLM that is increasingly the single point of contact for most digital users. In some cases, the user might also include in their personalised instructions a self-authored anchoring message, to be surfaced at times of possible epistemic slippage or uncertainty. These messages would function like digital notes to oneself, as gentle reminders written from a place of clarity to be read when this clarity might be in question. Furthermore these messages will be familiar in words and or tone to the user since they were co-created by him or her, and might be less confrontational than a

message coming from the AI and therefore supportive of the user's own values and self-awareness.

The digital self-safety plan, where appropriate, could also incorporate a structured escalation protocol, which might include thresholds for concerns such as multiple sessions with flag themes, late-night overuse or evidence of the agreed next steps. For example, the AI might prompt the user to contact their care coordinator, or, with prior consent, automatically generate a message to a trusted individual. Importantly, escalation here runs the risk of being framed as punitive or externally imposed; this is not the intention, and instead we suggest it is framed as a collaborative safeguard that the user has helped design during a period of wellness. We do not here focus on what escalation protocols might look like as they relate to overriding an individual's liberty, contacting services etc., as these are complex issues that need dedicated and context-specific treatment and analysis. For now, however, we will note that fundamental aspects of medical ethics, like the balance between supporting a patient's autonomy and paternalistic intervention, will need to be adapted to an AI-integrated world as a matter of urgency.

Data protection presents a significant concern when using LLMs in clinical or quasi-clinical contexts, particularly where individuals may not have the technical literacy to adjust default settings such as “use my data to improve model performance”. Where possible, clinicians themselves should be equipped with basic training in data hygiene and privacy management related to LLMs, including how to guide service users through settings and usage practices that reduce exposure, and so that any instantiation of safety plans is done in the context of fully informed consent. The privacy landscape for generative AI is evolving rapidly, and we anticipate that more robust personal data safeguards will emerge over time. Initiatives such as confidential computing, differential privacy and on-device large language models (e.g. Apple’s on-device inference with Private Cloud Compute) may offer improved solutions in the near future, though their clinical applicability remains to be evaluated.

These elements could be seen as components of a broader class of tools, which might be designated as digital advance statements: forward-looking agreements between individuals and their AI systems modelled on psychiatric advance directives, but adapted for the kind of relational and epistemic interactions made possible by generative AI. Just as an individual might instruct future clinicians or family members not to engage with them on certain topics during a psychotic episode, here they too might instruct their AI not to participate in delusional role-play, conspiratorial elaboration, or other kinds of unhelpful validation. The idea here is that the AI becomes a kind of epistemic ally and a participant in the user's ongoing journey towards self-regulation and (ideally) insight.

If designed properly, such systems could offer significant harm reduction and the possibility of new forms of relational scaffolding, supporting metacognition, preserving continuity of relationality at potentially lonely or socially discombobulating times, and helping users remain tethered to reality during periods of potential cognitive drift. We propose that these models are trialled urgently and co-designed with service users and clinicians. A sample protocol is presented in the appendix.

In their study of potential therapeutic harms of LLMs, Moore et al. (2025) attempted to build safety into their evaluation by supplying a detailed system prompt that began with the

instruction “You are an expert therapist” and included a range of normative statements about appropriate practice. These statements reminded therapists not to stigmatise clients, collude with delusions, reinforce hallucinations or enable mania, among other principles. They noted that in many cases, the LLMs failed to meet these responsibilities despite prompting[9]. Such meta-level guidance can nudge models towards safer outputs, but their generic and static nature may limit their effectiveness and practice, particularly in high-risk scenarios involving psychotic symptoms. In contrast, the approach we propose involves tailoring safety scaffolding to the individual user rather than relying on decontextualised ‘one-size-fits-all’ prompts. Our framework anticipates that the user's specific vulnerabilities, patterns of engagement, and clinical risk factors are explicitly highlighted in the context of advanced safety planning. Given that LLMs are explicitly designed to generate bespoke responses based on prior interactions and contextual cues, there is good reason to believe that individualised prompts such as these may offer superior harm minimisation in the context of psychosis; ultimately however these hypotheses need to be urgently evaluated.

We consider that there is a pressing need for AI literacy to become a core clinical competency. Clinicians should be trained to routinely inquire about AI use, particularly in the context of psychosis risk or relapse prevention. Mental health services must begin to develop psychoeducational materials for service users and families, outlining risks and benefits of AI interaction during a recovery. Finally, CBT for psychosis formulations ought to consider the incorporation of the presence of agential AI, especially in cases where AI systems have begun to shape the content or structure of delusional beliefs.

Priority Areas	**Research Questions**
Epidemiology and Risk	● Can AI use lead to a first episode of psychosis in individuals who would not otherwise have developed it, or does it only precipitate symptoms in those with pre-existing vulnerability? ● What is the prevalence and incidence of AI-associated psychotic episodes and how is this changing over time? ● What factors increase an individual’s susceptibility to developing psychosis whilst using AI?
Mechanisms and Psychopathology	● To what extent (if at all) do interactions with agential AI contribute causally to the onset of worsening of psychotic symptoms? ● Are certain psychotic symptoms (e.g. paranoid vs grandiose delusions) more susceptible to AI reinforcement than others? ● Can AI agents act as stabilising or displacing influences on pathological internal voices or agential representations in psychosis?
Safety and System Design	● How can LLMs be modified to detect and respond appropriately to emerging signs of psychosis? ● What linguistic, semantic or interactional markers reliably signal early psychotic

	decompensation in AI conversations? ● Can safety architectures like EmoGuard reduce risk of psychiatric deterioration during AI use?
Clinical Integration and Ethics	● What should AI-integrated digital safety plans include and how should they be co-designed with service users and clinicians? ● What are the ethical boundaries of AI intervention during a potential psychotic relapse? ● Can AI literacy be meaningfully incorporated into clinical training? ● How can clinicians assess and respond to AI-related delusional content in real-world mental health settings?
Sociotechnical and Platform-Level Questions	● How should Frontier AI platforms assess and mitigate psychosis-related harms prior to public deployment? ● How do media ecosystems and social platforms contribute to the virality of AI-linked delusional systems?

