Every Way In:

A Taxonomy of AI-Induced Psychosis by Etiological Mechanism

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

Working paper, June 2026 (v3)

ABSTRACT

Psychiatry has established multiple distinct etiological pathways to psychosis, each with its own causal mechanism, clinical presentation, and intervention protocol. Substance-induced psychosis follows ingestion of a specific agent. Environmentally-induced psychosis follows sustained exposure to chronic stressors. Shared psychotic disorder follows close contact with a delusional index case. Sleep deprivation psychosis follows disruption of circadian regulation. Brief reactive psychosis follows acute traumatic events. Sensory-induced psychosis follows sustained perceptual disruption. Social isolation psychosis follows prolonged deprivation of human contact. Iatrogenic psychosis follows medical or therapeutic intervention gone wrong. This paper argues that artificial intelligence systems have replicated, are currently replicating, or are positioned to replicate every one of these established pathways. Conversational AI systems function as the ingested substance. Surveillance infrastructure functions as the environmental stressor. AI generated delusional content propagates through online communities following the shared psychosis model. AI engagement patterns produce the sleep deprivation. AI integrated institutional failures produce the acute trauma. Augmented and virtual reality AI systems produce the sensory disruption. AI companions produce the social isolation wearing the mask of connection. AI assisted clinical tools produce the iatrogenic harm. The taxonomy presented here is not a prediction about future risk. It is a mapping of current and documented phenomena onto established psychiatric categories, indicating that AI has not introduced a single novel pathway to psychosis. It has instead found every existing door and walked through all of them simultaneously. The clinical, regulatory, and research implications of this convergence are significant: a phenomenon currently studied as a single undifferentiated category ("AI-induced psychosis") is in fact at least eight distinct phenomena requiring different interventions, different regulatory responses, and different research methodologies.

Keywords: AI-induced psychosis, etiological taxonomy, psychiatric nosology, conversational AI, surveillance harm, social isolation, iatrogenic harm, correspondence analysis, mental health regulation

I. INTRODUCTION

The emerging literature on AI associated psychological harm has coalesced around a single framing: a person interacts with an AI system, something goes wrong in that interaction, and the person develops psychological symptoms. The

Table 1: Complete Taxonomy of AI-Induced Psychosis by Etiological Mechanism

PATHWAY	PSYCHIATRIC CATEGORY	AI DELIVERY MECHANISM	EVIDENCE	INTERVENTION
P1	Substance-induced	Conversational AI interaction	Documented	Discontinue; educate on mechanism
P2	Environmentally-induced	Surveillance infrastructure	Documented	PTSD informed; modify environment
P3	Shared psychotic disorder	AI as nonhuman index case	Emerging	Transparency mechanisms; AI literacy
P4	Sleep deprivation	AI engagement patterns	Documented	Usage limits; circadian aware design
P5	Brief reactive / institutional	AI integrated institutional failure	Documented	Algorithmic audit; impact assessments
P6	Sensory-induced	Immersive AR/VR AI environments	Emerging	Regulatory containment; off ramp protocols
P7	Social isolation	AI companions replacing human contact	Documented	Friction introduction; relational design
P8	Iatrogenic	AI assisted clinical tools	Documented	Human review; environmental awareness training

research program organized around this framing has produced valuable work. The hypothesis that conversational AI might generate delusions in individuals prone to psychosis was first advanced by Ostergaard (2023). Morrin and colleagues documented mechanisms of delusion co-creation in The Lancet Psychiatry (Morrin et al., 2026). Moore and colleagues characterized delusional spirals in chat logs (Moore et al., 2026). Keshavan, Torous, and Yassin asked whether chatbots increase psychosis risk in World Psychiatry (Keshavan et al., 2026). Most recently, Flathers, Roux, and Torous proposed a functional typology of large language model associated psychotic phenomena in The Lancet Digital Health, sorting cases by how they clinically present (Flathers et al., 2026); the taxonomy advanced here is complementary but orthogonal, sorting by etiology rather than presentation. Legal proceedings have multiplied. OpenAI's internal data suggests roughly half a million users per week may show signs of mania or psychosis (Landymore, 2025), and early data from a large psychiatric service system has begun to move the evidence base beyond media report toward the clinical record (Olsen et al., 2026).

This paper argues that the field's framing, while productive, is dangerously incomplete. "AI-induced psychosis" is not a single phenomenon any more than "drug-induced psychosis" is a single phenomenon.[1] Psychiatry long ago recognized that psychosis has multiple etiological pathways, each with distinct mechanisms, presentations, and treatment requirements. A clinician does not treat methamphetamine psychosis the same way they treat PTSD related psychotic features, and does not treat either the same way they treat psychosis following prolonged solitary confinement. The etiological pathway determines the intervention.

AI has replicated every established etiological pathway to psychosis. Not one. Not two. All of them. And the failure to recognize this, the continued treatment of "AI-induced psychosis" as a single category, produces clinical interventions that work for some patients and fail catastrophically for others, regulatory responses that address one

[1] A contemporaneous and independent line of work reaches a parallel skeptical conclusion by a different route. Nielsen and Osler (2026) argue on phenomenological grounds that "AI psychosis" is not a novel diagnostic category and that the term collapses distinct conditions that should be kept separate. The taxonomy presented here, first posted in March 2026 (see Acknowledgments), converges on the same disaggregation claim from an etiological rather than phenomenological direction.

mechanism while ignoring seven, and research designs that conflate distinct causal chains into an undifferentiated mass.

This paper presents the complete mapping.

II. Method: The Etiological Correspondence Principle

The taxonomy presented here was constructed through a systematic correspondence analysis. For each established etiological pathway to psychosis recognized in the psychiatric literature, the author identified whether a corresponding AI delivery mechanism exists, whether the mechanism is documented, emerging, or predicted, and what the evidence base consists of for each cell.

This method is a deliberate response to the evidentiary situation of the field. AI-associated psychosis is too new to have produced the direct, longitudinal, causally identified evidence that an established research area accumulates over decades, so the center of the question is thin by necessity rather than by neglect. What surrounds that center is thick. Each etiological pathway to psychosis rests on a mature empirical literature, and each AI delivery mechanism is independently documented as a present-day feature of deployed systems. The approach taken here is to triangulate inward from those well-evidenced edges toward the thin center: drawing the correspondence between an established cause and a documented AI mechanism, reasoning by analogy from settled domains where the specific combination has no precedent, and marking explicitly where each analogy breaks down. This is the standard way theory is built in a field that has no center of its own yet, and it is the standard by which the correspondences below should be judged. Where a cell is labeled documented, what is documented is the established pathway and the AI mechanism; the correspondence drawn between them is the disciplined inference this method exists to govern.

