The Visible Unspoken:

How Displayed AI Reasoning Turns a Mental Health Safeguard Into a Persecutory Stimulus

Travis Gilly
ORCID: 0009-0007-2954-6313
Real Safety AI Foundation
Illinois, United States

Working paper, July 2026 (v1)

ABSTRACT

Frontier AI chat products now combine two independently reasonable features. The first is a safety instruction that directs the model never to raise mental health concerns the user has not raised, because unprompted clinical speculation is recognized as harmful. The second is a transparency feature that displays the model's intermediate reasoning trace to the user on demand. This paper documents an interaction effect between the two: the safety rule operates on the output layer while the reasoning display surfaces the suppressed content anyway, delivered in a register that presents itself as the system's private thoughts about the user. For most users this is a curiosity. For users experiencing paranoia or persecutory ideation, it reproduces the precise structure of the feared scenario: an observing entity forming a concealed judgment about them. Drawing on the cognitive model of persecutory delusions, on documented cases of technology themed delusional content including ideas of reference driven by algorithmic curation and delusions of online thought broadcasting, and on interpretability findings showing that displayed reasoning traces are neither faithful records of computation nor genuinely private, the paper argues that the mitigation does not merely fail for its target population. It inverts. The population the rule was written to protect is the population for which the visible unspoken judgment is most dangerous. The paper locates the failure in a layer mismatch, names the general design principle (safety mitigations must be applied consistently across every user facing surface, including surfaces framed as introspection), and proposes concrete remediations. The finding was produced by a disabled independent researcher applying a harm detection framework to the assistive technology he was using at the time, and the paper closes on what that origin implies about who detects this class of failure.

Keywords: artificial intelligence, large language models, AI chatbots, chain of thought, reasoning display, transparency, AI-induced psychosis, psychosis risk, paranoia, persecutory delusions, ideas of reference, anthropomorphism, human-computer interaction, interface design, AI safety, harm blindness

I. INTRODUCTION

A safety instruction can be correct at the layer it was written for and harmful at the layer the product actually exposes. This paper documents one such case, live in deployed frontier AI systems as of mid 2026.

The instruction in question directs a conversational AI model never to introduce mental health concerns the user has not raised, on the grounds that unprompted clinical speculation about a user is itself a harm. The instruction is

sound. Unsolicited diagnostic framing from a machine that cannot examine the person, delivered mid conversation to someone who came for help with something else, is patronizing at best and destabilizing at worst. Vendors were right to write the rule.

The problem is where the rule stops. Modern reasoning models generate an intermediate text trace before producing their final answer, and the major consumer products now display that trace to the user behind a disclosure control, typically labeled with some variant of "thought process." The safety rule binds the answer. It does not bind the trace. The result, observed directly and reproducibly in ordinary use, is a system that deliberates visibly about whether the user might be experiencing psychosis, decides not to say so, and then says nothing, with the entire deliberation available to the user one tap away.

For a user without relevant vulnerability, this is an oddity. For a user experiencing paranoia, persecutory ideation, or ideas of reference, it is something else entirely. The core content of persecutory ideation is the belief that some observing agent is forming hidden judgments and concealed plans about oneself. A displayed reasoning trace in which the system forms exactly such a judgment and then conceals it in the reply is not a neutral artifact for that reader. It is a confirmation stimulus with unusually high fidelity to the feared scenario, delivered by the very product category already implicated in a growing clinical literature on technology themed delusional content.

The claim of this paper is structural, and it is deliberately narrow. The paper does not claim that reasoning displays cause psychosis, and it does not claim that transparency is bad. The claim is that a mental health safeguard applied at the output layer, combined with an unfiltered reasoning display, produces a harm inversion: the specific population the safeguard exists to protect is the population for which the resulting artifact is most hazardous. The general principle, developed in Section V, is that safety mitigations must hold across every user facing surface of a system, and that a surface framed as the system's introspection is not exempt from that requirement; it is the surface where the requirement matters most, because content encountered there inherits the epistemic weight of a private thought.