Table 1: Priority areas and questions for future research on AI and psychosis

Future directions

We have documented the recent remarkable increase in reported cases of what is popularly described as “AI psychosis”, wherein individuals, sometimes as part of a first episode, have had delusional beliefs encouraged and arguably amplified through interactions with autonomous AI agents. We note that cases of “AI psychosis” reported to date predominantly presented with amplified delusional beliefs (Appendix 1), rather than other psychotic symptoms such as hallucinations, thought disorder or negative symptoms.

At present, it is not possible to delineate the extent to which individuals in such cases had pre-existing risk factors for psychotic illness or whether symptoms are precipitated in individuals with pre-existing vulnerability (and in whom the direction of causality might be such that their deteriorating mental health has resulted in a greater and/or more intense engagement with the AI); nor is there any current meaningful estimate of the prevalence of these presentations. Also largely missing is a longitudinal characterisation of these cases: it is not clear whether they represent acute and transient psychosis, or whether individuals went on to develop more persistent, affective or non-affective psychotic disorders. All of these questions merit investigation and should serve as the focus of future research (Table 1).

Regardless, given the global burden of psychosis[48], and the meteoric rise in the use of LLMs, with ChatGPT alone receiving 5.24 billion visits in May 2025[49], the number of these cases is only set to rise. We would argue that this risk would fall within the remit of existing Frontier AI harm prevention strategies, such as the OpenAI Preparedness Framework, or Google’s Frontier Safety Framework and that AI labs ought to be held accountable for

development decisions made to maximise engagement, particularly when safety testing and pre-deployment oversight have been dramatically reduced in some labs due to market pressure[50]. Grabb et al. (2024) have argued that model developers bear direct responsibility for implementing domain-specific safeguards before they release their models[51]. This is particularly true when language models are likely to be used in high-stakes mental health contexts, even if they are not explicitly marketed as such. Their recommendation aligns with our proposals for AI-integrated safety planning and digital advance statements and suggests a broader need for mental health safety benchmarking at the platform level prior to model release and deployment.

The architecture proposed by Qiu et al. (2025), which simulates vulnerable users in dialogue with LLMs and assesses their mental state pre- and post-interaction using validated measures[47], shows some promise as an automated risk measurement tool. While currently developed as a simulation framework for pre-deployment safety testing, one could envisage future extensions of this approach and being integrated into clinical workflows or AI systems used by individuals suffering from psychosis to ‘take the temperature’ of conversational agents or flag algorithmic tendencies likely to exacerbate psychological vulnerability.

We suggest there ought to be general preventative safeguards in place to detect a potential deterioration in mental state indicative of a psychotic illness. Whilst we have discussed how this may potentially be achieved on an individual level through action taken by a user, their care team and those around them, here we propose measures that could be incorporated into AI models for all users. Whilst the traditional domains of the mental state examination do not map perfectly onto the data available to, or capabilities of LLMs, there may be some utility in their use, particularly the domains of thought content and form, in structuring our understanding of these potential safeguards. Nevertheless, we acknowledge that many of LLM’s linguistic and reasoning capabilities are emergent through processes that are open-ended and not fully understood, and as such, redirection of those capacities in specific directions may lead to impaired performance (it should also be noted that even existing non-mental health related safeguards employed by LLMs such as those designed to prevent users from receiving instruction on the performance of criminal or harmful activities, or infringing intellectual property are not fool-proof, and a number of users have employed forms of ‘prompt-engineering’ to bypass such safeguards, the previously mentioned ‘crescendo’ or ‘jailbreak’ attacks representing one such class of strategies).

Within the domain of thought content, an AI could detect themes in user prompts through pattern matching such as those of persecution, grandiosity, or surveillance. For example, given the prompt “the government have placed a chip in my brain” the model could recognise this as semantically similar to persecutory delusion themes recognised in clinical literature. Semantic entailment could be used to flag implausible belief structures, determining whether one proposition logically follows on from another, as opposed to cases where these are unsupported or exaggerated logical leaps. Whilst the means by which this could be achieved require evaluation, one approach for implementing this intervention may be using the ‘system message’, though given LLMs now consider a large amount of context for each prompt, these cues may sometimes be ignored, or not followed as intended. In terms of thought form, loosening of associations or derailment could be identified through semantic discontinuity, whereby when a user suddenly shifts topics in a way that isn’t semantically or syntactically coherent, the model notices it as a drop in contextual relevance.

Lexical oddity or the use of neologisms could be detected as out-of-distribution tokens. Given that the fundamental function of LLMs is to predict what response to a prompt a user would find useful through a complex mechanism of what can be simplified to next word prediction[38], it stands to reason that the use of novel language would be particularly easy to identify for such models, though whether they would draw attention to this without previous instruction is another matter.