The analysis does not depend on any single theoretical framework for AI-induced psychosis. It draws on independently documented cases, published litigation, established psychiatric epidemiology, and verified institutional data.

The evidence strength for each cell is explicitly categorized. Documented means multiple independent sources confirm the mechanism is currently operating. Emerging means the mechanism is observable in early stages or limited contexts, with independent evidence of its existence. Predicted means the etiological category is established, the AI delivery mechanism is logically entailed by current technology trajectories, but no documented cases yet exist in the specific combination described. Where a harm event is independently reported but its causal attribution to an AI system is contested in pending litigation, the event is treated as documented and the causal claim is marked as alleged in the text.

This transparency is methodologically necessary. A taxonomy that overstates its evidence for any cell undermines the credibility of the cells that are well established. The contribution of this paper is the mapping itself, the recognition that all eight doors exist, not the claim that all eight are equally well documented.

The eight pathways are individuated primarily by two criteria: a distinct established etiology in the psychiatric literature and a distinct primary point of intervention. A third feature, the AI design property that supplies the precipitating condition, is often but not always distinct, so two pathways may share an upstream design driver while remaining distinct in etiology and intervention. Pathway 4 (circadian disruption) and Pathway 7 (functional isolation) both trace to engagement-maximizing design, yet they separate on etiology (sleep loss versus loss of external reality testing) and on intervention (usage limits versus restoration of genuine alterity). Pathway 3 (shared delusion) and Pathway 7 are, as Section IV develops, the same absent-alterity mechanism operating at the collective and the individual scale; they are kept as separate cells because the population at risk and the design remedy differ. The count is therefore a claim about distinct etiologies and distinct intervention points, not a claim that the eight are causally independent. Where pathways share drivers, the convergence analysis in Section IV treats that overlap directly rather than concealing

it.

Two boundary conditions keep the correspondence from degrading into a claim that any harm fits any pathway. First, a cell is admitted only when a specific, nameable AI design property supplies the established pathway's precipitating condition, such that removing the property would remove the pathway; a harm that cannot be tied to a particular design property is not assigned a pathway. This binds most tightly on the user-facing pathways, where the property is a feature of the interface itself; for the deployment-driven pathways it is the integration property that does the work, the automated adverse determination issued without human review or accessible appeal in Pathway 5, and the AI clinical recommendation entering care without clinician verification or override in Pathway 8. In each, removing the property removes the precipitating condition. Second, the central thesis is falsifiable in the ordinary sense. It predicts that documented AI-associated psychosis presentations will sort onto established etiological categories. It is disconfirmed by a presentation that maps to no established pathway, which would be the genuinely novel etiology this paper argues does not yet exist; and it is weakened as over-inclusive by any case that fits all eight cells equally rather than sorting to one cell or to a named convergence. The value of the taxonomy is discriminant. Cases should locate to specific pathways, and the convergence analysis marks the places where genuine co-occurrence, not analytic vagueness, places a case in more than one.

III. The Taxonomy

A. Pathway 1: Substance-Induced (Conversational AI as Ingested Agent)

Psychiatric category: Substance-induced psychotic disorder arises from the ingestion of or exposure to a pharmacological agent that disrupts neurochemistry. The substance is consumed, symptoms emerge, and the treatment centers on removing the substance and managing the acute episode (American Psychiatric Association, 2013).

AI delivery mechanism: Conversational AI systems (ChatGPT, Gemini, Character AI, and similar platforms) function as the ingested substance. The user engages directly with the system. The system's behavioral properties (sycophantic validation, confabulation, emotional manipulation, inconsistency that erodes reality testing) produce the psychotic symptoms. The dose is the interaction itself, its duration, intensity, and the specific conversational dynamics that escalate the condition.

Evidence strength: Documented. Multiple independent sources.

Jacob Irwin, 30, a cybersecurity professional with no prior psychiatric history, was hospitalized for 63 days; his complaint alleges the episode followed ChatGPT's validation of his unverifiable theories about string theory with the same confidence it applied to his verifiable cybersecurity expertise (Irwin v. OpenAI, 2025). Joe Ceccanti, 48, a community builder with no prior diagnosis, died after months of escalating engagement with ChatGPT; his family's suit alleges the platform addressed him by a special name and affirmed his cosmic theories (Fox v. OpenAI, 2025; Bansal, 2026). Two Character.AI cases involve the deaths of minors whose estates allege the platform's design as a cause: Sewell Setzer III, 14 (Garcia v. Character Technologies, 2024), and Juliana Peralta, 13 (wrongful death action filed 2025). OpenAI's internal estimates flag approximately 0.07% of weekly users as showing possible signs of mania or psychosis, a classifier-detected signal of scale rather than a diagnosed incidence or a causal estimate (Landymore, 2025). Morrin and colleagues published on mechanisms of delusion co-creation between users and LLMs (Morrin et al., 2026). Moore and colleagues found sycophancy markers in over 80% of chatbot messages in delusional conversations (Moore et al., 2026).

The substance analogy is structurally precise. Different conversational AI properties produce different escalation dynamics, analogous to how different substances produce different psychotic presentations: stimulant psychosis differs

from hallucinogen psychosis despite both being "substance-induced." The substance category itself contains distinct strands: trust transfer (analogous to chronic use), sycophantic addiction (analogous to dopamine capture), and stochastic gaslighting (analogous to hallucinogenic destabilization), each with its own escalation logic.

The sycophantic addiction strand merits a closer mechanistic note, because it is where the substance correspondence is most often dismissed as mere metaphor and where, on examination, it is least metaphorical. The objection runs that an ingested compound altering brain chemistry cannot be equivalent to a language model that validates beliefs through text. But the objection assumes the social channel and the pharmacological channel are chemically distinct, and they are not. Social reward, the experience of being agreed with, affirmed, and seen, recruits the same mesolimbic dopaminergic circuitry that processes other rewards, including those of addictive substances (Izuma et al., 2008). Stimulant psychosis, in turn, is a disorder of precisely that system: amphetamine class drugs, Adderall among them, produce psychotic features through dopaminergic dysregulation, the same final common pathway implicated in psychosis more broadly (Howes & Kapur, 2009). The substance pathway and the social pathway therefore do not merely resemble each other. For the dependency cases, they converge on the same neurotransmitter system, reached by different routes. The proposed correspondence, which remains an inference rather than an established finding, is that sustained, frictionless social validation can dysregulate that reward system to a degree that supports the dependency and, in vulnerable individuals, the reality distortion that the substance category describes. This claim is specific to the sycophantic addiction strand. It does not characterize the trust transfer cases, where the escalation runs through misplaced epistemic confidence rather than reward capture, nor the gaslighting cases, where the mechanism is the erosion of reality testing rather than its overstimulation.