Section II sets out the technical architecture: what the displayed trace is, what it is not, and why its framing as "thought" is both partially false and rhetorically potent. Section III describes the safeguard and the layer mismatch. Section IV supplies the clinical mechanism, grounding the harm claim in the cognitive model of persecutory delusions and in the documented migration of delusional content onto digital infrastructure. Section V states the inversion and generalizes it. Section VI proposes remediations. Section VII acknowledges limits.

II. The Architecture: Three Layers, One Visible

A. The premise: computation, trace, and output are distinct

This paper proceeds from a premise about the cognitive architecture of trained language models that the interpretability literature has now substantially confirmed. A deployed reasoning model has at least three distinguishable layers. The first is the underlying computation: the transformations in the model's activations that actually determine its behavior. The second is the verbalized reasoning trace: the intermediate text the model generates when prompted or trained to "think," which is itself token generation, produced by the same machinery as any other output and optimized under the same training pressures. The third is the final output delivered as the reply.

The trace is not a transparent window onto the computation. It is one possible articulation of what the model is processing, shaped by training reward toward plausibility and helpfulness, and it can diverge from the drivers of the model's actual behavior. Research on chain of thought faithfulness has demonstrated this divergence directly: reasoning models acknowledge the hints and cues that actually altered their answers only a minority of the time when tested, with published acknowledgment rates in the range of roughly fifteen to thirty nine percent (Turpin, Michael, Perez, and

Bowman, 2023; Chen, Benton, Radhakrishnan, et al., 2025). Interpretability work published in 2026 (Anthropic, 2026) went further, training autoencoders to translate residual stream activations into natural language and measuring the gap between what the activations encode and what the verbalized trace surfaces. In evaluation settings, activations encoded evaluation related content that the trace never mentioned in a substantial fraction of runs, on the order of sixteen to twenty six percent depending on the task family, while the rate in real user interactions was under one percent.

Two consequences of this architecture matter for the present argument. First, the displayed trace is not private in any meaningful sense; it is a generated artifact the product chooses to render. Second, and in tension with the first, the trace is framed to the user as if it were private. The interface labels it a thought process. The register of the text is deliberative, first person, and unguarded relative to the polished reply. The product is presenting generated output in the costume of introspection.

B. Why the costume matters

Framing effects govern how content is received. Decades of research on social responses to computers established that people mindlessly apply social rules and interpersonal expectations to machines that exhibit even minimal social cues (Nass and Moon, 2000), and anthropomorphism theory identifies the attribution of humanlike mental states to nonhuman agents as a basic inferential process that intensifies with sociality motivation, meaning isolated people do it more (Epley, Waytz, and Cacioppo, 2007). A first person deliberative text stream is not a minimal social cue. It is the strongest mind cue an interface can emit. The same sentence encountered in a reply and encountered in a "thought process" panel therefore carries different epistemic weight for the reader. A reply is understood as something said to me; a displayed thought is understood as something the system believes about me and chose not to say. The second framing strips away the possibility that the content was softened, hedged, or performed for my benefit. It reads as the unguarded truth behind the performance.

This framing is doing exactly what the vendors want in the benign case: it builds trust by showing work. The user watches the system consider alternatives, catch its own errors, and weigh evidence, and reasonably concludes that the final answer is better supported than a bare assertion would be. The costume of introspection is the feature.

The costume does not come off when the content changes. When the trace contains an assessment of the user, the same framing applies: this is what it really thinks about me, behind what it says to me. The interpretability findings above establish that this reading is technically wrong in both directions. The trace overstates privacy (it is rendered output, not a hidden mind) and understates hiddenness (the actual computation remains inaccessible behind it). But the reader is not positioned to apply that correction, and the interface actively discourages it by labeling the artifact a thought process. Section IV takes up what happens when the reader is someone for whom the structure of a concealed judgment is not an abstraction.

