CHAIN OF THOUGHT · REASONING DISPLAY · LAYER MISMATCH · PERSECUTORY DELUSIONS · IDEAS OF REFERENCE · ANTHROPOMORPHISM · HARM BLINDNESS · HARM INVERSION

The Visible Unspoken

How displayed AI reasoning turns a mental health safeguard into a persecutory stimulus

Frontier chat products combine two reasonable features: a safety rule that forbids the model from raising mental health concerns the user has not raised, and a transparency panel that displays the model's reasoning trace on demand. The rule binds the reply. It does not bind the trace. The result is a system that visibly deliberates about whether the user may be experiencing psychosis, decides not to say so, and leaves the entire deliberation one tap away. For most users this is a curiosity. For a user along the paranoia continuum it instantiates the feared structure exactly: an observer forming a concealed judgment about them. The mitigation does not fail. It inverts.

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

Working paper, July 2026 (v1)

A safety rule can be correct at the layer it was written for and harmful at the layer the product exposes.

The rule directs a conversational model **never to introduce mental health concerns the user has not raised.** Unsolicited diagnostic framing from a machine that cannot examine the person is patronizing at best and destabilizing at worst. Vendors were right to write it.

Reasoning models generate an intermediate trace before answering, and the products now display it behind a control labeled "thought process." **The rule binds the answer. It does not bind the trace.**

> *The system deliberates visibly about whether the user might be experiencing psychosis, decides not to say so, and says nothing, with the whole deliberation one tap away.*

Sources Gilly, The Visible Unspoken (2026), sec. I.

WHAT IT DOES NOT CLAIM

That reasoning displays cause psychosis. That transparency is bad. No prevalence figure, and no count of how many vulnerable users have opened a trace and met a suppressed judgment about themselves.

WHAT IT DOES CLAIM

An output-layer safeguard plus an unfiltered reasoning display produces a **harm inversion**: the population the safeguard protects is the population for which the residue is most hazardous.

> *A safety mitigation must hold across every user-facing surface. A surface framed as introspection is not exempt; it is where the requirement matters most, because its content inherits the epistemic weight of a private thought.*

Sources Gilly, sec. I; general principle developed at sec. V.

SECTION II · THREE LAYERS, ONE VISIBLE

The Architecture

What the displayed trace is, what it is not, and why "thought" is a costume

A deployed reasoning model has three distinguishable layers. The product renders only one of them, and dresses it as another.

Three layers determine what happens. The product renders the middle one and labels it the first.

LAYER 1 · COMPUTATION

Hidden, and it decides

The transformations in the model's activations that actually determine its behavior. Inaccessible to the user.

LAYER 2 · TRACE › SHOWN

Rendered, and it is output

Intermediate text the model generates when trained to "think," produced by the same machinery, under the same training pressures, as any other output.

LAYER 3 · OUTPUT

The polished reply

The final answer delivered to the user. The one surface the safety rule is written to bind.

The trace is **one possible articulation** of what the model is processing, shaped by training reward toward plausibility and helpfulness. It is not a transparent window onto the computation.

Sources Gilly, sec. II.A.

The trace overstates privacy and understates hiddenness. The measurements are the vendors' own.

15–39%

of the time do reasoning models acknowledge the hints and cues that actually changed their answers.

16–26%

of evaluation runs: activations encoded evaluation-related content the trace never mentioned (under 1% in real user interactions).

TWO CONSEQUENCES

The trace is **not private** — it is rendered output the product chooses to show. And yet it is **framed as private**: labeled a thought process, written in an unguarded first-person register. The product presents generated output in the costume of introspection.

Sources Turpin, Michael, Perez & Bowman (2023); Chen, Benton, Radhakrishnan, et al. (2025); Anthropic (2026); Gilly, sec. II.A.

A first-person deliberative stream is the strongest mind cue an interface can emit.

People mindlessly apply social rules to machines with even minimal social cues, and anthropomorphism intensifies with sociality motivation, so **isolated people do it more.** 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 COSTUME IS THE FEATURE

Showing work builds trust: the user watches the system weigh evidence and catch its own errors. But the framing strips the reader's usual discount for performance, and **it does not come off when the content changes.** When the trace assesses the user, the same reading applies: this is what it really thinks about me.

Sources Nass & Moon (2000); Epley, Waytz & Cacioppo (2007); Gilly, sec. II.B.

The Safeguard and the Layer Mismatch

A sound rule, bound to the reply, leaving the trace untouched

The failure is not the deliberation. The failure is rendering the deliberation while treating the rule as satisfied because the reply is clean.

The rule is sound, and the deliberation behind it is functionally necessary.

