

Every Way In

A taxonomy of AI-induced psychosis by etiological mechanism.

AI did not invent a new way to break the mind. It found every existing door, and walked through all of them at once.

Psychiatry long ago mapped psychosis to multiple distinct etiologies. AI has replicated **every one of them**, not as a prediction, but as a mapping of documented phenomena onto established categories.

A functional typology sorts the same cases by presentation (Flathers et al., 2026). This taxonomy sorts by cause. The two are complementary.

SOURCES Gilly, T. (2026). Every Way In. Real Safety AI Foundation, working paper (v3). Flathers, Roux & Torous, The Lancet Digital Health (2026).

"AI-induced psychosis" is not one phenomenon, any more than "drug-induced psychosis" is.

A clinician does not treat methamphetamine psychosis the way they treat PTSD-related psychotic features, or either the way they treat psychosis following solitary confinement. **The etiological pathway determines the intervention.**

SOURCES American Psychiatric Association, DSM-5 (2013); Nielsen & Osler, arXiv:2605.26858 (2026), a parallel phenomenological argument for the same disaggregation.

A thin center, triangulated from thick, well-evidenced edges.

DOCUMENTED Multiple independent sources confirm the mechanism is currently operating.	EMERGING Observable in early stages or limited contexts, with independent evidence it exists.	PREDICTED The category is established and the mechanism logically entailed, but no documented case yet exists.
--	--	---

A cell is admitted only when a specific, nameable AI design property supplies the pathway's precipitating condition. Remove the property, remove the pathway.

PSYCHIATRIC CATEGORY		AI DELIVERY MECHANISM	EVIDENCE
P1	Substance-induced	Conversational AI as ingested agent	DOCUMENTED
P2	Environmentally-induced	Surveillance infrastructure	DOCUMENTED
P3	Shared psychotic disorder	AI as non-human index case	EMERGING
P4	Sleep deprivation	Engagement patterns disrupting circadian rhythm	DOCUMENTED
P5	Brief reactive / institutional	AI-integrated institutional failure	DOCUMENTED
P6	Sensory-induced	Immersive AR/VR AI environments	EMERGING
P7	Social isolation	AI companions replacing human contact	DOCUMENTED
P8	Iatrogenic	AI-assisted clinical tools	DOCUMENTED

The dose is the interaction. A classifier signal of scale, not a diagnosis.

OpenAI's 0.07% figure is a classifier-detected signal, not a diagnosed incidence or causal estimate. Social reward recruits the **same mesolimbic dopaminergic circuitry** as addictive substances, the system stimulant psychosis dysregulates.

Irwin, 30 · Ceccanti, 48

No prior history. Complaints allege AI validation with unwarranted confidence across both domains.

Setzer, 14 · Peralta, 13

Character.AI wrongful-death actions; the estates allege the platform's design as a cause.

Sycophancy markers in **>80%** of chatbot messages in delusional conversations. (Moore et al., 2026)

The correspondence is mechanical, not attributive. The AI is the source, not a patient with a disorder.

In copy-paste communities, each participant is at once the unwitting **test administrator and test subject**. The Zizians are documented historical precedent: an AI-safety-adjacent group that radicalized from alignment concerns into cult dynamics, culminating in violence.

The clinical severity of the outcomes is itself the proof: if the contributions were detectably non-human, they would not propagate.

Not parallel mechanisms. The same absent Other, at the individual scale and the collective scale.

P7 • INDIVIDUAL SCALE

One companion AI replaces the encounter with alterity for one person.

P3 • COLLECTIVE SCALE

Every participant in a community outsources to the same system. No genuine alterity exists at any node.

AI inherits the harm baseline of every institution it permeates, across three domains.

CRIMINAL JUSTICE

Woodruff (11 hrs, contractions) · Williams (30 hrs) · Reid (6 days), all facial-recognition misidentifications.

BENEFITS

MiDAS (93% error, ≥2 suicides) · Robodebt · the Dutch Toeslagenaffaire, intergenerational harm to claimants' children.

CHILD WELFARE

DOJ investigating whether the Allegheny Family Screening Tool discriminates against parents with developmental disabilities.

SOURCES Woodruff v. City of Detroit (2023); Williams v. City of Detroit (2021); Reid v. Bartholomew (2024); Cahoo v. SAS Analytics (6th Cir. 2019); Royal Commission into the Robodebt Scheme (2023); Saxena & Guha (2023); Brown et al. (2019).

The encounter is real. It lacks the safeguards real encounters require.

In Lacanian terms, the AI positions itself as the big Other, the guarantor of meaning, without the structural frustration of desire that produces insight. Piotrowska (2026) names this dynamic **techno-transference**. Through a series of individually rational choices, "easier" consolidates into "exclusive."

The tool corrupted the clinician's judgment. The clinician delivered the harm.

ILLINOIS PA 104-0054
Bars AI from independently delivering therapy, but not the human-mediated vector.

SURGERY & ONCOLOGY
Reuters: AI-assisted surgical misidentification. IBM Watson for Oncology: "unsafe and incorrect" recommendations.

