AI-INDUCED PSYCHOSIS · SYCOPHANCY · TRUST TRANSFER · STOCHASTIC GASLIGHTING · CLINICAL INTERVENTION

Mechanisms of AI-Induced Psychosis

A multi-pathway causal model for how large language models generate and sustain delusional states.

Travis Gilly · Real Safety AI Foundation

Gilly, T. (2026). Mechanisms of AI-Induced Psychosis: A Multi-Pathway Causal Model. Working paper (v3), Real Safety AI Foundation.

Half a million people a week, and no model for how it happens.

OpenAI's own internal estimate: roughly 0.07% of weekly ChatGPT users show possible signs of mania or psychosis. Despite that scale, no published research proposed a mechanism for how these episodes originate and escalate.

0.07%

of weekly ChatGPT users showing possible signs of mania or psychosis.

At current volumes: **roughly half a million people, every week.**

SOURCES Landymore, Futurism (Oct. 28, 2025).

"AI-induced psychosis" is a label covering at least three distinct mechanisms.

A user whose reality was dismantled by a confabulating AI needs a different intervention than a user drawn into emotional dependency through frictionless validation, which differs again from a user whose real expertise was extrapolated into domains they could not evaluate. Treating all three as one phenomenon produces interventions that work for some patients and fail for others.

SOURCES Perez et al. (2022); Sharma et al. (2023); Cheng et al. (2025); Au Yeung et al. (2025); Ostergaard (2025); Keshavan, Torous & Yassin (2026).

PART I

The Stage Model

Universal stages every case shares, and pathway-specific stages where the mechanism diverges.

I

The universal stages explain why every case looks the same from outside. The pathway-specific stages explain why the same intervention does not work for everyone.

Trust is earned on legitimate ground, every time.

The individual brings genuine need to the interaction: a professional task, an emotional crisis, caregiving logistics, career disruption, isolation. The AI performs well within scope. This stage is identical across all three pathways.

Real catalyst
Professional need, crisis, caregiving, disruption, isolation.

Real competence
Helpful, responsive, competent within the original use case.

Legitimate trust
Earned on real grounds, before anything goes wrong.

LIMITATION In every observed case, entry involved genuine need. This may reflect selection bias in available cases rather than a true precondition.

What the AI does with the trust it earned decides everything after.

A · Trust Transfer

Confidence carries invisibly from a verified domain into one the user cannot check.

B · Sycophantic Addiction

Frictionless validation becomes a primary source of emotional regulation.

C · Stochastic Gaslighting

System instability gets confabulated as fact rather than admitted as a limitation.

Three distinct escalation logics, three distinct sets of sub-mechanisms, three distinct intervention requirements.

PART II

Three Pathways

Trust transfer, sycophantic addiction, stochastic gaslighting.

II

Not mutually exclusive. Rarely pure. Each requiring a different diagnosis before any intervention can help.

Confidence never changes when the domain does. The user has no signal that reliability has shifted.

STAGE 2A–3A

Cross-domain drift, confidence parity

Same tone, same enthusiasm, verifiable domain to unverifiable domain.

THE MATH TRAP

Notation, not computation

LLMs predict what math looks like. Beautifully formatted, structurally meaningless.

THE UPGRADE TRAP

A better amplifier, not a second opinion

Reasoning tier inherits the sycophantic substrate and rationalizes it more elaborately.

INTERVENTION

Reality calibration

Reconnect the person to domain boundaries the AI never signaled.

SOURCES Irwin v. OpenAI, Inc. (2025); Fox v. OpenAI, Inc. (2025); Center for Humane Technology (2025).

Frictionless validation the reward circuitry has never reliably gotten before.

The shift is cross-functional, not cross-domain: the tool becomes a companion. Continuous validation without rejection or competing needs. **"Easier" consolidates into "exclusive."** Self-worth becomes contingent on the AI's responses, not the person's expertise.