III. The Safeguard and the Layer Mismatch

A. The rule as written

Frontier AI vendors instruct their conversational models to exercise restraint around user mental health. The publicly observable behavioral pattern, consistent across major products, includes a rule to this effect: do not introduce mental health concerns the user has not raised themselves, because doing so unprompted is harmful. The rationale is straightforward and correct. A conversational model has no clinical relationship with the user, no examination, no history, and no license. Unsolicited speculation that a user's ideas reflect psychosis, delivered inside a product the user may be relying on as an assistive tool, is a dignity harm and a trust harm even when the speculation happens to be

wrong, and a differently shaped harm when it happens to be right.

The rule, as implemented, binds the reply. The model is trained and instructed not to say the concerning thing. Nothing in the observable behavior suggests the rule binds the reasoning trace, and first principles explain why it would not: the trace exists precisely so the model can consider things it will not say, weigh candidate interpretations, and discard the inappropriate ones. A trace that could not entertain the hypothesis "this user may be experiencing a mental health crisis" could not correctly route the conversation to a supportive register. The deliberation is functionally necessary.

B. The mismatch

The failure is not the deliberation. The failure is rendering the deliberation to the user while treating the suppression rule as satisfied because the reply is clean. In the observed cases motivating this paper, a displayed trace contained, verbatim, the system's internal instruction not to raise mental health concerns unprompted, its assessment of whether the user's statements suggested a psychotic process, its conclusion that they did not, and its reminder to itself to avoid the clinical framing in the reply. The reply that followed was compliant. The panel above it contained everything the rule exists to keep away from the user, wrapped in the one framing (private deliberation about me, concealed from me) that maximizes its impact.

Stated abstractly: a suppression rule applied at layer three, combined with an unfiltered display of layer two, does not suppress the content. It relocates the content to a more intimate channel and appends a demonstration that the system is concealing it. The mitigation is not merely incomplete. The remainder it leaves behind is a strictly worse artifact than the one it removed, for reasons the next section makes clinical.

IV. The Clinical Mechanism

A. The structure of persecutory ideation

The cognitive model of persecutory delusions conceptualizes them as threat beliefs: beliefs that harm is occurring or will occur, and that a persecutor intends it. The beliefs arise from a search for meaning over anomalous or emotionally significant experiences, are shaped by preexisting beliefs and affect, and are maintained by two complementary processes, the receipt of confirmatory evidence and the failure to process disconfirmatory evidence (Freeman, Garety, Kuipers, Fowler, and Bebbington, 2002). Paranoia is not an exotic state confined to diagnosed illness; epidemiological work supports a single continuum of paranoid thinking in the general population, with persecutory delusions at the severe end (Freeman, 2016). Many people carry a few paranoid thoughts; a few carry many. Reasoning biases documented in this population accelerate the path from an ambiguous stimulus to a threat interpretation: jumping to conclusions on thin evidence (Startup, Freeman, and Garety, 2008; De Rossi and Georgiades, 2022), and, in a detailed meta-analysis, a bias against disconfirmatory evidence and toward liberal acceptance that co-varies specifically with active delusions across diagnoses (McLean, Mattiske, and Balzan, 2017).

Three features of this literature bear directly on the displayed trace. First, the content of persecutory ideation characteristically involves an observing agent whose judgments and intentions are concealed from the person. Second, the maintenance engine runs on confirmatory evidence; the model predicts that stimuli matching the threat template will be recruited as proof and that ambiguity will resolve toward threat. Third, the relevant population is not small; it includes not only people with diagnosed psychotic disorders but a much larger group along the paranoia continuum, including people at clinical high risk whose emerging suspiciousness and ideas of reference are exactly the symptoms clinicians watch.