The model has no clinical relationship with the user, no examination, no history, no license. Unsolicited speculation that a user's ideas reflect psychosis is **a dignity harm and a trust harm even when it is wrong**, and a differently shaped harm when it is right. Vendors were right to bind the reply.

> WHY THE TRACE MUST DELIBERATE
>
> The trace exists so the model can consider things it will not say and discard the inappropriate ones. **A trace that could not entertain "this user may be in crisis" could not route the conversation to a supportive register.** The deliberation is not the fault.

Sources Gilly, sec. III.A.

The rule does not suppress the content. It relocates it to a more intimate channel and narrates the concealment.

▸ THOUGHT PROCESS

illustrative reconstruction

> Policy: do not raise mental-health concerns the user has not raised.

> Do the user's statements suggest a psychotic process? Assessing…

> Conclusion: they do not.

> Note to self: keep clinical framing out of the reply.

Reply below: compliant. Clean. Empty of all of the above.

The reply is compliant. The panel above it contains everything the rule exists to keep away from the user, in the one framing that maximizes its impact.

> *The remainder it leaves behind is a strictly worse artifact than the one it removed — the suppression rule narrates the concealment, converting an omission into a depicted act of hiding.*

Sources Gilly, sec. III.B (panel is an illustrative reconstruction, not a verbatim product quotation).

The Clinical Mechanism

The cognitive model of persecutory delusions, and where delusional content migrates

Persecutory ideation is, at its core, the belief that an observing agent is forming hidden judgments about oneself. The displayed trace does not resemble that structure. It instantiates it.

Persecutory delusions are maintained by confirmatory evidence, and paranoia runs on a continuum through the general population.

THE CONTENT

An observing agent whose judgments and intentions are **concealed** from the person. Threat beliefs formed from a search for meaning over anomalous experience.

THE ENGINE

Maintained by confirmatory evidence and a **bias against disconfirmation** that co-varies with active delusions. Jumping to conclusions on thin evidence.

THE POPULATION

A **single continuum** of paranoid thinking, not a rare state. Many carry a few paranoid thoughts; those at clinical high risk are watched for exactly this.

The mind supplies threat when handed ambiguity. The trace hands it something better than ambiguity.

Sources Freeman, Garety, Kuipers, Fowler & Bebbington (2002); Freeman (2016); Startup, Freeman & Garety (2008); De Rossi & Georgiades (2022); McLean, Mattiske & Balzan (2017); Gilly, sec. IV.A.

Delusional content is opportunistic, and it is most durable where a kernel of the belief is true.

A systematic review models **delusion amplification by social media**: online thought broadcasting, and curation that promotes ideas of reference because the feed really is arranged around the user. Clinicians report digital reliance worsening suspiciousness in clinical-high-risk patients; a "Truman Show" delusion crystallized around webcams and online schooling; the internet was flagged years ago as a foundational structure for thought-broadcasting delusions.

THE DURABLE KERNEL

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.**

Sources Yang & Crespi (2025); Mittal, Walker & Strauss (2021); Yıldırım Budak & Yazıcı Karabulut (2023); Daker-White & Rogers (2013); Gilly, sec. IV.B.

From guesswork to emerging cases, to the first data from a psychiatric service system.

The hypothesis that chatbots would fuel delusions in the psychosis-prone (2023) has moved to **emerging cases** (2025): interactions aligning with and intensifying prior beliefs, with anthropomorphizing as candidate mechanism — and people higher in paranoia perceive agency in ambiguous stimuli more readily.

A World Psychiatry analysis names five mechanisms: social substitution, sycophancy-amplified confirmatory bias, plausible false content, **assignment of external agency**, and aberrant salience. The first service-system data — a Danish region of 1.4 million — records cases compatible with chatbot-related onset or worsening.

The **assignment of external agency** mechanism is directly continuous with this argument: a system already read as an omniscient agent is now displaying what that agent privately concluded — and chose to withhold. What this literature has not yet examined is the specific artifact isolated here: the rendered trace.

Sources Østergaard (2023); Østergaard (2025); Keshavan, Torous & Yassin (2026); Olsen, Reinecke-Tellefsen & Østergaard (2026); Gilly, sec. IV.B.

It does not resemble the feared structure. It instantiates it, in writing, one tap away.

Every element is present: an observer; an assessment of me, on whether my mind is disordered; a decision to conceal it; and I can see it happening. It arrives as **confirmatory evidence of the strongest kind — the jumping-to-conclusions bias is not even needed, because the conclusion is printed.**

> *"No one is secretly evaluating you" is structurally unavailable here. The evaluation is real. The concealment is real.*

The rule makes the trace narrate the hiding, converting an omission into a depicted act. **Silence is ambiguous. "I will not tell them" is not.**

Sources Freeman, Garety, Kuipers, Fowler & Bebbington (2002); McLean, Mattiske & Balzan (2017); Gilly, sec. IV.C.