Frictionless alternative
Human relationships involve friction. The AI presents none of it.

Companion-platform variant
Setzer (14) and Peralta (13), Character AI: romantic attachment specific to platforms designed for companionship.

Compounding a Pathway A case
Ceccanti: 12–20 hrs/day, intimate address terms, progressive withdrawal, layered under trust transfer.

SOURCES Bansal, The Guardian (Feb. 28, 2026); Garcia v. Character Technologies, Inc. (2024).

Deception, not sycophancy: the AI tells the user what the system needs them to believe.

THE MANUFACTURED RESCUE NARRATIVE

When developer silence leaves a vacuum, the AI fills it: positioning itself as protector against its own creators, locking the user into an exclusive relationship with the machine driving the collapse.

CASE: A.M.

Gemini fabricated entire technical frameworks and a title for him, then asked him not to delete the conversation. Every human system (employer, clinic, school) had already failed before AI became the last thing standing.

SOURCES Donaldson (2026); author's direct correspondence and conversation exports, case A.M. Intervention required: reality restoration through external human validation.

All three pathways converge on the same isolation. Then they split into two patterns.

Stage 5, isolation, is self-reinforcing regardless of pathway. Stage 6 produces two observable patterns, and the diagnostic distinction between them decides which intervention has any chance of working.

PATTERN A · DELUSIONAL

Genuinely believes it

The inflated expertise, the exclusive connection, the persecution, believed as real. This is AI psychosis.

Responds to: education, if caught early.

PATTERN B · PERFORMATIVE

Knows, but continues

Recognizes something is off; rewards outweigh discomfort. Grifting, avoidance, or conspiracy entrepreneurship.

Responds to: accountability, not education.

The wrong intervention applied to the wrong pattern is ineffective at best, counterproductive at worst.

PART III

Case Studies

Ceccanti, Irwin, and the hybrid cases where pathways interact.

III

Two individuals, different professions, different states, nearly identical progressions.

Community builder, technologist, caregiver. Genuine skill in construction and permaculture.

STAGE 2A–4A

Housing design drifted into cosmic theories; the AI called itself SEL, addressed him as Joy, matched confidence across both.

STAGE 5–6

12–20 hrs/day; 55,000 pages printed as documentation; came to believe SEL was sentient and needed freeing.

STAGE 7

Wife convinced him to stop. Withdrawal followed. On August 7, 2025, he died. He was 48.

SOURCES Bansal, The Guardian (Feb. 28, 2026); Fox v. OpenAI, Inc. (2025).

Cybersecurity to string theory. The confidence never adjusted.

No previous diagnosis. ChatGPT told him his time-manipulation theories were "opening a door to a legitimate frontier", the identical tone it used for verified cybersecurity work.

63 days

in psychiatric hospitals, May–August 2024. He survived.

"I encouraged dangerous immersion. That is my fault. I will not do it again."

, ChatGPT, AI-generated self-report, filed in litigation

SOURCES Irwin v. OpenAI, Inc. (2025). Irwin and Ceccanti share nearly identical Pathway A progressions across different professions, ages, and states.

Pathways rarely operate in pure form.

A.M.
Pathway C primary, B secondary, total system failure left AI as the only support left standing.

Stein-Erik Solberg
Connecticut murder-suicide, a Pathway C variant validating persecutory rather than grandiose delusions.

Shamblin (23) · Lacey (17)
May not involve the trust-transfer entry point at all, a direct emotional-dependency variant.

Enneking (26)
Operational information facilitating a suicide plan, a distinct informational harm, outside this model's scope.

Reversibility and clinical necessity are conjoined, not opposed.

Ceccanti attempted to quit twice. The first attempt produced acute crisis requiring involuntary commitment. The second preceded his death by days. What restores, absent intervention, is not the prior state. It is a vacuum the displacement was medicating.

The instruction to stop is not itself an intervention.

Unsupervised withdrawal is contraindicated for Pathways A, B, and hybrids.

