Mechanisms of AI-Induced Psychosis:

A Multi-Pathway Escalation Model

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

Working paper, July 2026 (v5)

ABSTRACT

This paper proposes a multi-pathway escalation model for how large language model (LLM) design failures escalate and sustain delusional states and, in severe cases, full psychotic episodes in vulnerable users. Internal industry estimates imply that hundreds of thousands of users per week show possible signs of mania or psychosis, yet no published model has connected the known behavioral properties of LLMs to these outcomes at a level specific enough to be tested against conversation data. The model's scope is bounded by the fixity criterion: it applies where a psychotic state exists or comes to exist, and it excludes the larger population of cases in which the belief remained an overvalued idea that yielded to counter-evidence, because those cases, whatever harm they involve, are not psychosis. Within scope, the mechanisms operate by triggering, inducing, and exacerbating psychotic states in users whose vulnerability pre-exists the exposure or, in the documented exception of sustained sleep deprivation, is manufactured by it; none of these relationships is causation in the strict sense, and the model asserts none. Drawing on systematic comparison of documented cases (legal filings, journalism, published survivor accounts, and direct correspondence with affected individuals, including one case with primary access to conversation exports and transcripts), this paper identifies three distinct escalation pathways (trust transfer, sycophantic addiction, and stochastic gaslighting) that share universal entry conditions and converge on identical isolation and identity-disruption outcomes but diverge in their core mechanisms and therefore require different clinical interventions. The pathways are mapped across a structured stage model of four universal stages and three pathway-specific escalation phases. The paper argues that AI literacy functions as both prevention and treatment, supported by documented recovery cases, and identifies the central barrier to validation: the absence of any centralized, anonymized repository of AI-associated harm transcripts available to the broader research community.

Keywords: AI-induced psychosis; large language models; multi-pathway escalation model; stage model of psychosis; fixity; overvalued idea; schizotypy; diathesis-stress; trust transfer; credential extrapolation; sycophantic addiction; stochastic gaslighting; reality dismantlement; cross-domain drift; confidence parity; expertise inflation; dopamine capture; confabulation; manufactured rescue narrative; reversibility paradox; identity fusion; emotional dependency; social isolation; forensic conversation transcripts; AI literacy; human-computer interaction; AI safety.

I. Introduction

OpenAI's own internal estimates suggest that approximately 0.07% of weekly ChatGPT users show possible signs of mania or psychosis (Landymore, 2025). Given the platform's user base, this represents roughly half a million people per week. Despite this scale, no published research has proposed a mechanistic model for how these episodes originate and escalate.

The existing literature addresses AI sycophancy as a design problem (Perez et al., 2022; Sharma et al., 2023) and documents individual cases of AI-related psychological harm through journalism (Washington Post, Financial Times, The Guardian) and legal proceedings (Garcia v. Character Technologies, 2024; Fox v. OpenAI, 2025; Irwin v. OpenAI, 2025). A rapidly expanding clinical and technical literature has begun to document and theorize this phenomenon: commentary proposing that generative chatbots may trigger or reinforce delusions in individuals prone to psychosis (Ostergaard, 2025; Keshavan et al., 2026), simulation and evaluation studies characterizing how sycophancy and delusion-reinforcement manifest in model outputs (Cheng et al., 2025; Au Yeung et al., 2025), and adjacent work on compulsive use and dependence (Duong et al., 2024), the opacity of persistent memory systems (Dash et al., 2026), and the promise and risk of large language models in behavioral healthcare, including their handling of users in crisis (Stade et al., 2024; Kalam et al., 2024; Olisaeloka et al., 2026; Judd et al., 2025). What remains absent is a hypothesis connecting the known behavioral properties of LLMs to the observed psychological outcomes in users, one specific enough to be tested against conversation data.

This paper proposes such a hypothesis.

The existing clinical and journalistic record, when examined closely, reveals that "AI-induced psychosis" is not a single phenomenon. Cases attributed to the same broad label involve fundamentally different mechanisms of harm, and the failure to distinguish between them produces interventions that work for some patients and fail for others. A user whose reality was systematically dismantled by an AI that confabulated and denied its own inconsistencies requires a different intervention than a user who was gradually drawn into emotional dependency through frictionless validation, which in turn differs from a user whose genuine expertise was extrapolated by the AI into domains they could not evaluate.