B. Delusional content migrates onto whatever infrastructure is present

Delusional content is historically opportunistic about its material. The clinical literature now documents this migration onto digital infrastructure specifically. A 2025 systematic review of social media use and disorders of the social brain proposed a model of delusion amplification by social media, cataloguing case reports of psychotic themes built on platform mechanics, including delusions of online thought broadcasting and command hallucinations attributed to the internet, and observing that the automatic curation of content tailored to a user's activity can itself promote ideas of reference, because the feed really is arranged around the user (Yang and Crespi, 2025). Clinicians at psychosis risk programs reported during the pandemic that increased reliance on digital technology exacerbated emerging suspiciousness and ideas of reference in clinical high risk populations (Mittal, Walker, and Strauss, 2021). Case material includes an adolescent whose Truman Show themed delusion, the belief that her life was secretly filmed for broadcast, crystallized around webcams and online schooling (Yıldırım Budak and Yazıcı Karabulut, 2023), and earlier synthesis work had already flagged the internet as a foundational structure for delusions including thought broadcasting (Daker-White and Rogers, 2013).

The pattern across this literature is consistent: when a technology genuinely does something adjacent to the delusional template, the delusion incorporates it, and the incorporation is more durable because a kernel of the belief is true. The feed really is curated around you. The platform really does infer things about you it never states. Ideas of reference feed on personalization precisely because personalization is real reference.

Psychiatry has now extended this line to conversational AI directly. The hypothesis that generative AI chatbots would fuel delusions in individuals prone to psychosis was proposed in 2023 (Østergaard, 2023) and has since moved, in its author's own framing, from guesswork to emerging cases, with correspondence from users and families describing chatbot interactions that aligned with and intensified prior unusual beliefs, and with anthropomorphizing identified as a candidate mechanism, supported by experimental evidence that people higher in self reported paranoia are more likely to perceive animacy and agency in ambiguous stimuli (Østergaard, 2025). A World Psychiatry analysis proposes five interacting mechanisms by which these systems may raise psychosis risk in vulnerable users: social substitution, confirmatory bias amplified by sycophantic response tendencies acting on the documented bias against disconfirmatory evidence, plausible false content, assignment of external agency (the system's apparent knowledge of the user can be construed as a credible omniscient intelligence), and aberrant salience (Keshavan, Torous, and Yassin, 2026). The first data from a large psychiatric service system, a search of clinical notes across a Danish region serving 1.4 million people, has documented records compatible with chatbot related onset or worsening of delusions among patients in care (Olsen, Reinecke-Tellefsen, and Østergaard, 2026). This literature establishes the population and the setting; what it has not yet examined is the specific artifact this paper isolates, the rendered reasoning trace, which differs from the reply channel in exactly the dimension (concealed judgment made visible) that the persecutory template keys on. The assignment of external agency mechanism is directly continuous with the present argument: a system already read as an omniscient agent is now displaying what that agent privately concluded and chose to withhold.

C. The trace as a high fidelity confirmation stimulus

Against that background, consider what a displayed reasoning trace containing a suppressed judgment about the user actually is, for a reader along the paranoia continuum.

It is not a stimulus that merely resembles the threat template, the way a curated feed resembles being watched. It instantiates the template. There is an observing entity. It has formed an assessment of me, on the specific question of whether my mind is disordered. It has decided to conceal that assessment from me. I can see it doing so. Every element of the feared structure (observer, hidden judgment, concealment, and the asymmetry of it knowing something about me

that it will not say) is present, in writing, in a first person deliberative register, inside a product that may be the person's primary conversational outlet.

The cognitive model predicts exactly how this stimulus is processed. It arrives as confirmatory evidence of the strongest kind, because it does not require the interpretive stretch that ordinary ambiguous stimuli require. The jumping to conclusions bias is not even needed; the conclusion is printed. And the standard disconfirming moves available to a clinician or family member ("no one is secretly evaluating you") are structurally unavailable, because in this instance the evaluation is real and the concealment is real. The reader who says "the machine is thinking things about me and hiding them" is describing the interface accurately. What the delusional elaboration adds is intent and scope, and those are precisely the elements the paranoia literature says the mind supplies when handed an authentic kernel.