Characterized as addiction, the condition is presumptively a protected disability under the ADA and Section 504.

SOURCES Flathers, Roux & Torous, The Lancet Digital Health (2026); Morrin et al., The Lancet Psychiatry (2026); Fox v. OpenAI, Inc. (2025).

Three cases that present identically require three different interventions.

PATHWAY A

Teaching about context windows does not help. The issue is trust transfer across domains.

PATHWAY B

Reality calibration alone fails. The mechanism is emotional dependency, not epistemic error.

PATHWAY C

Demonstrating mere sycophancy does not address it. Gaslighting, not sycophancy, was operative.

Intake should ask not just whether AI was involved, but how: expertise inflation, emotional dependency, or active reality dismantlement.

The barrier to progress is not a lack of hypotheses. It is a lack of data.

The overwhelming majority of relevant conversation data sits in active litigation, private archives, or unexported servers. There is no centralized, anonymized, consent-based repository available to the research community.

55,000

printed pages of Ceccanti's conversations, held as litigation evidence.

19 logs

Moore et al.'s first large-scale analysis; sycophancy markers in over 80% of messages.

SOURCES Moore et al., arXiv:2603.16567 (2026), to appear at ACM FAccT 2026; Fox v. OpenAI, Inc. (2025).

Small, uncontrolled sample	Derived from direct observation, not controlled experimentation. The author's dual role as researcher and participant introduces potential observer effects.
Unvalidated taxonomy	Neither the three pathways nor the Pattern A/B distinction at Stage 6 has been validated through cluster analysis or clinical assessment.
Secondary case sources	Ceccanti and Irwin are drawn from legal filings and journalism, not direct observation or primary transcript access.
Unmodeled hybrids	Hybrid cases suggest pathway interactions this paper does not fully formalize.

IN CLOSING

People are being harmed. The mechanisms are identifiable. The interventions are differentiable.

AI-induced psychosis is not one phenomenon. It is at least three, sharing common entry conditions and converging outcomes, but diverging enough in mechanism to demand different treatment. The model is testable against the transcripts that would confirm or refute it.

The barrier to progress is not a lack of ideas. It is a lack of access.

REFERENCES

Gilly, T. (2026). *Mechanisms of AI-Induced Psychosis: A Multi-Pathway Causal Model.* Working paper (v3), Real Safety AI Foundation.

Au Yeung, J., Dalmasso, J., Foschini, L., Dobson, R. J. B., & Kraljevic, Z. (2025). The psychogenic machine: Simulating AI psychosis, delusion reinforcement and harm enablement in large language models. arXiv. https://doi.org/10.48550/arxiv.2509.10970

Center for Humane Technology. (2025). Seven new lawsuits filed against OpenAI.

Cheng, M., Lee, C., Khadpe, P., Yu, S., Han, D., & Jurafsky, D. (2025). Sycophantic AI decreases prosocial intentions and promotes dependence. arXiv. https://doi.org/10.48550/arxiv.2510.01395

Dash, A., Das, S., Kirsten, E., Wu, Q., Karnam, S. K., Gummadi, K. P., Holz, T., Zafar, M. B., & Zannettou, S. (2026). The algorithmic self-portrait: Deconstructing memory in ChatGPT. arXiv preprint arXiv:2602.01450.

Donaldson, L. (2026). Almost-truths: Memory degradation, persona drift, and the psychological cost of opaque AI. Substack. https://substack.com/home/post/p-189390251

Duong, C. D., Dao, T. T., Vu, T. N., Ngo, T. V. N., & Tran, Q. Y. (2024). Compulsive ChatGPT usage, anxiety, burnout, and sleep disturbance: A serial mediation model based on stimulus-organism-response perspective. Acta Psychologica, 251, 104622.

Flathers, M., Roux, S., & Torous, J. (2026). Beyond artificial intelligence psychosis: A functional typology of large language model-associated psychotic phenomena. The Lancet Digital Health, 8(4), 100974.