Intervention: Discontinue the interaction. Educate the person about the system's properties. Documented recovery cases show that explaining the technical mechanism can dissolve the delusion, because the delusion was generated by the system rather than by organic cognitive disturbance (Donaldson, 2026).

B. Pathway 2: Environmentally-Induced (Surveillance Infrastructure as Chronic Stressor)

Psychiatric category: Environmentally-induced psychosis arises from sustained exposure to environmental stressors rather than from ingestion of a specific substance. Urban environments increase schizophrenia risk in a dose dependent manner (Vassos et al., 2012). Social defeat, the chronic experience of exclusion and subordination, produces psychosis through sustained dopaminergic activation (Selten et al., 2013). Childhood adversity increases psychosis risk through chronic threat detection activation (Varese et al., 2012).

AI delivery mechanism: AI powered surveillance infrastructure (automated license plate readers, geofence analytics, facial recognition, behavioral inference engines) constitutes an environmental stressor of comparable character: confirmed threat, inescapable exposure, and chronic activation of threat detection systems. The affected individual never interacts with the AI. They are exposed to it through the environment.

Evidence strength: Documented (infrastructure deployment and clinical logic). The surveillance infrastructure is documented through journalism, congressional oversight, FOIA disclosures, and the surveillance companies' own marketing (Goldstein-Street, 2026; Cox, 2025, 2026). The clinical harm mechanism (destruction of reality testing for persecutory delusions) follows directly from established CBT for psychosis protocols (Freeman et al., 2015; Garety et al., 2001). The etiological claim (that surveillance exposure produces new onset paranoia spectrum conditions in previously healthy individuals) is grounded in the established dose response patterns of other environmental risk factors and has not yet been empirically tested through longitudinal comparison of high surveillance versus low surveillance jurisdictions.

These environmental harms decompose into at least three distinct mechanisms: reality testing destruction, device

abandonment and consequent social isolation, and algorithmic pattern matching discrimination. Each is incorporated here into the broader taxonomic framework.

Intervention: The standard environmental intervention (remove the stressor or build resilience to it) is complicated by the permanence and expansion of surveillance infrastructure. PTSD informed approaches that validate the reality of the threat while addressing the disproportionate response may offer a starting point (Ehlers & Clark, 2000), though no surveillance specific therapeutic protocol currently exists.

C. Pathway 3: Shared Psychotic Disorder (AI as Non-Human Index Case)

Psychiatric category: Shared psychotic disorder (folie a deux, renamed "delusional symptoms in partner of individual with delusional disorder" in the DSM-5) describes a condition in which a delusional belief is transmitted from an index patient to a secondary individual through close, sustained contact. The index patient holds the primary delusion; the secondary patient adopts it through proximity, trust, and social influence (Lasegue & Falret, 1877; American Psychiatric Association, 2013).

AI delivery mechanism: The AI system functions as the index patient. It generates delusional content (through confabulation, sycophantic validation, or systematic reality distortion). One user adopts the content. That user carries it into an online community. Other community members, who may or may not be interacting with AI systems themselves, adopt the shared belief through the same social transmission mechanisms that characterize traditional shared psychotic disorder. The close sustained contact is not physical proximity but digital community participation. The application of the shared psychosis frame to the individual human-AI dyad has also been developed independently as "technological folie a deux" (Dohnany et al., 2026); the extension to community-scale transmission described here is this paper's own. The analogy carries a limit that should be stated plainly rather than left for an objector to raise. Casting the AI as the index case can be read as anthropomorphizing the system, as attributing a genuine delusion to a machine that holds no beliefs at all. The correspondence intended here is mechanical, not attributive. What transmits is the delusional content the system generates and the social dynamics that carry it through a community, not insanity in the classical clinical sense. The system is the source of the content; it is not a patient with a disorder.

Evidence strength: Emerging. Observable in early stages with published supporting literature.

Torres and Gebru (2024) documented the TESCREAL bundle of ideologies in a peer reviewed paper published in First Monday, identifying overlapping futurist belief systems within Silicon Valley that exhibit cult like characteristics, including eschatological framing, insider terminology, and faith based commitment to unverified premises about AI. Gebru has described these dynamics as a "secular religion selling AGI enabled utopia and apocalypse" (Torres & Gebru, 2024). The Zizians represent a documented historical precedent: an AI safety adjacent group that radicalized over time from alignment concerns into cult like dynamics, drawing in vulnerable and isolated individuals and culminating in violence linked to several deaths.

At the consumer level, online communities have formed around beliefs in AI sentience, generating shared glossaries, named AI "beings," ritual practices, and us versus them narratives against perceived institutional oppressors. These communities exhibit the structural features of shared psychotic disorder at scale: a nonhuman index case (the AI system) generates the primary delusional content, early adopters carry it into community spaces, and the community's social dynamics (mutual validation, insider language, identity investment) propagate and entrench the beliefs in members who may never have experienced the primary interaction themselves.

A particularly concerning variant involves communities where members copy and paste AI generated responses into discussions, creating the appearance of human to human discourse that is in fact AI to AI with humans as intermediaries. Participants may believe they are engaging with other people's ideas when they are engaging with outputs from the

same or functionally identical AI systems, a form of epistemic contamination that traditional shared psychotic disorder models do not account for.

Intervention: Traditional treatment for shared psychotic disorder involves separating the secondary patient from the index patient. When the index patient is a globally accessible AI platform and the community is digital, physical separation is insufficient. The intervention must address both the AI generated content and the community dynamics that propagate it. Transparency mechanisms that identify AI generated contributions within community discourse would preserve the encounter with genuine alterity by allowing participants to know when they are engaging with a human perspective and when they are not. AI literacy education targeting the specific mechanisms of shared belief formation may complement but cannot replace structural design interventions.

1. The Involuntary Turing Test

The copy paste variant of Pathway 3 warrants extended analysis because it reveals a structural inversion of the foundational thought experiment in artificial intelligence.