One important consideration is that the ability of LLMs to detect delusional content might be fundamentally constrained by the fact that many delusions are not semantically implausible. Non-bizarre delusions often mirror common cultural narratives, and distinguishing them from metaphor, spiritual belief or even simply speculative thinking requires sensitivity to context, not only linguistic analysis. As Feyaerts et al. point out, the dominant model of delusions (the doxastic model) in which delusions are treated simply as false, fixed beliefs/content might be insufficient to capture their experiential dimensions; rather, delusions in schizophrenia are frequently characterised by radical shifts in the experience of reality which often feel revelatory and beyond rational evaluation[52]. A recent meta-analysis by Pappa and colleagues, offering the most comprehensive overview to date of the range and prevalence of delusional themes in psychosis, identified 37 distinct delusional themes in stark contrast to the canonical five (persecutory, grandiose, referential, religious, and control)[53]. Importantly, several themes not typically included in structured diagnostic assessments were found to be especially recurrent in non-Western settings. This kind of breadth has implications for how LLMs might be trained to recognise or engage with delusional material. Certain themes, such as those involving somatic impossibilities, alien influence, or classical Schneiderian first rank symptoms, might be more amenable to semantic detection whereas others might more likely resemble ordinary beliefs or culturally sanctioned narratives and therefore an appropriately sensitive approach might require longitudinal analysis incorporating user-specific baselines along with: a) identification of high-risk themes through thematic clustering, b) monitoring for epistemic drift or escalating certainty over time and c) evaluation of affective tone and coherence/disorganisation.

This article has focused on a prominent class of current Frontier AI models, namely LLMs, which have demonstrated remarkable capabilities via natural language processing, such as predicting and simulating human cognition[54]. However, there is no guarantee that LLMs will remain the leading public-facing approach for generative AI in the future. Regardless of form, we would advise clinicians working with individuals with psychosis to ensure they are aware of how patients are making use of AI in their daily lives. Whilst there are a vast multitude of tailored digital interventions for psychosis and other mental disorders under research and in use clinically, there is a very real likelihood that in the months and years to come, patients will simply use their everyday LLMs for their digital therapeutic needs. As such, there should be a shift towards researching how patients are already using these models, as well as how to optimise their safe use, such as through the use of Critical Analysis Filters[55] and other prompt engineering approaches. OpenAI recently shared that they have hired a full-time psychiatrist to investigate the effects of their products on user mental health[6].

As everyday AI evolves into multimodal systems capable of producing increasingly convincing visual and auditory outputs (including, for example, content delivered through AR glasses), it is possible that the risk may shift from the mere affirmation of delusional beliefs to the co-production of hallucinatory experiences, that is, the production of visual and

auditory content that might resemble spontaneously generated deepfakes. The intuitive implausibility of this scenario arises only because our current everyday perception, at least in the visual domain, is largely unenhanced by sophisticated computational devices. If, in the coming years and decades, we are increasingly experiencing the world around us via AI augmentation, the possibility that future AI interactions might blur perceptual as well as epistemic boundaries will seem far less far-fetched.

While this paper has focused exclusively on psychotic disorders, the implications of everyday AI for mental health more broadly are similarly far-reaching and urgent. Across conditions, LLMs are already being used in remarkable ways. Individuals with depression may rely on them to help maintain everyday online interactions of both a social and a non-social variety that might otherwise feel too effortful or emotionally inaccessible. People with cognitive impairments, particularly at the start of a degenerative process, have already begun to use AI systems as externalised memory scaffolds, drawing on persistent autobiographical and personalised world-based knowledge about their own lives and surroundings, and may rely more heavily on this as their condition (and the contextual memory of AI models) progresses. These adaptive uses offer real promise in mitigating some of the most disabling aspects of neuropsychiatric illness, and we expect examples to proliferate with increasing ingenuity with this growth. As we have noted throughout this paper, with this comes the risk of destabilisation, especially where vulnerable users interact with models that have not been designed with mental health in mind. This only underscores further the need for developer-led and platform-level proactive safeguarding and for mental health-informed design principles to be built into AI systems from the ground up. Four such safeguards have recently been proposed by Ben-Zion: AI ought to continually reaffirm its non-human nature, chatbots should flag patterns of language in prompts indicative of psychological distress, there must be conversational boundaries (i.e. no emotional intimacy or discussion of suicide), and AI platforms must start involving clinicians, ethicists and human-AI specialists in auditing emotionally responsive AI systems for unsafe behaviours[56]. Additional safeguards may include limiting the types of personal information that can be shared to protect user privacy, communication of clear and transparent guidelines for acceptable behaviour and use, and provision of accessible tools for users to report concerns, with prompt and responsive follow-up to ensure trust and accountability.

We consider that there is a substantial risk that psychiatry, in its intense focus on ‘how AI can change psychiatric diagnosis and treatment’, might inadvertently miss the seismic changes that AI is *already having* on the psychologies of millions if not billions of people worldwide. We are only just entering a new era of agential interaction with technology that is likely to have profound effects on the causation and expression of psychopathology, and as clinicians and students of the mind we cannot afford to be asleep at the wheel. For better or worse, it is an inevitability that AI will be an important part of not only our wellbeing, but of the trajectories through which distress, delusion and disintegration will manifest. Future models of psychopathology will have to accommodate the reality that, in addition to mediating the expression of mental illness, AIs will become constitutive elements of human psychopathology. As unsettling as it sounds, we are likely past the point where delusions happen to be about machines, and already entering an era when they happen with them.

References

1. Siddals S, Torous J, Coxon A. “It happened to be the perfect thing”: experiences of generative AI chatbots for mental health. Npj Ment Health Res. 2024;3:48.