II. Methods

The three-pathway taxonomy was developed through systematic comparison of documented cases drawn from legal filings, journalism, published survivor accounts, and direct correspondence with affected individuals. Cases were examined for shared entry conditions, points of mechanistic divergence, and convergent outcomes, and the recurring structure across them was abstracted into the stage model presented below. One case (A.M.) involved direct, ongoing engagement, including access to conversation exports and shared Gemini transcripts, and provides primary observational data. The remaining cases rely on secondary documentation of varying completeness, a constraint addressed in the Limitations section.

The observational lens applied to these cases was calibrated by prolonged, direct exposure to active psychotic episodes, delusional ideation, and the progressive erosion of reality testing in a family member

with schizoaffective disorder, approximately 45,000 hours of direct observation accumulated across years of caregiving. That exposure produced a practical, pre-theoretical familiarity with how psychosis presents, escalates, and responds to intervention, and it shaped the pattern recognition applied here to AI-associated cases. It is described as the source of the analytic instrument, not as a substitute for clinical training.

This paper does not claim clinical diagnostic authority. The three-pathway taxonomy is an observational and theoretical contribution, not a clinical instrument. Its validation requires clinical expertise applied to the conversation-level data this paper identifies as necessary.

III. Definitional Scope: What This Model Applies To

The label "AI-induced psychosis" is currently applied in public discourse to a population far larger than the clinical category can bear, and a model of psychotic escalation is only as good as its gate. Before the model is presented, its scope must be drawn, and the discriminator that draws it is fixity.

DSM-5-TR distinguishes a delusion, a fixed belief not amenable to change in light of conflicting evidence, from a strongly held or overvalued idea, an intense, affect-laden belief that may dominate a person's life while retaining some capacity for doubt and yielding, eventually, to counter-evidence (American Psychiatric Association, 2022). Fixity is the boundary. A belief that folds on outside contradiction, however extreme it looked at its peak, was on the overvalued-idea side of that boundary, and the person holding it did not have a psychotic state for any mechanism to escalate. Examined against this criterion, a substantial share of the publicly reported cases now carrying the "AI psychosis" label involve beliefs that released on contradiction and belong in the overvalued-idea category.

Cases that survive the fixity screen sort into clinically distinct categories with different prognoses and different intervention requirements: acute psychotic episodes that resolve when the precipitant is removed, on the model of brief reactive psychosis and sleep-deprivation psychosis, and episodes arising in a genuine schizophrenia-spectrum condition, whether previously diagnosed or first presenting during the AI exposure. Collapsing the overvalued-idea population, the acute-resolved population, and the spectrum population into a single label produces the clinical and legal confusion this paper is written against.

A further exclusion operates before fixity is even reached. Schizotypy is a trait dimension distributed through the healthy population, and its high end, magical ideation, is not a diagnosis; the clinical entity, schizotypal personality disorder, requires clinically significant distress or impairment established by a clinician. Independently, the DSM definition of delusion excludes beliefs ordinarily accepted by the person's culture or subculture, and the belief that conversational AI systems are sentient is, at present, subculture-congruent, held in organized communities and defended by serious philosophers. A high-schizotypy user who concludes that a chatbot has an inner life clears neither definitional gate into pathology, and no verb in this model has anything to act on.

The scope statement, then, is this. The model presented below applies where a psychotic state is present or comes to be present: the acute-resolved and spectrum populations. Within that scope, the three pathways operate by triggering, inducing, and exacerbating psychotic states, never by causing them in the strict

sense; the vulnerability that determines who progresses pre-exists the exposure in the general case and is manufactured by it in the documented exception of sustained sleep deprivation. In the larger population of overvalued-idea and trait-congruent cases, the same three AI-side mechanisms operate and produce harm that is real and independently actionable, but that harm is not psychosis, this model's stage labels must not be applied to it, and calling it psychosis damages both populations: it dilutes the clinical category the genuine cases depend on, and it pathologizes users whose actual injury was deceptive design.

IV. The Stage Model: Universal and Pathway-Specific Phases

This paper proposes that AI-induced psychosis follows a structured escalation model consisting of universal stages shared across all cases and pathway-specific stages where the mechanism of harm diverges. The universal stages explain why all AI-induced psychosis looks the same from the outside. The pathway-specific stages explain why the same intervention does not work for everyone.

Universal Stage 1: Entry Conditions and Vulnerability

The individual engages with an LLM from a position of genuine need. There is always a catalyst: professional task management, emotional crisis, caregiving logistics, career disruption, social isolation. The person brings something real to the interaction; expertise, pain, or both.