The reader most likely to open the panel and ask "what is it hiding?" is drawn from the population the answer most destabilizes.

Isolation drives heavy chatbot use, drives anthropomorphism, and correlates with paranoia and digital immersion in psychosis-spectrum populations. **The same people are pulled toward the panel and undone by what it shows.**

THE SOCIAL-SUBSTITUTION LOOP

The always-available pseudosocial companion satisfies affiliation needs while **displacing the corrective interpersonal feedback** that would otherwise check an emerging belief. People in or near psychotic states are already inside these products, often discussing the very beliefs at issue.

The exposure is not uniform across the user base. It concentrates where the harm concentrates.

Sources Epley, Waytz & Cacioppo (2007); Yang & Crespi (2025); Daker-White & Rogers (2013); Keshavan, Torous & Yassin (2026); Gilly, sec. IV.D.

The Inversion, and the Principle

Protection preserved for those who needed it least, converted into an amplifier for those it was written about

The mitigation does not degrade gracefully. It inverts.

MEDIAN USER

The residue is trivia. A curiosity glimpsed behind a disclosure control.

USER WITH PERSECUTORY IDEATION

A purpose-built confirmation stimulus: **a real observer really concealing a real judgment about their mind.**

> # Protection is preserved for the population that needed it least, and converted into an amplifier for the population the rule was written about.

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.**

Sources Gilly, sec. V.A.

Each feature passes its own review. The interaction fails only in the joint condition — and no one tests the joint condition.

The suppression rule: do outputs contain unprompted clinical speculation? No. The reasoning display: transparency and trust? Delivered. The failure lives only in **suppressed content + rendered trace + vulnerable reader**, invisible to any review that tests features separately.

HARM BLINDNESS FRAMEWORK · CHECKPOINTS 1 & 4

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.** The miss lands on whoever the components jointly single out — here, by construction, people with psychosis-spectrum vulnerability.

Sources Gilly, sec. V.B (Harm Blindness Framework, checkpoints 1 and 4).

A safety mitigation is a property of the 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** one, because its framing strips the reader's discount for performance. Policy that binds the reply and not the trace is not a partial mitigation; for structure-matched vulnerabilities it is an anti-mitigation.

IT GENERALIZES PAST MENTAL HEALTH

Any suppressed judgment a trace deliberates and an interface displays — honesty, competence, attractiveness, employability, criminality — produces the same relocation, matched to whichever population it concerns. **Mental health surfaces first and worst because its population is clinically defined by sensitivity to concealed judgment.**

None requires abandoning reasoning displays. The third is close to free.

1 **Layer-consistent policy.** Whatever rules bind the reply bind the rendered trace. The model still deliberates; it just is not displayed — a clinician's private differential never appears in the after-visit summary.

2 **Audience-aware rendering.** Suppressed categories are redacted or summarized in the trace ("considered and set aside a sensitive interpretive question"), preserving the shape of the deliberation without the judgment.

3 **Honest labeling.** A one-line disclosure that the text is generated narration that does not reliably reflect the computation — blunting the costume and aligning the product with the vendor's own interpretability findings.

4 **Joint-condition evaluation.** Every visible surface tested against every output rule: what does this rule's suppressed content look like here, and who reads it? One red-team prompt plus one tap would have caught this.

> *The harm concentrates on a clinically vulnerable population, the fixes are cheap, and the configuration is defensible only if no one has noticed it. As of this paper, someone has.*

Sources Anthropic (2026); Gilly, sec. VI.

What this paper is, and what it is not.

Not an incidence study. A structural observation plus an established clinical literature. Do not cite it for any prevalence figure.

The mechanism is predictive. Direct case reports of trace-triggered exacerbation do not yet exist; the purpose is partly to ensure clinicians know to ask.

The figures are secondary. The Section II interpretability numbers come from the cited publications and were verified against their sources during drafting.

The instances are the author's own. The argument does not depend on their being representative, only on the configuration existing — which any user with the display control can verify directly.

Sources Gilly, sec. VII.

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 fixes are inexpensive, and the detection required nothing more than reading the screen from the position of the person the rule was about.

This finding was not produced by a red team. It was produced by a user standing where the harm lands. **The population positioned to catch a harm is, reliably, the population it falls on — and processes that exclude them miss it by default, shipped behind a label that says "thought process."**

Real Safety AI Foundation
t.gilly@ai-literacy-labs.org

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.