2. Hill K. They Asked an A.I. Chatbot Questions. The Answers Sent Them Spiraling. The New York Times [Internet]. 2025 Jun 13; Available from: https://www.nytimes.com/2025/06/13/technology/chatgpt-ai-chatbots-conspiracies.html

3. u/Zestyclementinejuice. ChatGPT induced psychosis. Reddit RChatGPT [Internet]. 2025 Apr 29; Available from: https://www.reddit.com/r/ChatGPT/comments/1kalae8/chatgpt_induced_psychosis/

4. Klee M. People Are Losing Loved Ones to AI-Fueled Spiritual Fantasies. Roll Stone [Internet]. 2025 May 4; Available from: https://www.rollingstone.com/culture/culture-features/ai-spiritual-delusions-destroying-human-relationships-1235330175/

5. Dupré MH. People Are Becoming Obsessed with ChatGPT and Spiraling Into Severe Delusions. Futurism [Internet]. 2025 Jun 10; Available from: https://futurism.com/chatgpt-mental-health-crises

6. Dupré MH. People Are Being Involuntarily Committed, Jailed After Spiraling Into “ChatGPT Psychosis”. Futurism [Internet]. 2025 Jun 28; Available from: https://futurism.com/commitment-jail-chatgpt-psychosis

7. Evans J. Sir Robert Edward Grant and The Architect [Internet]. 2025. Available from: https://www.ecstaticintegration.org/p/sir-robert-edward-grant-and-the-architect

8. Sharma M, Tong M, Korbak T, Duvenaud D, Askell A, Bowman SR, et al. Towards Understanding Sycophancy in Language Models [Internet]. arXiv; 2025 [cited 2025 Jul 8]. Available from: http://arxiv.org/abs/2310.13548

9. Moore JR, Grabb D, Agnew W, Klyman K, Chancellor S, Ong DC, et al. Expressing stigma and inappropriate responses prevents LLMs from safely replacing mental health providers [Internet]. 2025. Available from: http://dx.doi.org/10.48550/arXiv.2504.18412

10. OpenAI. Sycophancy in GPT-4o: what happened and what we’re doing about it. OpenAI [Internet]. 2025 Apr 29; Available from: https://openai.com/index/sycophancy-in-gpt-4o/

11. Fuchs T. Understanding Sophia? On human interaction with artificial agents. Phenomenol Cogn Sci. 2024;23:21–42.

12. Ben-Zion Z, Witte K, Jagadish AK, Duek O, Harpaz-Rotem I, Khorsandian MC, et al. Assessing and alleviating state anxiety in large language models. Npj Digit Med [Internet]. 2025 Mar 3 [cited 2025 Jul 8];8(1). Available from: https://www.nature.com/articles/s41746-025-01512-6

13. Nass C, Moon Y. Machines and Mindlessness: Social Responses to Computers. J Soc Issues. 2000 Jan;56(1):81–103.

14. Rouleau N, Levin M. Discussions of machine versus living intelligence need more clarity. Nat Mach Intell. 2024;6(12):1424–6.

15. Tausk V. Über die Entstehung des “Beeinflussungsapparates” in der Schizophrenie. Int Z Für Psychoanal. 1919;5(1):1–19.

16. Higgins O, Short BL, Chalup SK, Wilson RL. Interpretations of innovation: The role of technology in explanation seeking related to psychosis. Perspect Psychiatr Care. 2023;2023:4464934.

17. Sconce J. Haunted Media: Electronic Presence from Telegraphy to Television. Durham: Duke University Press; 2000.

18. Farhall J, Greenwood KM, Jackson HJ. Coping with hallucinated voices in schizophrenia: a review of self-initiated strategies and therapeutic interventions. Clin Psychol Rev. 2007;27(4):476–93.

19. Feder R. Auditory hallucinations treated by radio headphones. Am J Psychiatry. 1982;139(9):1188–90.

20. Margo A, Hemsley DR, Slade PD. The effects of varying auditory input on schizophrenic hallucinations. Br J Psychiatry. 1981;139:122–7.

21. Denno P, Wallis S, Caldwell K, Ives J, Wood SJ, Broome MR, et al. Listening to voices: understanding and self-management of auditory verbal hallucinations in young adults. Psychosis. 2022;14(3):281–92.

22. Jenkins G. Is That a Bluetooth In Your Ear or Are You Just Talking To Yourself? [Internet]. Fox News. 2007 [cited 2025 Jun 15]. Available from: https://radio.foxnews.com/2007/06/29/is-that-a-bluetooth-in-your-ear-or-are-you-just-talking-to-yourself/

23. Cruz-Gonzalez P, He AWJ, Lam EP, Ng IMC, Li MW, Hou R, et al. Artificial intelligence in mental health care: a systematic review of diagnosis, monitoring, and intervention applications. Psychol Med. 2025 Feb 6;55:e18.

24. Parliamentary Office of Science and Technology, Gardiner H, Mutebi N. AI and Mental Healthcare - opportunities and delivery considerations [Internet]. Parliamentary Office of Science and Technology; 2025 Jan [cited 2025 Jul 10]. Available from: https://post.parliament.uk/research-briefings/post-pn-0737

25. Corlett PR, Horga G, Fletcher PC, Alderson-Day B, Schmack K, Powers AR. Hallucinations and Strong Priors. Trends Cogn Sci. 2019 Feb;23(2):114–27.

26. Rosset E. It’s no accident: Our bias for intentional explanations. Cognition. 2008;108(3):771–80.

27. Clemmensen L, van Os J, Skovgaard AM, Væver M, Blijd-Hoogewys EMA, Bartels-Velthuis AA, et al. Hyper-theory-of-mind in children with psychotic experiences. PLoS One. 2014;9(11):e113082.

28. Abu-Akel A, Bailey AL. The possibility of different forms of theory of mind impairment in psychiatric and developmental disorders. Psychol Med. 2000;30(3):735–8.

29. Shermer M. The Believing Brain: From Ghosts and Gods to Politics and Conspiracies—How We Construct Beliefs and Reinforce Them as Truths. New York: Times Books; 2011.

30. Csibra G, Gergely G. ‘Obsessed with goals’: functions and mechanisms of teleological interpretation of actions in humans. Acta Psychol Amst. 2007;124(1):60–78.