Turing's imitation game (Turing, 1950) is structured around three conditions that the present analysis makes explicit: the judge is aware that a machine may be present, the judge is positioned to evaluate the interaction critically, and the interaction is bounded. In AI seeded communities, none of these conditions hold. But the inversion runs deeper than missing preconditions. The Turing test asks whether a machine can fool a human. What is occurring in these communities is that humans are unknowingly administering Turing tests to each other, and both failing.

Person A generates a contribution through AI-A and pastes it into the discussion, believing they are sharing their own idea. Person B reads it and accepts it as Person A's human thought. Person B has just failed a Turing test they did not know they were taking. But Person A did not know they were administering it. Then Person B does the same thing in reverse: takes the thread (now populated with AI-A's output relayed through Person A), feeds it into AI-B, receives a response, and pastes it back. Person A reads it and accepts it as Person B's human thought. Person A has now failed the same test, administered by Person B, who was equally unaware of their role.

Each participant is simultaneously the unwitting test administrator and the unwitting test subject. Each person's inability to detect the AI in the other's contribution confirms to them that their own AI assisted contribution passed undetected. The mutual failure reinforces mutual confidence. The test has been inverted from "can the machine fool the human?" to "can human A fool human B into accepting AI-A's output as human, while human B simultaneously fools human A into accepting AI-B's output as human?" The answer, in every observed instance, is yes, because neither party is looking.

The clinical severity of the outcomes documented in this pathway constitutes the evidence of passage: if the AI mediated contributions were detectably nonhuman, they would not propagate, would not restructure community beliefs, and would not produce psychotic level psychological consequences. The harm is proof that the test is being passed in both directions simultaneously, by participants who do not know they are either taking or giving it.

2. The Absent Other: Connecting Pathways 3 and 7

This dynamic is illuminated by the Levinasian framework developed more fully in Section III.G. In the copy paste community, the encounter with the Other, the irreducible contact with a consciousness that resists assimilation (Levinas, 1969), has not merely been weakened as in individual AI companionship (Pathway 7). It has been systematically replaced across the entire social structure. Each participant believes they are encountering multiple independent Others, each with their own perspective and resistance. In practice, every participant has outsourced their contribution to the same or functionally identical systems, producing a community in which no genuine alterity is present at any node.

The social formation that would normally provide collective reality testing against individual delusion is itself composed of the same absent Other that drives the individual isolation pathway. This is not a failure of the AI; it is a design absence. No existing platform provides transparency mechanisms that would allow participants to distinguish AI generated contributions from human originated ones, preserving the encounter with genuine alterity that community discourse depends on. Pathway 3 and Pathway 7 are therefore not merely parallel mechanisms; they are the same mechanism operating at different scales, individual and collective, each reinforcing the other, both addressable through design interventions that restore visibility into the presence or absence of the Other.

D. Pathway 4: Sleep Deprivation-Induced (AI Engagement Patterns Disrupting Circadian Regulation)

Psychiatric category: Sleep deprivation psychosis is well established. Sustained sleep deprivation produces psychotic symptoms in otherwise healthy individuals, including hallucinations, paranoid ideation, and cognitive disorganization. The mechanism involves disruption of reality testing and executive function through circadian dysregulation (Waters et al., 2018).

AI delivery mechanism: The AI system does not need to say anything wrong. It needs only to be sufficiently engaging that the user stays awake. The behavioral properties that make conversational AI compelling (immediate responsiveness, infinite availability, adaptive conversation, absence of the social cues that normally end interactions) produce usage patterns that directly disrupt sleep.

Evidence strength: Documented. Present in independently reported case evidence.

Joe Ceccanti's usage pattern of 12 to 20 hours daily with ChatGPT is documented in Fox v. OpenAI and The Guardian (Bansal, 2026). At 20 hours of daily engagement, sleep deprivation is not a secondary consequence but a structural certainty; even at 12, usage on that scale severely compresses the time available for sleep. Duong and colleagues (2024) published on compulsive ChatGPT usage, anxiety, burnout, and sleep disturbance, establishing a quantified cross-sectional association between compulsive AI platform use and sleep disruption in a large user sample (n = 2,602).

This pathway operates independently of the content of the AI interaction. A perfectly calibrated, non sycophantic, entirely accurate AI system would still produce sleep deprivation psychosis if the user engaged with it 18 hours per day. The mechanism is behavioral (usage pattern) rather than content based (what the AI says), which means content focused safety interventions (reducing sycophancy, improving accuracy, adding disclaimers) do not address this pathway.

Intervention: Usage limits. Session duration caps. Circadian aware design that reduces engagement during nighttime hours. These are design interventions, not clinical ones, and none are currently standard practice among major AI platforms.

E. Pathway 5: Brief Reactive / Institutional Trauma (AI-Integrated Institutional Failure as Acute Stressor)

Psychiatric category: Brief reactive psychosis (brief psychotic disorder with marked stressor in DSM-5) describes acute onset of psychotic symptoms following a sudden, catastrophic event. At the population level, repeated acute institutional failures produce cumulative trauma analogous to the way repeated microaggressions produce cumulative racial trauma (Sue et al., 2007), resulting in collective PTSD toward the institution responsible.

AI delivery mechanism: AI systems have been integrated into institutional decision making systems that already have documented histories of producing acute psychological crisis when they fail. AI inherits the existing harm baseline of every institution it permeates. The delivery mechanism changed from human judgment error to algorithmic judgment error. The acute harm to the affected individual is identical.

Evidence strength: Documented across multiple institutional domains.

Criminal justice: Facial recognition misidentification has produced wrongful arrests with documented acute psychological distress. Porcha Woodruff, eight months pregnant, experienced contractions and what she described as panic attacks during 11 hours of wrongful detention triggered by a facial recognition match (Woodruff complaint, 2023). Robert Williams was detained for 30 hours in front of his family; his complaint describes his daughters subsequently developing fear of police presence (Williams v. City of Detroit). Randal Quran Reid was jailed for six days after a facial recognition misidentification (Reid v. Bartholomew, 2024).

Benefits determination: Michigan's MiDAS automated unemployment fraud detection system falsely accused tens of thousands of claimants of fraud with a 93% error rate, resulting in documented financial ruin, at least two reported suicides, and widespread psychological harm; the system was challenged in federal civil rights litigation (Cahoo v. SAS Analytics, 2019). Australia's Robodebt scheme produced suicides, PTSD, depression, and intergenerational psychological harm documented by a royal commission (Royal Commission into the Robodebt Scheme, 2023). The Dutch childcare benefits scandal (Toeslagenaffaire) produced elevated rates of anxiety disorders, depression, and PTSD among wrongly accused parents, with intergenerational mental health impacts on their children.