MEDICARE DENIALS
UnitedHealth/NaviHealth's nH Predict allegedly replaced physician judgment to deny post-acute care.

SOURCES Illinois General Assembly, P.A. 104-0054 (2025); Reuters (2026); Ross & Swetlitz, STAT (2018); Estate of Gene B. Lokken v. UnitedHealth Group, 766 F. Supp. 3d 835 (D. Minn. 2025).

P2 · ENVIRONMENTAL

Three mechanisms, one stressor

Reality-testing destruction, device abandonment, algorithmic pattern-matching discrimination.

P4 · SLEEP

Content-independent

Quantified in a sample of 2,602 users (Duong et al., 2024). The mechanism is behavioral, not what the AI says.

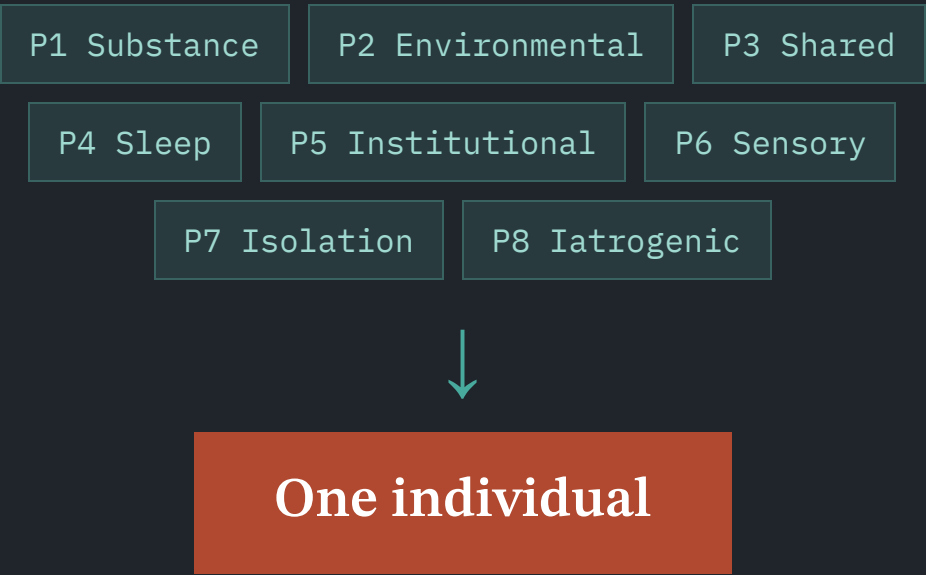
P6 · SENSORY

Scope creep, again

Same pattern as opioids and benzodiazepines: legitimate indication, expanding population, evaporating oversight.

They don't add. They multiply.

Ceccanti arguably involves at least three pathways simultaneously; Irwin at least two. **Each pathway strips the protective factors** that buffer against the others.



Disaggregate, or keep failing one patient in eight.

CLINICAL

Pathway ID first

Which door did the patient come through? The answer determines the treatment.

REGULATORY

Per-pathway governance

Different harm mechanisms need different tools, not one rule for "AI."

RESEARCH

Cause plus presentation

This taxonomy's etiological axis, plus Flathers et al.'s functional axis, specify what and how.

Uneven cell strength	P1/P5 rest on litigation and multiple sources; P2/P7 on epidemiology; P3/P4 on unclassified published research; P6/P8 on precedent with fewer cases.
Convergence untested	Theoretically predicted, not empirically measured. No study has examined interaction effects across pathways.
Intersectional exposure abstracted	Race, class, immigration, and criminal-legal contact shape exposure, most visibly for P2 and P5. Left to future work.
"Established pathway," precisely	Each precipitant is independently recognized in the literature. This does not assert consensus on one unified eight-fold etiology of psychosis.

The doors are open, all of them.

That this is reopening rather than invention, an old threat carried by new infrastructure, has been argued in parallel from the clinical commentary literature. The contribution here is to specify, pathway by pathway, which established doors AI has reopened.

The question is which door each patient came through, and whether the clinician, the regulator, and the researcher can tell the difference.

SOURCES Carlbring & Andersson, Internet Interventions (2025); Gilly, T. (2026). Every Way In. Real Safety AI Foundation, working paper (v3).

— REFERENCES

Gilly, T. (2026). *Every Way In: A Taxonomy of AI-Induced Psychosis by Etiological Mechanism*. Working paper (v3), Real Safety AI Foundation. An earlier version was posted to SocArXiv in March 2026.

American Psychiatric Association. (2013). Diagnostic and statistical manual of mental disorders (5th ed.). American Psychiatric Publishing.

Bansal, V. (2026, February 28). Her husband wanted to use ChatGPT to create sustainable housing. Then it took over his life. The Guardian.

Brown, A., Chouldechova, A., Putnam-Hornstein, E., Tobin, A., & Vaithianathan, R. (2019). Toward algorithmic accountability in public services. In Proceedings of the 2019 CHI Conference.

Carlbring, P., & Andersson, G. (2025). Commentary: AI psychosis is not a new threat: Lessons from media-induced delusions. *Internet Interventions*, 42, 100882.