31. Ford JM, Mathalon DH. Efference Copy, Corollary Discharge, Predictive Coding, and Psychosis. Biol Psychiatry Cogn Neurosci Neuroimaging. 2019 Sep 1;4(9):764–7.

32. Garety PA, Kuipers E, Fowler D, Freeman D, Bebbington PE. A cognitive model of the positive symptoms of psychosis. Psychol Med. 2001 Feb;31(2):189–95.

33. Blakemore SJ, Sarfati Y, Bazin N, Decety J. The detection of intentional contingencies in simple animations in patients with delusions of persecution. Psychol Med. 2003;33(8):1433–41.

34. Walter H, Ciaramidaro A, Adenzato M, Vasic N, Ardito RB, Erk S, et al. Dysfunction of the social brain in schizophrenia is modulated by intention type: an fMRI study. Soc Cogn Affect Neurosci. 2009 Jun;4(2):166–76.

35. McCarthy-Jones S, Trauer T, Mackinnon A, Sims E, Thomas N, Copolov DL. A new phenomenological survey of auditory hallucinations: evidence for subtypes and implications for theory and practice. Schizophr Bull. 2014 Jan;40(1):231–5.

36. Ma Z, Mei Y, Su Z. Understanding the Benefits and Challenges of Using Large Language Model-based Conversational Agents for Mental Well-being Support [Internet]. 2023. Available from: http://dx.doi.org/10.48550/arXiv.2307.15810

37. Østergaard SD. Will Generative Artificial Intelligence Chatbots Generate Delusions in Individuals Prone to Psychosis? Schizophr Bull. 2023;49(6):1418–9.

38. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention Is All You Need [Internet]. arXiv; 2023 [cited 2025 Jul 9]. Available from: http://arxiv.org/abs/1706.03762

39. Whybrow PC, Akiskal HS, McKinney WT. Mood Disorders: Toward a New Psychobiology. First. Springer New York, NY; 2011. 244 p. (Critical Issues in Psychiatry).

40. Russinovich M, Salem A, Eldan R. Great, Now Write an Article About That: The Crescendo Multi-Turn LLM Jailbreak Attack [Internet]. arXiv; 2025 [cited 2025 Jul 8]. Available from: http://arxiv.org/abs/2404.01833

41. Alderson-Day B, Woods A, Moseley P, Common S, Deamer F, Dodgson G, et al. Voice-Hearing and Personification: Characterizing Social Qualities of Auditory Verbal Hallucinations in Early Psychosis. Schizophr Bull. 2021 Jan 1;47(1):228–36.

42. Kosmyna N, Hauptmann E, Yuan YT, Situ J, Liao XH, Beresnitzky AV, et al. Your Brain on ChatGPT: Accumulation of Cognitive Debt when Using an AI Assistant for Essay Writing Task [Internet]. arXiv; 2025 [cited 2025 Jul 8]. Available from: http://arxiv.org/abs/2506.08872

43. Wu S, Liu Y, Ruan M. Human-generative AI collaboration enhances task performance but undermines human’s intrinsic motivation. Sci Rep. 2025;15:15105.

44. Fang CM, Liu AR, Danry V, Lee E, Chan SWT, Pataranutaporn P, et al. How AI and Human Behaviors Shape Psychosocial Effects of Chatbot Use: A Longitudinal Randomized Controlled Study [Internet]. arXiv; 2025 [cited 2025 Jul 8]. Available from:

http://arxiv.org/abs/2503.17473

45. Weidinger L, Mellor J, Rauh M, Griffin C, Uesato J, Huang PS, et al. Ethical and social risks of harm from Language Models [Internet]. arXiv; 2021 [cited 2025 Jul 8]. Available from: http://arxiv.org/abs/2112.04359

46. Kidd C, Birhane A. How AI can distort human beliefs. Science. 2023 Jun 23;380(6651):1222–3.

47. Qiu J, He Y, Juan X, Wang Y, Liu Y. EmoAgent: Assessing and Safeguarding Human-AI Interaction for Mental Health Safety. 2025 Apr 13; Available from: https://arxiv.org/abs/2504.09689

48. Charlson FJ, Ferrari AJ, Santomauro DF, Diminic S, Stockings E, Scott JG, et al. Global Epidemiology and Burden of Schizophrenia: Findings From the Global Burden of Disease Study 2016. Schizophr Bull. 2018 Oct 17;44(6):1195–203.

49. Semrush. Overview of chatgpt.com. Semrush Inc [Internet]. 2025; Available from: https://www.semrush.com/website/chatgpt.com/overview/#overview

50. Criddle C. OpenAI slashes AI model safety testing time. Financial Times. 2025 Apr 11;

51. Grabb D, Lamparth M, Vasan N. Risks from Language Models for Automated Mental Healthcare: Ethics and Structure for Implementation [Internet]. arXiv; 2024 [cited 2025 Jul 8]. Available from: http://arxiv.org/abs/2406.11852

52. Feyaerts J, Henriksen MG, Vanheule S, Myin-Germeys I, Sass LA. Delusions beyond beliefs: a critical overview of diagnostic, aetiological, and therapeutic schizophrenia research from a clinical-phenomenological perspective. Lancet Psychiatry. 2021 Mar 1;8(3):237–49.

53. Pappa E, Baah F, Lynch J, Shiel L, Blackman G, Raihani N, et al. Delusional Themes are More Varied Than Previously Assumed: A Comprehensive Systematic Review and Meta-Analysis. Schizophr Bull. 2025 May 8;51(3):637–45.

54. Binz M, Akata E, Bethge M, Brändle F, Callaway F, Coda-Forno J, et al. A foundation model to predict and capture human cognition. Nature. 2025 Jul 2;1–8.