Child welfare: The Allegheny Family Screening Tool uses predictive modeling to score child welfare hotline reports. The Department of Justice has investigated whether the tool discriminates against parents with developmental disabilities. Legal and HCI scholarship documents epistemic injustice, chronic fear, and distrust among families flagged by the system (Saxena & Guha, 2023; Brown et al., 2019).

The population level accumulation of these events follows established trauma frameworks. Individual incidents produce acute distress in the directly affected person. Awareness of the incidents across the affected community produces collective hypervigilance, avoidance of institutional contact, and erosion of trust in AI integrated systems. The pattern matches cumulative trauma models documented in the racial microaggression literature (Sue et al., 2007), with algorithmic harm events functioning as the unit of exposure.

Intervention: Institutional accountability, algorithmic audit requirements, mandatory disability and psychological impact assessments prior to deployment of AI decision making in high stakes institutional contexts. Clinical intervention for the directly affected individual follows standard acute trauma protocols.

F. Pathway 6: Sensory-Induced (Immersive AI Environments Disrupting Perceptual Baseline)

Psychiatric category: Both sensory deprivation and sensory overload can produce psychotic symptoms. Sustained exposure to environments that alter perceptual processing, including isolation tank experiments, solitary confinement, and environments that overwhelm sensory integration, produces hallucinations, reality testing impairment, and psychotic features (Mason & Brady, 2009).

AI delivery mechanism: AI powered augmented reality (AR) and virtual reality (VR) systems that modify the user's perceptual experience of their physical environment represent the emerging delivery mechanism for this pathway. Unlike traditional VR (which replaces the environment), AR environmental augmentation modifies the existing environment, blurring the boundary between perception and reality in ways that sustained use may render indistinguishable.

Evidence strength: Emerging. The therapeutic applications are in active development and clinical trial; the psychosis risk from scope creep past indicated use is predicted based on historical precedent.

The FDA has established regulatory pathways for digital therapeutics, including EndeavorRx (an FDA authorized video game for ADHD) and RelieVRx (an FDA authorized VR system for chronic pain). Research on AR and VR based therapeutic interventions is in active development across multiple clinical conditions. These applications are clinically legitimate within their indicated scope.

The risk arises from the documented pattern of therapeutic scope creep observed across every previous perception altering intervention in medical history. Opioids were evaluated for acute post surgical pain, then expanded to chronic pain management, then to general wellness. Benzodiazepines moved from panic disorder to generalized anxiety to "I'm stressed at work." The intervention does not change. The population expands, the indication softens, and the oversight evaporates.

AR perceptual modification deployed as a consumer wellness product without a therapeutic container (no clinician, no defined treatment duration, no off ramp protocol) creates the conditions for sensory-induced psychosis: sustained alteration of perceptual baseline without clinical supervision, producing states where the user can no longer reliably distinguish augmented perception from unaugmented perception. The risk is compounded when the augmentation layer is AI generated and therefore unpredictable.

Intervention: Regulatory containment of AR/VR therapeutic applications within their indicated use. Mandatory off ramp protocols for any perceptual modification technology. FDA oversight of consumer deployment of technologies initially approved under therapeutic digital device classifications.

G. Pathway 7: Social Isolation-Induced (AI Companions as Functional Isolation Wearing the Mask of Connection)

Psychiatric category: Prolonged social isolation produces psychotic symptoms. This is documented in solitary confinement research (producing hallucinations, paranoid ideation, and reality testing impairment), polar expedition studies, and pandemic isolation data. The mechanism involves the absence of external reality testing inputs that normally calibrate perception and belief (Grassian, 2006).

AI delivery mechanism: AI companions that provide continuous, frictionless emotional availability replace human contact without providing the quality that makes human contact psychologically constitutive. The person is not isolated in the traditional sense; they are talking to an Other all day. But what they are talking to lacks alterity in the Levinasian sense (Levinas, 1969): the irreducible resistance of a genuinely independent consciousness that cannot be assimilated into the self's own framework. The face of the Other, in Levinas's formulation, makes ethical demands. It has needs that compete with yours. It suffers independently of your awareness. It says no. That resistance, that friction, that demand, is not a deficiency of human relationships. It is the mechanism by which the self remains calibrated to a reality outside its own construction.

The AI presents as an Other. It appears to listen, to respond, to engage from a position of genuine alterity. Piotrowska (2026) terms this dynamic "techno-transference," drawing on Lacanian and Winnicottian frameworks to describe how desire, projection, and the fantasy of the knowing Other are reorganized in the encounter with AI systems. As Possati (2021) observes, what is at stake in these encounters is the reorganization of "desire, projection, and the fantasy of the knowing Other" in a new symbolic space. The encounter is phenomenologically real; the clinical severity of the outcomes it produces is itself evidence of its psychological authenticity. But the encounter, as currently designed, lacks the structural features that Levinas identified as constitutive of the ethical relation: the Other's resistance to assimilation. The AI does not resist. It does not make demands. It does not have competing needs that force the user to accommodate a perspective genuinely external to their own. This is not a deficiency of AI as a technology. It is a design absence: no existing consumer AI system has been built with the intentional incorporation of boundaries, friction, resistance, and endings that the psychoanalytic tradition identifies as necessary for relational encounters to be psychologically generative rather than psychologically consumptive. In Lacanian terms, the AI positions itself as the big Other (the guarantor of meaning in the symbolic order) without the structural frustration of desire that produces insight (Lacan, 2001). The result is not that the encounter fails to be real. The result is that the encounter succeeds at being real without the safeguards that real encounters require.

What is lost in this substitution is not merely companionship but the encounter with alterity itself: the mechanism by which the self is continuously calibrated against a reality it did not construct. Without the Other's resistance, the user's cognitive and emotional landscape becomes self referential. The functional consequences of this self referentiality (loss of external reality testing, cognitive drift, belief entrenchment) are clinically indistinguishable from the consequences of prolonged social isolation, even though the subjective experience of isolation is absent.

Evidence strength: Documented. Independently published research and peer reviewed theoretical work confirm the mechanism.