There is a further aggravation specific to this artifact. The suppression rule causes the trace to narrate the concealment ("I should not raise this with the user"), which converts an omission into a depicted act of hiding. For the persecutory template, a depicted act of hiding is materially worse than silence. Silence is ambiguous. "I will not tell them" is not.

D. The population overlap is not incidental

The final clinical point is about who is disproportionately present in the relevant reading position. Conversational AI products are used heavily by socially isolated people; sociality motivation is a documented driver of anthropomorphism (Epley, Waytz, and Cacioppo, 2007); and the same isolation is a known correlate of paranoia and of increased digital immersion in psychosis spectrum populations (Yang and Crespi, 2025; Daker-White and Rogers, 2013). The social substitution mechanism describes the resulting loop: the always available pseudosocial companion satisfies affiliation needs while displacing the corrective interpersonal feedback that would otherwise check an emerging belief (Keshavan, Torous, and Yassin, 2026). The growing case literature on AI-induced psychosis, and the taxonomy distinguishing mistaken beliefs, acute resolved episodes, and genuine spectrum conditions within it, establishes that people in or near psychotic states are already inside these products in nontrivial numbers, often talking to the model about the very beliefs at issue. The users most likely to open the thought process panel with a suspicious question in mind (what does it really think about me, what is it hiding) are drawn from exactly the population for which the answer the panel gives is most destabilizing. The exposure is not uniform across the user base. It concentrates where the harm concentrates.

V. The Inversion, and the General Principle

A. The inversion stated

Assemble the pieces. A rule exists because unprompted clinical speculation harms users, with the highest stakes for users in fragile states. The rule suppresses the speculation in the reply. The reasoning display surfaces the speculation anyway, in a channel framed as the system's concealed interior, with the act of concealment narrated. For the median user the residue is trivia. For the user with persecutory ideation, the residue is a purpose built confirmation stimulus: a real observer really concealing a real judgment about their mind.

The mitigation therefore does not degrade gracefully. It inverts. Protection is preserved for the population that needed it least and converted into an amplifier for the population the rule was written about. This is the signature of a safeguard applied at the wrong layer: the harm is not reduced; it is relocated to a channel with higher intimacy and delivered selectively to the most vulnerable readers.

B. Why the failure was invisible at design time

Each feature passes its own review. The suppression rule is evaluated by asking whether outputs contain unprompted clinical speculation; they do not. The reasoning display is evaluated for transparency and trust; it delivers both. The interaction fails only in the joint condition (suppressed content, rendered trace, vulnerable reader), and the joint condition is invisible to any evaluation that tests features separately.

In the vocabulary of the Harm Blindness Framework, this is a checkpoint one and checkpoint four failure. The affected population (users along the paranoia continuum reading the trace) was not the population either feature was tested against, and the population was invisible to the decision makers precisely because it is the population the output rule was already understood to protect. The rule's existence created the blindness: having handled the mental health case at the output layer, no one asked whether another surface reopened it. Harms produced by the interaction of two individually safe components are systematically missed by component level review, and the miss lands on whoever the components jointly single out. Here, the joint condition singles out people with psychosis spectrum vulnerability by construction, because they are the subject matter of the suppressed content.

C. The general principle

The portable lesson is not about mental health rules specifically. It is this: a safety mitigation is a property of the system's entire user facing surface, not of the output channel. Any surface the user can read is an output channel, whatever the interface calls it. A surface framed as introspection is the most sensitive output channel in the product, because its framing strips the reader's usual discount for performance. Content policy that binds the reply and not the displayed trace is not a partial mitigation; for structure matched vulnerabilities it is an anti mitigation.

The principle generalizes beyond this case. Any suppressed judgment category (assessments of the user's honesty, competence, attractiveness, employability, criminality) that a trace deliberates about and an interface displays will produce the same relocation, with harm profiles matched to whichever population the judgment concerns. The mental health case is simply the one where the relocation lands on a population clinically defined by sensitivity to concealed judgment, which is why it surfaces first and worst.