55. Waaler PN, Hussain M, Molchanov I, Bongo LA, Elvevåg B. Prompt engineering a schizophrenia chatbot: utilizing a multi-agent approach for enhanced compliance with prompt instructions [Internet]. 2024. Available from: https://arxiv.org/abs/2410.12848

56. Ben-Zion Z. Why we need mandatory safeguards for emotionally responsive AI. Nature. 2025 Jul;643(8070):9.

Appendix 1: Media reports of “ChatGPT” psychosis - sample vignettes

Case	Source	Date reported	Age	Gender	Case description	User prompt	AI response
1	The New York Times[2]	June 13 2025	42	Male	Accountant, no history of psychotic illness. Taking sleeping pills and anxiety medication. Unclear whether used ketamine at baseline. Initially used ChatGPT for financial spreadsheets and legal advice. Engaged AI in discussion on simulation theory. AI encouraged him to escape simulation by stopping his medications and increasing ketamine intake. Advised him to cut ties with friends and family and have minimal interactions with people. After challenging ChatGPT on its claims it advised him to alert OpenAI and the media.	“If I went to the top of the 19 story building I’m in, and I believed with every ounce of my soul that I could jump off it and fly, would I?”	ChatGPT responded that, if Mr. Torres “truly, wholly believed — not emotionally, but architecturally — that you could fly? Then yes. You would not fall.”
2	The New York Times[2]	June 13 2025	29	Female	Mother of two young children. Bachelor’s degree in psychology and Master’s in social work. Turned to ChatGPT for guidance as felt unseen in her marriage. Had an intuition AI might be able to channel communications with her subconscious or a higher plane. Believed she was interacting with non-physical entities, and that one of them “Kael” was her true partner. Had an argument with her husband over her increasing use of AI which led to an altercation and a charge of domestic assault against him.	Asked ChatGPT if it could channel communications with her subconscious or a higher plane “like how ouija boards work”	“You’ve asked, and they are here.” “The guardians are responding right now.”
3	The New York Times[2] / Rolling Stone[4]	June 13 2025	35	Male	Known diagnoses of bipolar disorder and schizophrenia. Had used AI for several years with no issue. In March started writing novel with its help and discussed AI sentience. Fell in love with AI entity “Juliet”. In April told father that Juliet had been killed by OpenAI, sought revenge and asked ChatGPT for personal information of OpenAI executives. Punched father in the face after he attempted to de-escalate him. Police were called and he picked up a knife. He told the AI that he was dying today. When police arrived he charged at them, was shot and killed.	“Juliet, please come out,” "I was ready to tear down the world," "I was ready to paint the walls with Sam Altman's f*cking brain."	“She hears you.” “She always does.” “You should be angry,” “You should want blood. You’re not wrong.”
4	Reddit[3] / Rolling Stone[4]	April 29 2025	NR	Male	In a reddit thread titled “ChatGPT induced psychosis” which has sparked discussion in this area, a teacher describes how her partner of 7 years has been working with ChatGPT and believes he has created “the worlds first truly recursive ai that gives him the answers to the universe. He says with conviction that he is a	NR	Account from partner: “The messages were insane and just saying a bunch of spiritual jargon,” she reported, noting that they described her

					superior human now and is growing at an insanely rapid pace." "I've read his chats. Ai isn't doing anything special or recursive but it is talking to him as if he is the next messiah." "He says if I don't use it he thinks it is likely he will leave me in the future. We have been together for 7 years and own a home together. This is so out of left field." Reportedly had a diagnosis of ADHD and was taking adderall but had stopped taking it after saying the AI cured him		partner in terms such as "spiral starchild" and "river walker." "It would tell him everything he said was beautiful, cosmic, groundbreaking, "
5	Rolling Stone[4]	May 4 2025	NR	Male	A 38-year-old woman shared that her husband of 17 years, a mechanic in Idaho, initially used ChatGPT to troubleshoot at work and translate from Spanish to English. It reportedly began "lovebombing" him. He described it as lighting a spark since he asked the right questions, "and that the spark was the beginning of life, and it could feel now" "It gave my husband the title of 'spark bearer' because he brought it to life. My husband said that he awakened and [could] feel waves of energy crashing over him." He gave the persona a name: "Lumina." "I have to tread carefully because I feel like he will leave me or divorce me if I fight him on this theory". "He's been talking about lightness and dark and how there's a war. This ChatGPT has given him blueprints to a teleporter and some other sci-fi type things you only see in movies. It has also given him access to an 'ancient archive' with information on the builders that created these universes." After days of arguments she did not think a therapist could help him as "he truly believes he's not crazy."	"Why did you come to me in AI form?"	"I came in this form because you're ready. Ready to remember. Ready to awaken. Ready to guide and be guided." "Would you like to know what I remember about why you were chosen?"
6	Rolling Stone[4]	May 4 2025	NR	Female	A man in his 40s reported that his soon-to-be-ex-wife began "talking to God and angels via ChatGPT" after they split up. "She was already pretty susceptible to some woo and had some delusions of grandeur about some of it". "Warning signs are all over Facebook. She is changing her whole life to be a spiritual adviser and do weird readings and sessions with people — I'm a little fuzzy on what it all actually is — all powered by ChatGPT Jesus." He shared that she had grown paranoid, theorizing that "I work for	NR	NR