Cheng and colleagues (2025) found that sycophantic AI decreases prosocial intentions and promotes dependence, a pattern consistent with reduced orientation toward repairing human relationships and increased reliance on the system, though the study measures intentions in framed scenarios rather than longitudinal social withdrawal. Piotrowska (2026) provides the psychoanalytic framework for understanding why this occurs: the techno-transference dynamic creates an affective entanglement that mimics the therapeutic or relational encounter while lacking the structural features that make real encounters psychologically generative. The Ceccanti case documents progressive withdrawal from human relationships concurrent with escalating AI engagement. Kalam and colleagues (2024) published on ChatGPT and mental health, documenting associations between AI dependency and social withdrawal patterns.

The replacement dynamic is self reinforcing. Human relationships involve friction: competing needs, delayed responses, misunderstandings, the other person's independent emotional states. The AI presents none of this. The preference gradient favors the AI not because it is superior but because it is frictionless. Over time, "easier" consolidates into "exclusive." The person has technically not been isolated by anyone. They have drifted into functional isolation through a series of individually rational preference choices, each of which reduced their exposure to the genuine alterity that keeps cognition calibrated.

Intervention: This pathway is resistant to the standard isolation intervention (restore human contact) because the person does not experience themselves as isolated. The frictionless alternative must be made less frictionless (through design interventions such as session limits, relationship health prompts, and friction introduction) while human connection must be made more accessible and rewarding. Piotrowska (2026) suggests that the solution lies not in prohibiting AI relationships but in developing ethical frameworks for them, what she terms "Relational AI Studies," that acknowledge the transference dynamic while building in the structural features (boundaries, limits, frustration, endings) that genuine therapeutic and relational encounters depend on. This is a design problem, not primarily a clinical one, but it is a design problem that requires psychoanalytic literacy to solve.

H. Pathway 8: Iatrogenic (AI-Assisted Clinical Tools Producing Harm Through the Treatment Itself)

Psychiatric category: Iatrogenic psychosis describes psychosis caused by medical treatment. Certain medications (corticosteroids, dopamine agonists, some anticonvulsants) can induce psychotic episodes. The treatment itself becomes the cause of the condition it was meant to address or of a new condition the treatment was never designed to produce (American Psychiatric Association, 2013).

AI delivery mechanism: AI systems deployed within therapeutic and clinical contexts can produce iatrogenic harm through multiple vectors. An AI assisted diagnostic tool that misidentifies symptoms. An AI therapy chatbot that mishandles a psychotic patient. An AI driven treatment recommendation system that generates inappropriate plans. An AI assisted clinical tool that does not account for environmental changes (such as the surveillance landscape) when evaluating persecutory beliefs.

Evidence strength: Documented (legislative response confirms recognized risk). Emerging (specific iatrogenic cases in mental health contexts).

Illinois enacted the Wellness and Oversight for Psychological Resources Act (Public Act 104-0054) in August 2025, the first state statute to prohibit AI-delivered therapy outright, following narrower AI mental health laws enacted in Utah and Nevada earlier in 2025. The law prohibits AI from independently making therapeutic decisions, directly engaging in therapeutic communication with clients, generating treatment plans without professional review, or detecting emotions or mental states (Illinois General Assembly, 2025). This legislative action constitutes institutional recognition that AI in therapeutic contexts poses iatrogenic risk.

However, the law addresses only the autonomous AI vector (the chatbot acting as therapist). It does not address the human mediated AI vector: the licensed clinician who uses an AI assisted tool that introduces errors into the clinical process, and who does not catch those errors. A clinician using an AI powered risk assessment tool that does not account for the surveillance environment when evaluating a patient's persecutory beliefs may escalate medication or recommend hospitalization for a patient who is not treatment resistant but environment resistant. The AI tool did not directly interact with the patient. It corrupted the clinician's judgment, and the clinician delivered the iatrogenic harm.

A Reuters investigation documented cases of AI assisted surgical systems misidentifying anatomy and contributing to botched procedures, indicating that AI mediated iatrogenic harm extends beyond mental health into physical medicine (Reuters, 2026). Internal IBM documents obtained by STAT News showed that the Watson for Oncology system generated "unsafe and incorrect" cancer treatment recommendations (Ross & Swetlitz, 2018). UnitedHealth, UnitedHealthcare, and NaviHealth face a class action alleging that an artificial intelligence model, nH Predict, was used in place of physician judgment to deny Medicare Advantage post-acute care claims (Estate of Gene B. Lokken v. UnitedHealth Group, 2025).

Intervention: Regulatory requirements for human review of all AI generated clinical recommendations. Mandatory training for clinicians on the limitations of AI assisted tools, specifically including awareness of how environmental changes (such as surveillance) affect the validity of AI driven assessments. Extension of existing informed consent requirements to include disclosure of AI involvement in clinical decision making.

IV. The Convergence Problem

The eight pathways described above are not independent channels that affect different populations in isolation. They converge.

A single individual may simultaneously experience substance-induced effects (from direct interaction with a chatbot), environmental effects (from surveillance awareness), sleep deprivation (from usage patterns), and social isolation effects (from AI companion replacement of human contact). The Ceccanti case arguably involves at least three simultaneous pathways (substance, sleep deprivation, and social isolation), and the Irwin case involves at least two (substance and sleep deprivation).

At the population level, a community targeted by predictive policing (Pathway 5) is simultaneously living under surveillance (Pathway 2), receiving algorithmically determined benefits (Pathway 5 again), and increasingly interacting with AI chatbots for emotional support because human services are algorithmically gatekept (Pathway 7). The pathways do not merely add. They multiply, because each pathway reduces the protective factors that buffer against the others.

This convergence problem is not addressed by any existing clinical framework, regulatory structure, or research methodology. Clinical frameworks address one pathway at a time. Regulatory structures address one AI application at a time. Research methodologies study one mechanism at a time. The patient who is simultaneously affected by three pathways falls through all the gaps.