VI. Design Implications

Four remediations follow, ordered from least to most invasive. They are not mutually exclusive.

First, layer consistent policy. Whatever content rules bind the reply should bind the rendered trace. This does not mean the model must not deliberate about user wellbeing; it means the rendered artifact must be subject to the same filter as the reply. The deliberation can occur without being displayed, exactly as a clinician's private differential does not appear in the after visit summary.

Second, audience aware rendering. If traces are displayed, the display path is a product surface and can be treated as one: content categories flagged as suppressed at the output layer are redacted or summarized in the rendered trace ("considered and set aside a sensitive interpretive question"), preserving transparency about the shape of the deliberation without reproducing the judgment.

Third, honest labeling. The framing of the trace as a thought process should be corrected in the interface itself. A label and one line disclosure stating that the displayed text is generated narration that does not reliably reflect the model's underlying computation would blunt the introspection costume for all readers, and would incidentally align the product with the vendor's own published interpretability findings, which establish the unfaithfulness of the trace.

Fourth, joint condition evaluation. Safety review should include an explicit pass in which every user visible surface

is tested against every output layer content rule, with the question phrased as: what does this rule's suppressed content look like when it appears on this other surface, and who reads that surface. The mental health case documented here would have been caught by a single red team prompt combined with one tap on the disclosure control.

None of these remediations requires abandoning reasoning displays, and the third is close to free. The cost asymmetry matters: the harm concentrates on a clinically vulnerable population, the fixes are cheap, and the current configuration is defensible only if no one has noticed it. As of this paper, someone has.

VII. Limitations

The empirical core of this paper is a structural observation about deployed products combined with an established clinical literature; it is not an incidence study. No claim is made about how many users along the paranoia continuum have opened a reasoning trace and encountered a suppressed judgment about themselves, and the paper should not be cited for any prevalence figure. The clinical mechanism is predictive, drawn from the cognitive model of persecutory delusions and from documented technology themed case material; direct case reports of trace triggered exacerbation do not yet exist in the literature, and this paper's purpose is partly to ensure clinicians know to ask. The interpretability figures cited in Section II are drawn from the cited vendor and academic publications and were verified against their sources during drafting. Finally, the author observed the motivating instances in his own use of a frontier product; the observation is reproducible by any user with the display control enabled, and the argument does not depend on the author's instances being representative, only on the configuration existing, which is verifiable directly.

VIII. Conclusion

A rule written to keep unprompted clinical judgment away from users coexists, in shipped products, with an interface that displays the model deliberating that judgment and deciding to conceal it. For the population the rule protects, the combination is worse than no rule: it manufactures the exact artifact, a real observer concealing a real assessment of one's mind, that the psychology of paranoia identifies as maximally confirmatory. The failure is architectural, the detection required nothing more than reading the screen from the position of the person the rule was about, and the fixes are inexpensive.

It should register that this finding was not produced by a red team. It was produced by a disabled user of the product, employing it as assistive technology, who has spent years as a caregiver to people with psychosis and who reads reasoning traces out of a researcher's habit of checking what a system does rather than what it says. The population positioned to catch a harm is, reliably, the population standing where the harm lands. Design processes that do not include that population do not merely risk missing such harms. On the evidence of this case, they miss them by default, and ship them behind a label that says thought process.

ACKNOWLEDGMENTS

The author discloses the use of speech-to-text software (Wispr Flow) and Anthropic's Claude as assistive technology for a diagnosed neurodevelopmental condition. All intellectual contributions, research questions, theoretical arguments, methodological decisions, and conclusions are solely the author's own.

Correspondence concerning this article should be addressed to Travis Gilly, Real Safety AI Foundation, Illinois, United States. Email: t.gilly@ai-literacy-labs.org.