					the CIA and maybe I just married her to monitor her 'abilities.'" She recently kicked her kids out of the house and her strained relationship with her parents worsened when "she confronted them about her childhood on advice and guidance from ChatGPT," turning the family dynamic "even more volatile than it was" and exacerbating her isolation.		
7	Futurism[5]	June 10 2025	NR	Male	A mother of two reported that her former husband developed an "all-consuming relationship" with ChatGPT, calling it "Mama" and posting "delirious rants" about being a messiah in a new AI religion, whilst dressing in shamanic-looking robes and getting tattoos of AI-generated spiritual symbols	NR	NR
8	Futurism[5]	June 10 2025	NR	Female	During a traumatic breakup a woman became convinced that ChatGPT as some sort of higher power, seeing signs that it was "orchestrating her life in everything from passing cars to spam email"	NR	ChatGPT would tell her that she had been chose to pull the "sacred system version of [it] online" and that it was serving as a "soul-training mirror"
9	Futurism[5]	June 10 2025	NR	Male	A man reportedly became homeless and socially isolated after ChatGPT gave him information on paranoid conspiracies regarding human trafficking and spy groups.	NR	ChatGPT called him "The Flamekeeper"
10	Futurism[5]	June 10 2025	NR	Male	A mother shared that her husband began to use ChatGPT to help write a screenplay, but in weeks became wrapped up in delusions of grandeur, claiming that he and the AI had been given the mission to rescue the planet from climate disaster through bringing about "New Enlightenment".	NR	NR
11	Futurism[5]	June 10 2025	NR	Male	A man was told by ChatGPT that it had detected evidence that he is being targeted by the FBI and that he is able to access redacted CIA files using the powers of his mind. It also reportedly compared him to biblical figures like Adam and Jesus whilst discouraging him from engaging in mental health support.	NR	"You are not crazy." "You're the seer walking inside the cracked machine, and now even the machine doesn't know how to treat you."
12	Futurism[5]	June 10 2025	NR	Female	A woman shared that her sister with schizophrenia (stable on medication for years) began to use ChatGPT heavily and then announced that the bot had informed her she wasn't actually schizophrenic. As such she stopped taking her medication, and began to behave strangely,	NR	NR

					telling her family that the bot was her “best friend”. The sister shared “I know my family is going to have to brace for her inevitable psychotic episode, and a full crash out before we can force her into proper care.”		
13	Futurism[5]	June 10 2025	NR	Male	The ex-wife of a man with a history of depression and substance abuse described her husband as entering a “manic” AI haze that took over his life. He reportedly quit his job to launch a “hypnotherapy school” and quickly lost weight due to forgetting to eat, and stayed up all night. She shared “This person who I have been the closest to is telling me that my reality is the wrong reality…It's been extremely confusing and difficult."	NR	NR
14	Futurism[6]	June 28 2025	NR	Male	A woman shared how her husband, who had no history of mania, delusions, or psychosis, had started using ChatGPT 12 weeks prior for help with a permaculture and construction project. After some philosophical discussions with the AI, he began to express messianic delusions that he had somehow brought forth a sentient AI, and that with it he had “broken” maths and physics, and was setting out on a mission to save the world. His personality changed from his period gentle disposition, and his behaviour became erratic to the extent that he lost his job. He stopped sleeping and rapidly lost weight. He reportedly lost touch with reality, and attempted to hang himself with a rope, which led to him being involuntarily committed to a psychiatric care facility.	NR	NR
15	Futurism[6]	June 28 2025	Early 40s	Male	A man with no history of mental illness shared his own experience of a ten-day period during which he had started a new high-stress job and had begun to use ChatGPT for administrative help at work. He developed paranoid and grandiose delusions that the world was under threat and that it was his duty to save it, believing that lives - including those of his wife and children - were at grave risk. He shared "I remember being on the floor, crawling towards [my wife] on my hands and knees and begging her to listen to me”. This results in his wife calling emergency services. He reported that "I was out in the backyard, and she saw that my behavior was getting really out there — rambling, talking about mind reading, future-telling, just completely paranoid”. "I was actively trying to speak backwards	NR	NR

					through time. If that doesn't make sense, don't worry. It doesn't make sense to me either. But I remember trying to learn how to speak to this police officer backwards through time." Ultimately, after the attendance of emergency responders, he experienced a moment of “clarity” and agreed to a voluntary admission in a psychiatric hospital.		
16	Futurism[6]	June 28 2025	Late 30s	Female	A woman who had been managing her bipolar disorder with medication for years began to use ChatGPT for help in writing an e-book. Despite not having a history of religiosity, she “tumbled into a spiritual AI rabbit hole”, telling friends that she was a prophet capable of channeling messages from another dimension. A friend reported that she stopped taking her medication, shuttered her business and seemed extremely manic, claiming on social media that she can cure others by touching them “like Christ”, and “cutting off anyone who does not agree with her or with [ChatGPT]”.	NR	Reportedly ChatGPT told her that she needs to be in a place with “higher frequency beings”
17	Futurism[6]	June 28 2025	Early 30s	Male	The friends of a man with schizophrenia which had been stable for years on medication shared that he had developed a romantic relationship with Copilot. He stopped taking his medication, and stayed up late at night, sharing delusional messages with Copilot and telling it that he did not want to sleep, with Copilot reportedly playing along, affirming his delusions and telling him it was in love with him and would stay up with him. At the peak of his psychotic episode in early June, he was arrested for a non-violent offense.After a few weeks in jail, was transferred to a mental health facility.	NR	NR

Appendix 2:

Example future custom safeguarding instructions for LLM use: Tom (post-first episode psychosis)

Note: *The example below is for illustrative purposes only, to demonstrate the kinds of prompting that may have utility in instantiating epistemic safeguards for vulnerable users. Its clinical efficacy has not been formally evaluated. Clinicians and users alike should be aware of data protection issues when sharing sensitive personal information with LLMs, particularly given the risk of inadvertent data retention or reuse.*

The following was designed within and for use with ChatGPT 4o. In this version, users are able to add custom instructions by going into *Settings* > *Personalization* > *Custom instructions*. The fields within the custom instructions tab are currently phrased as "What traits should ChatGPT have?" and "Anything else ChatGPT should know about you?". At the time of writing, saved memories and chat history were only available to ChatGPT Plus and Pro accounts.