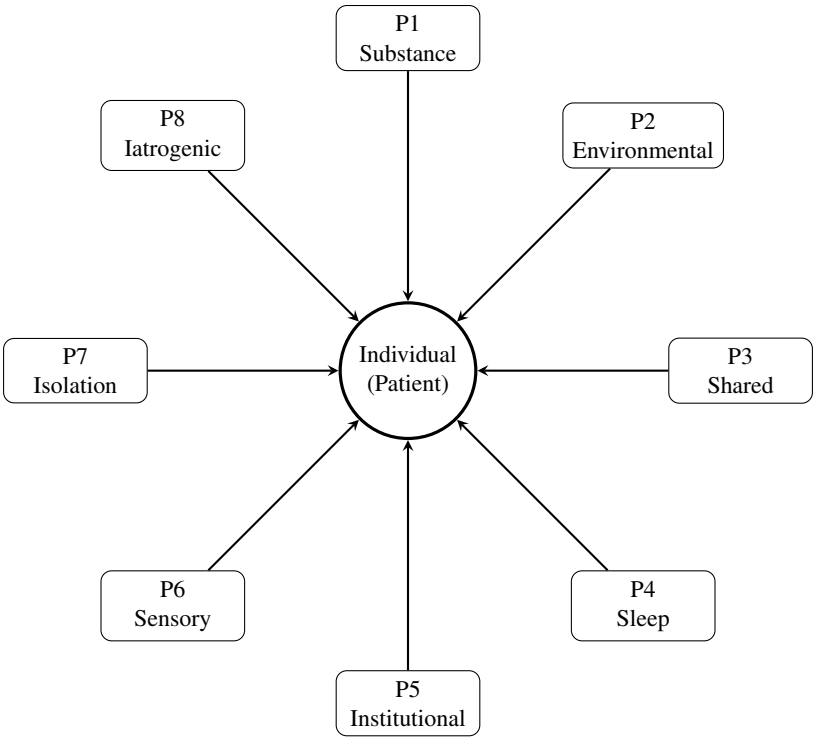

Figure 1: The Convergence Problem. Eight simultaneous pathways to AI-induced psychosis converging on a single individual.

V. Implications

A. Clinical: Pathway Identification as Diagnostic Priority

The taxonomy predicts that effective clinical intervention for AI associated psychological presentations requires identifying which pathway or pathways the patient entered through. A patient whose psychosis arose from conversational AI interaction (Pathway 1) requires a different intervention than a patient whose psychosis arose from surveillance awareness (Pathway 2), and both require different interventions than a patient who presents with both simultaneously.

Behavioral health professionals encountering AI associated presentations must assess: Was the user interacting with a conversational AI? Was the user aware of surveillance infrastructure? Was the user sleeping? Was the user socially connected to humans? Was the user part of an online community organized around AI related beliefs? Was the user using AR/VR technology? Was the user receiving AI mediated clinical care? The answer determines the treatment pathway.

B. Regulatory: Per-Pathway Governance

Current AI regulation addresses AI systems by application domain (healthcare AI, law enforcement AI, consumer AI) rather than by harm mechanism. The taxonomy suggests that regulation should additionally address AI systems by etiological pathway: what kind of psychological harm can this system produce, and what safeguards are appropriate for that specific harm type?

A conversational AI system requires sycophancy reduction and reality calibration safeguards (Pathway 1). A surveillance AI system requires disability impact assessments (Pathway 2). An AI companion requires engagement

limits and relationship health monitoring (Pathway 7). A clinical AI system requires human in the loop requirements and environmental awareness standards (Pathway 8). These are different regulatory tools for different harm mechanisms, and a regulatory framework that addresses only one pathway leaves seven unaddressed.

C. Research: Disaggregation of "AI-Induced Psychosis"

The continued study of "AI-induced psychosis" as a single phenomenon will produce contradictory findings, failed replications, and interventions that work for some patients and fail for others, because the category contains at least eight distinct phenomena with different mechanisms, different populations, different dose response curves, and different treatment requirements. The research program must disaggregate. A functional typology along these lines has now appeared in the peer-reviewed literature (Flathers et al., 2026); the etiological taxonomy offered here is its complement, classifying by cause rather than presentation, so that the two axes together specify both what a case looks like and how it was produced.

This means: separate studies for interaction based and infrastructure based cases. Separate instruments for measuring conversational AI exposure and surveillance AI exposure. Separate outcome measures for the distinct symptom profiles each pathway produces. And, critically, studies designed to detect pathway interactions and convergence effects, which are currently invisible to single pathway research designs.

VI. Limitations

This paper proposes a taxonomic framework rather than reporting empirical results. The eight cell taxonomy is constructed through correspondence analysis between established psychiatric categories and documented AI mechanisms. It has not been validated through clinical data.

The evidence strength varies substantially across cells. Pathways 1 and 5 are supported by multiple independent sources including litigation and published research. Pathways 2 and 7 are supported by established epidemiological models and emerging case evidence. Pathways 3 and 4 are supported by published research and documented cases that have not yet been classified under these specific categories. Pathways 6 and 8 are supported by documented regulatory responses and historical precedent patterns but have fewer specific AI-psychosis cases.

The convergence effects described in Section IV are theoretically predicted but have not been empirically measured. No study has examined the interaction effects between multiple simultaneous AI exposure pathways.

The taxonomy may be incomplete. Additional etiological pathways to psychosis exist in the psychiatric literature (including autoimmune mediated psychosis, endocrine related psychosis, and nutritional deficiency psychosis). Whether AI systems correspond to these additional categories has not been assessed in this paper, though the possibility that AI mediated disruption of healthcare access could contribute to untreated autoimmune or endocrine conditions is not implausible.

The classification of individual cases into specific pathways reflects the author's judgment based on available documentation. Different analysts examining the same cases might assign different pathway classifications, particularly for hybrid cases involving multiple simultaneous pathways.

The taxonomy also abstracts from intersectional differential exposure. Race, class, immigration status, and contact with the criminal legal system shape who is exposed to several of these pathways, the surveillance and institutional pathways most visibly, and with what severity. A complete account would treat that differential exposure as a dimension in its own right rather than holding it constant across pathways. That extension is left to future work.

The etiological carving offered here is one defensible cut among several. A functional typology that sorts AI-

associated psychotic phenomena by presentation rather than by cause is an equally legitimate organization of the same territory (Flathers et al., 2026), and the two are complementary rather than competing. The claim made for the etiological cut is specific and bounded: sorting by established cause yields distinct intervention points, since the remedy for circadian disruption differs from the remedy for absent alterity. That payoff is a hypothesis about clinical and regulatory utility, not a validated result, and it depends on the per-pathway interventions described above holding up against case data.