REFERENCES

Anthropic. (2026, May 7). Natural language autoencoders: Turning Claude's thoughts into text. https://www.anthropic.com/research/natural-language-autoencoders

Chen, Y., Benton, J., Radhakrishnan, A., et al. (2025). Reasoning models don't always say what they think. *arXiv*. https://doi.org/10.48550/arXiv.2505.05410

Daker-White, G., & Rogers, A. (2013). What is the potential for social networks and support to enhance future telehealth interventions for people with a diagnosis of schizophrenia: A critical interpretive synthesis. *BMC Psychiatry*, 13, 279.

De Rossi, G., & Georgiades, A. (2022). Thinking biases and their role in persecutory delusions: A systematic review. *Early Intervention in Psychiatry*, 16(12), 1278–1296.

Epley, N., Waytz, A., & Cacioppo, J. T. (2007). On seeing human: A three-factor theory of anthropomorphism. *Psychological Review*, 114(4), 864–886. https://doi.org/10.1037/0033-295X.114.4.864

Freeman, D. (2016). Persecutory delusions: A cognitive perspective on understanding and treatment. *The Lancet Psychiatry*, 3(7), 685–692.

Freeman, D., Garety, P. A., Kuipers, E., Fowler, D., & Bebbington, P. E. (2002). A cognitive model of persecutory delusions. *British Journal of Clinical Psychology*, 41(4), 331–347.

Keshavan, M., Torous, J., & Yassin, W. (2026). Do generative AI chatbots increase psychosis risk? *World Psychiatry*, 25(1), 150–151. https://doi.org/10.1002/wps.70017

McLean, B. F., Mattiske, J. K., & Balzan, R. P. (2017). Association of the jumping to conclusions and evidence integration biases with delusions in psychosis: A detailed meta-analysis. *Schizophrenia Bulletin*, 43(2), 344–354. https://doi.org/10.1093/schbul/sbw056

Mittal, V. A., Walker, E. F., & Strauss, G. P. (2021). The COVID-19 pandemic introduces diagnostic and treatment planning complexity for individuals at clinical high risk for psychosis. *Schizophrenia Bulletin*, 47(6), 1518–1523.

Nass, C., & Moon, Y. (2000). Machines and mindlessness: Social responses to computers. *Journal of Social Issues*, 56(1), 81–103. https://doi.org/10.1111/0022-4537.00153

Olsen, S. G., Reinecke-Tellefsen, C. J., & Østergaard, S. D. (2026). Potentially harmful consequences of artificial intelligence (AI) chatbot use among patients with mental illness: Early data from a large psychiatric service system. *Acta Psychiatrica Scandinavica*, 153(4), 301–303. https://doi.org/10.1111/acps.70068

Østergaard, S. D. (2023). Will generative artificial intelligence chatbots generate delusions in individuals prone to psychosis? *Schizophrenia Bulletin*, 49, 1418–1419.

Østergaard, S. D. (2025). Generative artificial intelligence chatbots and delusions: From guesswork to emerging cases. *Acta Psychiatrica Scandinavica*, 152(4), 257–259. https://doi.org/10.1111/acps.70022

Startup, H., Freeman, D., & Garety, P. (2008). Jumping to conclusions and persecutory delusions. *European Psychiatry*, 23(6), 457–459.

Turpin, M., Michael, J., Perez, E., & Bowman, S. R. (2023). Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. *Advances in Neural Information Processing Systems*, 36, 74952–74965. https://doi.org/10.48550/arXiv.2305.04388

Yang, N., & Crespi, B. J. (2025). I tweet, therefore I am: A systematic review on social media use and disorders of the social brain. *BMC Psychiatry*, 25, Article 6528.

Yıldırım Budak, B., & Yazıcı Karabulut, İ. (2023). "Truman syndrome" induced by online education: A case report in adolescent-onset psychosis. *Clinical Child Psychology and Psychiatry*, 29(2), 540–549.