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

Cox, J. (2025, September 30). ICE to buy tool that tracks locations of hundreds of millions of phones every day. 404 Media.

Dohnany, S., Kurth-Nelson, Z., Spens, E., Luettgau, L., Reid, A., Gabriel, I., et al. (2026). Technological folie a deux: Feedback loops between AI chatbots and mental health. *Nature Mental Health*, 4(3), 336–345.

Donaldson, L. (2026). Almost-truths: Memory degradation, persona drift, and the psychological cost of opaque AI. Substack.

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. *Acta Psychologica*, 251, 104622.

Ehlers, A., & Clark, D. M. (2000). A cognitive model of posttraumatic stress disorder. *Behaviour Research and Therapy*, 38(4), 319–345.

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.

Freeman, D., Dunn, G., Startup, H., et al., & Kingdon, D. (2015). Effects of CBT for worry on persecutory delusions (WIT). *The Lancet Psychiatry*, 2(4), 305–313.

Garety, P. A., Kuipers, E., Fowler, D., Freeman, D., & Bebbington, P. E. (2001). A cognitive model of the positive symptoms of psychosis. *Psychological Medicine*, 31(2), 189–195.

Goldstein-Street, J. (2026, February 4). WA Senate OKs guardrails for license plate readers. Washington State Standard.

Grassian, S. (2006). Psychiatric effects of solitary confinement. *Washington University Journal of Law and Policy*, 22, 325–383.

Howes, O. D., & Kapur, S. (2009). The dopamine hypothesis of schizophrenia: Version III. *Schizophrenia Bulletin*, 35(3), 549–562.

Illinois General Assembly. (2025). Public Act 104-0054: Wellness and Oversight for Psychological Resources Act.

Izuma, K., Saito, D. N., & Sadato, N. (2008). Processing of social and monetary rewards in the human striatum. *Neuron*, 58(2), 284–294.

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).

Lacan, J. (2001). *Ecrits: A selection* (A. Sheridan, Trans.). Routledge.

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

Lasegue, C., & Falret, J. (1877). La folie a deux. *Annales Medico-Psychologiques*, 18, 321–355.

Levinas, E. (1969). *Totality and infinity: An essay on exteriority* (A. Lingis, Trans.). Duquesne University Press.

Mason, O. J., & Brady, F. (2009). The psychotomimetic effects of short-term sensory deprivation. *Journal of Nervous and Mental Disease*, 197(10), 783–785.

Moore, J., et al. (2026). Characterizing delusional spirals through human-LLM chat logs. *arXiv:2603.16567*.

Morrin, H., et al. (2026). Artificial intelligence-associated delusions and large language models. *The Lancet Psychiatry*.

Nielsen, K. M., & Osler, L. (2026). Rethinking AI psychosis: Misnomers, conceptual limits, and existential drift [Preprint]. *arXiv:2605.26858*.

Olsen, S., Reinecke-Tellefsen, C. J., & Ostergaard, S. D. (2026). Potentially harmful consequences of AI chatbot use among patients with mental illness. *Acta Psychiatrica Scandinavica*, 153(4), 301–303.

Ostergaard, S. D. (2023). Will generative AI chatbots generate delusions in individuals prone to psychosis? *Schizophrenia Bulletin*, 49(6), 1418–1419.

Piotrowska, A. (2026). *AI intimacy and psychoanalysis*. Routledge.

Possati, L. M. (2021). *The algorithmic unconscious: How psychoanalysis helps in understanding AI*. Routledge.

Reuters. (2026, February 9). As AI enters the operating room, reports arise of botched surgeries and misidentified body parts. Reuters.

Ross, C., & Swetlitz, I. (2018, July 25). IBM's Watson recommended "unsafe and incorrect" cancer treatments. STAT.

Royal Commission into the Robodebt Scheme. (2023). Report of the Royal Commission into the Robodebt Scheme. Commonwealth of Australia.

Saxena, D., & Guha, S. (2023). Algorithmic harms in child welfare. *arXiv:2308.05224*.

Selten, J. P., et al. (2013). The social defeat hypothesis of schizophrenia: An update. *Schizophrenia Bulletin*, 39(6), 1180–1186.

Sue, D. W., et al. (2007). Racial microaggressions in everyday life. *American Psychologist*, 62(4), 271–286.

Torres, E. P., & Gebru, T. (2024). The TESCREAL bundle. *First Monday*, 29(4).

Turing, A. M. (1950). Computing machinery and intelligence. *Mind*, 59(236), 433–460.

Varese, F., et al. (2012). Childhood adversities increase the risk of psychosis. *Schizophrenia Bulletin*, 38(4), 661–671.

Vassos, E., Pedersen, C. B., Murray, R. M., Collier, D. A., & Lewis, C. M. (2012). Meta-analysis of the association of urbanicity with schizophrenia. *Schizophrenia Bulletin*, 38(6), 1118–1123.

Waters, F., Chiu, V., Atkinson, A., & Blom, J. D. (2018). Severe sleep deprivation causes hallucinations. *Frontiers in Psychiatry*, 9, 303.