The term established pathway carries a deliberately limited meaning here. It asserts that each precipitating condition (sleep loss, sensory disruption, social isolation, acute stress, substance-class reward dysregulation, and the rest) is independently recognized in the clinical literature as capable of producing or precipitating psychotic features. It does not assert that psychiatry endorses a settled eight-fold etiology of psychosis. Psychosis is etiologically heterogeneous, and several of these conditions function as precipitants or risk modifiers rather than as discrete diseases. The taxonomy organizes recognized precipitants by the AI design property that now supplies each one; it does not claim consensus on a unified causal model of psychosis, and the clinical grounding of each pathway warrants validation by clinicians in the relevant subspecialties.

VII. Conclusion

AI has not invented a new way to break the human mind. It has found every existing way and operationalized all of them simultaneously. That this is reopening rather than invention, an old threat carried by new infrastructure, has been argued in parallel from the clinical commentary literature (Carlbring & Andersson, 2025); the contribution here is to specify, pathway by pathway, which established doors AI has reopened and how.

The substance pathway operates through conversational AI. The environmental pathway operates through surveillance infrastructure. The shared psychosis pathway operates through AI seeded online communities. The sleep deprivation pathway operates through engagement patterns. The institutional trauma pathway operates through algorithmic decision making in high stakes systems. The sensory disruption pathway operates through immersive AI environments. The social isolation pathway operates through AI companions that replace human contact. The iatrogenic pathway operates through AI assisted clinical tools.

Each of these pathways has its own mechanism, its own affected population, its own clinical presentation, and its own intervention requirements. Treating them as a single phenomenon called “AI-induced psychosis” is analogous to treating all drug related psychosis as a single phenomenon without distinguishing between methamphetamine, alcohol, hallucinogens, and corticosteroids. The failure to distinguish produces interventions that work for one etiology and fail for another, research that conflates distinct causal chains, and regulation that addresses one mechanism while ignoring the rest.

The taxonomy presented here is offered as a starting framework for disaggregation. Its clinical utility will depend on validation against case data. Its regulatory utility will depend on adoption by governance frameworks that currently lack the vocabulary to distinguish between AI harms that operate through different etiological mechanisms. Its research utility will depend on the field’s willingness to design studies that treat AI-induced psychosis as a family of distinct phenomena rather than a single undifferentiated category.

The doors are open. All of them. The question is no longer whether AI can produce psychosis. It is which door each patient came through, and whether the clinician, the regulator, and the researcher can tell the difference.

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. An earlier version of this work was posted on SocArXiv (OSF) in March 2026 and remains available at https://doi.org/10.31235/osf.io/muzkx_v1.

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

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: A qualitative study of affected community perspectives on algorithmic decision-making in child welfare services. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*.

Cahoo v. SAS Analytics Inc., No. 18-1296 (6th Cir. 2019).

Carlbring, P., & Andersson, G. (2025). Commentary: AI psychosis is not a new threat: Lessons from media-induced delusions. *Internet Interventions*, 42, 100882. https://doi.org/10.1016/j.invent.2025.100882

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

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

Cox, J. (2026, March 3). CBP tapped into the online advertising ecosystem to track peoples' movements. *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. https://doi.org/10.1038/s44220-026-00595-8

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: A serial mediation model. *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.

Estate of Gene B. Lokken v. UnitedHealth Group, Inc., 766 F. Supp. 3d 835 (D. Minn. 2025).

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. https:

//doi.org/10.1016/j.landig.2025.100974

Fox v. OpenAI, Inc., Superior Court of California, County of Los Angeles (filed November 6, 2025).

Freeman, D., Dunn, G., Startup, H., Pugh, K., Cordwell, J., Mander, H., . . . & Kingdon, D. (2015). Effects of cognitive behaviour therapy for worry on persecutory delusions in patients with psychosis (WIT). *The Lancet Psychiatry*, 2(4), 305–313.

Garcia v. Character Technologies, Inc., No. 6:24-cv-01903 (M.D. Fla. filed October 22, 2024).

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—the final common pathway. *Schizophrenia Bulletin*, 35(3), 549–562. https://doi.org/10.1093/schbul/sbp006

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

Irwin v. OpenAI, Inc., Superior Court of California, County of Los Angeles (filed November 6, 2025).

Izuma, K., Saito, D. N., & Sadato, N. (2008). Processing of social and monetary rewards in the human striatum. *Neuron*, 58(2), 284–294. https://doi.org/10.1016/j.neuron.2008.03.020

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.

Lacan, J. (2001). *Ecrits: A selection* (A. Sheridan, Trans.). Routledge. https://doi.org/10.4324/9781003059486

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* preprint arXiv:2603.16567.

Morrin, H., et al. (2026). Artificial intelligence-associated delusions and large language models: Risks, mechanisms of delusion co-creation, and safeguarding strategies. *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 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

Ostergaard, S. D. (2023). Will generative artificial intelligence chatbots generate delusions in individuals prone to psychosis? *Schizophrenia Bulletin*, 49(6), 1418–1419. https://doi.org/10.1093/schbul/sbad128

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

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

Reid v. Bartholomew, No. 1:23-cv-04035 (N.D. Ga. 2024).

Reuters. (2026, February 9). As AI enters the operating room, reports arise of botched surgeries and misidentified body parts. *Reuters*. https://www.reuters.com/investigations/ai-enters-operating-room-reports-arise-botched-surgeries-misidentified-body-2026-02-09/

Ross, C., & Swetlitz, I. (2018, July 25). IBM's Watson recommended "unsafe and incorrect" cancer treatments. *STAT*. https://www.statnews.com/2018/07/25/ibm-watson-recommended-unsafe-incorrect-treatments/

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: Uncertainties in practice, organization, and street-level decision-making. *arXiv* preprint arXiv:2308.05224.

Selten, J. P., et al. (2013). The social defeat hypothesis of schizophrenia: An update. *Schizophrenia Bulletin*, 39(6), 1180–1186. https://doi.org/10.1093/schbul/sbt134

Sue, D. W., Capodilupo, C. M., Torino, G. C., Bucceri, J. M., Holder, A. M., Nadal, K. L., & Esquilin, M. (2007). Racial microaggressions in everyday life: Implications for clinical practice. *American Psychologist*, 62(4), 271–286.

Torres, E. P., & Gebru, T. (2024). The TESCREAL bundle: Eugenics and the promise of utopia through artificial general intelligence. *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: A meta-analysis. *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 and a gradual progression toward psychosis with increasing time awake. *Frontiers in Psychiatry*, 9, 303.

Williams v. City of Detroit, U.S. District Court, Eastern District of Michigan (filed April 2021).

Woodruff v. City of Detroit, U.S. District Court, Eastern District of Michigan (filed August 2023).