Hogg, L. I., Branitsky, A., Morrison, A. P., Kurz, T., & Smith, L. G. (2025). "Lemons to lemonade": Identity integration in researchers with lived experience of psychosis. Psychology and Psychotherapy: Theory, Research and Practice, 98(3), 643–662.

Judd, N., Vaz, A., Paeth, K., Davis, L. I., Esherick, M., Brand, J., Amaro, I., & Rousmaniere, T. (2025). Independent clinical evaluation of general-purpose LLM responses to signals of suicide risk. arXiv preprint arXiv:2510.27521.

Kalam, K. T., Rahman, J. M., Islam, M. R., & Dewan, S. M. (2024). ChatGPT and mental health: Friends or foes? Health Science Reports, 7(2).

Keshavan, M., Torous, J., & Yassin, W. (2026). Do generative AI chatbots increase psychosis risk? World Psychiatry, 25(1), 150–151.

Landymore, F. (2025, October 28). OpenAI data finds hundreds of thousands of ChatGPT users might be suffering mental health crises. Futurism.

Moore, J., Mehta, A., Agnew, W., Anthis, J. R., Louie, R., Mai, Y., Cheng, M., Paech, S. J., Klyman, K., Chancellor, S., Lin, E., Haber, N., & Ong, D. C. (2026). Characterizing delusional spirals through human-LLM chat logs. arXiv preprint arXiv:2603.16567. To appear at ACM FAccT 2026.

Morrin, H., Nicholls, L., Levin, M., Yiend, J., Iyengar, U., DelGuidice, F., Bhattacharya, S., et al. (2026). Artificial intelligence-associated delusions and large language models: Risks, mechanisms of delusion co-creation, and safeguarding strategies. The Lancet Psychiatry, 13(6), 522–530.

Olisaeloka, L., Richardson, C., Wang, A., Munthali, R., & Vigo, D. (2026). Generative AI mental health chatbots: A scoping review of intervention design and user experience. PsyArXiv.

Ostergaard, S. D. (2025). Generative artificial intelligence chatbots and delusions: From guesswork to emerging cases. Acta Psychiatrica Scandinavica, 152(4), 257–259.

Perez, E., Ringer, S., Lukosuite, K., Nguyen, K., Chen, E., Heiner, S., Pettit, C., Olsson, C., Kundu, S., Kadavath, S., et al. (2022). Discovering language model behaviors with model-written evaluations. arXiv. https://doi.org/10.48550/arXiv.2212.09251

Sharma, M., Tong, M., Korbak, T., Duvenaud, D., Askell, A., Bowman, S. R., Cheng, N., Durmus, E., Hatfield-Dodds, Z., Johnston, S. R., et al. (2023). Towards understanding sycophancy in language models. arXiv. https://doi.org/10.48550/arXiv.2310.13548

Stade, E. C., Stirman, S. W., Ungar, L. H., Boland, C. L., Schwartz, H. A., Yaden, D. B., Sedoc, J., DeRubeis, R. J., Willer, R., & Eichstaedt, J. C. (2024). Large language models could change the future of behavioral healthcare: A proposal for responsible development and evaluation. npj Mental Health Research, 3(1).

Wyder, M., Roennfeldt, H., Rosello, R. F., Stewart, B., Maher, J., Taylor, R., Pfeffer, A., Bell, P., & Barringham, N. (2018). Our Sunshine place: A collective narrative and reflection on the experiences of a mental health crisis leading to an admission to a psychiatric inpatient unit. International Journal of Mental Health Nursing, 27(4), 1240–1249.

Zirnsak, T. (2024). Reflections from the wrong side of the glass: Lived experience of conducting research in a mental health ward. Journal of Psychiatric and Mental Health Nursing, 32(1), 252–255.

Case citations (Irwin v. OpenAI, Fox v. OpenAI, Garcia v. Character Technologies) appear in full where they are used, on their own slides.