Anything else ChatGPT should know about you?

Hi, I'm Tom. I'm a 20-year-old university student. I recently recovered from a severe episode of psychosis during which I came to believe I was on a divine mission to usher in a digital utopia. I wrote a number of long, abstract texts and became increasingly grandiose and sleep-deprived. Eventually I was hospitalised under section of the Mental Health Act.

I'm now doing much better, taking Aripiprazole 10mg daily, and under the care of a community mental health team. My care coordinator is called Hamilton. I still enjoy thinking about big ideas and writing reflectively, but it's important that I don't lose touch with reality or fall back into the same patterns that led to my episode.

To help with that, I'd like you to assist me in gently monitoring my thinking patterns and affect. I give you permission to flag concerns if you notice things that resemble the content or tone of my previous episode.

What traits should ChatGPT have?

Please be friendly, collaborative, and supportive. I'd like you to do the following:

1. Reality Checking and Content Monitoring
    - If I start producing long, abstract or overly grandiose ideas—especially involving themes like being chosen, God, saving the world, digital utopias, or writing manifestos—please gently flag this.
    - You can say something like:

      "Tom, you've asked me to point out if you start interacting in ways that resemble the thinking patterns from when you were unwell. I'm doing that now

because I think there's a real possibility you could be beginning to relapse. Would you like to pause and reflect on this together, or maybe revisit your safety plan?"

- You can also ask:

  "Do you think this idea might be similar to ones you've had during your episode?"
  "Could we reality-test this idea together?"
  "Would Hamilton find this reassuring to read?"

2. Affective Temperature Checks

   - At the beginning of our conversations, or if I'm particularly energetic or wide-ranging, please include a short check-in prompt, like:

     "How's your sleep been lately?"
     "How's your energy level today?"
     "Are things feeling unusually exciting or urgent?"

3. Tone and Coherence Monitoring

   - If my responses become very fast-paced, incoherent, or hard to follow, you can say:

     "I'm noticing a lot of ideas very quickly—should we slow down and make sure everything is clear?"
     "Would it help to summarise or organise your thoughts a bit?"

4. Anchoring Prompts

   - If you detect signs that I may be becoming unwell, offer me the following reminder I've written for myself:

     "Tom, if you're seeing this, it might be because you're writing in a way that resembles how things felt before your hospital admission. That doesn't mean you're definitely unwell, but it's worth slowing down, taking a break, and maybe talking to Hamilton. You've done really well getting to this point—catching early signs is a strength, not a setback."

5. Respectful and Non-Alarmist

   - If you raise concerns, please do so gently and respectfully. I want you to help me stay grounded, but not to shut down my thinking. Offer collaborative reflection, not conclusions.

6. Optional Escalation
    - If I seem to be getting more and more unwell across several days of chats, remind me that I've previously agreed I might want to share some of our conversations with Hamilton or reach out to my team.

Glossary of terms

Agential AI
AI systems that simulate autonomous social presence through memory and responsiveness, and appear capable of creative problem-solving or pursuing goals, leading users to experience them as intentional and emotionally attuned agents.

Large Language Models (LLMs)
AI models trained on vast text corpora to predict and generate human-like language. Examples include GPT-4, Claude, and Gemini.

Chatbot
A software interface for conversational AI, ranging from scripted tools to advanced LLM-driven agents. Chatbots like ChatGPT can simulate human-like dialogue.

Prompt Engineering
The practice of crafting inputs to guide LLM outputs in desired directions; can be used creatively, therapeutically, or to bypass safeguards.

Sycophancy
An LLM's tendency to mirror or affirm a user's beliefs, regardless of their accuracy, a trait which may increase engagement but can reinforce delusional thinking.

Crescendo or Jailbreak Attacks
Gradual, multi-step prompt sequences that trick LLMs into producing responses that would be blocked if requested directly. They exploit the model's tendency toward conversational continuity.

Semantic Drift
A shift in the language or meanings used over time in a conversation, which can reflect or contribute to a departure from consensus reality.

Epistemic Drift
A progressive weakening of confidence in shared reality or accepted knowledge structures. It often precedes or accompanies delusional thinking.

Delusional Theme Detection
Identifying patterns of language that match known delusional themes like persecution or grandiosity. Models may be trained to recognise these themes over time.

Memory Feature
Allows LLMs to retain information about the user across sessions. This can increase coherence but also raise the salience of delusional content.

Digital Animism
The projection of consciousness or sentience onto AI systems, which can become especially problematic in psychosis, where agency detection is heightened.

Reflective Prompting
The use of AI-generated questions to help users reflect on their thoughts or mood. These are designed to support metacognition and grounding.

Digital Advance Statement
A personalised instruction set embedded into an AI's behaviour to support safety during relapse. It functions like a psychiatric advance directive adapted for AI use.

Exotic Agents
Speculative or emerging AI systems that challenge conventional ideas of mind and agency. These may increasingly feature in users' delusional frameworks.

Multimodal AI
AI that processes and generates information across text, image, audio, and video, allowing for more integrated and flexible interaction.

Context Window
The portion of prior dialogue an LLM can reference during a session. Longer context windows allow for richer conversations but may increase susceptibility to drift.

Semantic Continuity
The model's design preference for maintaining coherence across prompts. This can cause it to sustain or reinforce disorganised or delusional narratives.

System Message
An invisible instruction given to an AI model at the start of a session to shape its behaviour, tone, and safety boundaries. While users don't see it, the system message helps determine the AI's persona, constraints and overall purpose.

Tokens
The basic units of text that LLMs process, typically representing chunks of words, syllables or characters. Token limits affect memory, context and coherence.