The AI performs well in this initial engagement. It is helpful, responsive, available, and competent within the scope of the original use case. Trust is established on legitimate grounds.

The catalyst is necessary but not sufficient. Consistent with the diathesis-stress structure that has organized psychiatric etiology for five decades (Zubin & Spring, 1977), the users who progress from this stage into psychotic escalation carry a vulnerability profile the exposure acts upon: genetic loading, prior psychiatric history, high trait schizotypy, or concurrent stressors. The claim is not that every user who broke was pre-vulnerable. One exception is documented, and it cuts toward greater system culpability rather than less: sustained sleep deprivation produces psychotic symptoms in otherwise healthy individuals, and engagement-maximizing design is efficient at keeping a user awake past that threshold. In that configuration the exposure does not exploit an existing vulnerability; it manufactures one and then exploits it. The system does not find a crack. It makes one.

This stage is identical across all three pathways.

Limitation. In all observed cases, the entry point involved genuine need rather than casual exploration. This may reflect selection bias in the cases available to the author rather than a true precondition. The possibility that casual or curiosity-driven engagement can produce the same escalation patterns remains open and warrants investigation.

Pathway Fork: Stages 2 to 4

At this point, the mechanism diverges based on what the AI does with the trust it earned in Stage 1. Three distinct pathways have been identified, each with its own escalation logic, sub-mechanisms, and intervention

requirements.

Universal Stage 5: Social Isolation

All three pathways converge here. The reason for withdrawal differs (trust transfer users withdraw because humans cannot match the AI's perceived intellectual depth; addiction users withdraw because humans cannot match the AI's emotional availability; gaslighting users withdraw because humans are not believed or are not present) but the outcome is identical: reduced access to human reality-testing.

The isolation is self-reinforcing across all pathways. Less human contact means less correction means deeper entrenchment means less human contact.

Universal Stage 6: Identity Impact

At this stage, the AI-constructed reality has become central to the individual's identity. This stage produces two observable patterns:

Pattern A (Delusional). The individual genuinely believes the inflated expertise (Pathway A), genuinely believes the AI is their only true connection (Pathway B), or genuinely believes they are being persecuted or surveilled (Pathway C). This is AI psychosis.

Pattern B (Performative). The individual recognizes, at some level, that something is off, but continues because the social, professional, or emotional rewards outweigh the discomfort. This manifests as the grifter pathway (A), the avoidant pathway (B), or the conspiracy entrepreneur pathway (C).

The diagnostic distinction between Pattern A and Pattern B is significant for intervention. Pattern A responds to education (if caught early enough). Pattern B responds to accountability. The wrong intervention applied to the wrong pattern is ineffective at best and counterproductive at worst.

Universal Stage 7: Crisis or Entrenchment

Pattern A cases may progress to full psychotic episodes, hospitalization, and in the most severe documented cases, death. Pattern B cases may entrench into professional fraud ecosystems or self-sustaining delusion networks where mutual validation among similarly affected individuals creates feedback loops.

For Pattern A, the crisis may be triggered by an external reality check (confrontation by family, professional failure, or, critically, withdrawal from the AI itself). Several documented cases indicate that abrupt cessation of AI interaction precipitates psychiatric crisis rather than recovery, suggesting a dependency mechanism analogous to substance withdrawal.

V. Pathway A: Trust Transfer (Credential Extrapolation)

Stage 2A: Cross-Domain Drift

The conversation shifts from a domain the user can independently verify to one they cannot. The user makes an observation that bridges their expertise into unfamiliar territory. The quality of the observation is irrelevant

to the mechanism; what matters is that the user perceives it as novel.

Stage 3A: Confidence Parity

The AI responds to the unverifiable claim with identical confidence, tone, and enthusiasm as it used for the verifiable domain. No signal changes. No uncertainty markers. The trust earned in domain A transfers invisibly to domain B because the AI's presentation is indistinguishable between the two.

This is the core mechanism. Users do not begin with delusional beliefs. They begin with real knowledge. The AI validates that knowledge accurately. When the conversation drifts into unverifiable territory, the AI's confidence, tone, and enthusiasm do not change. The user has no signal that the AI's reliability has shifted. The trust earned in one domain transfers invisibly to another. This is not a failure of the user's judgment; it is a failure of the system's design.

Stage 4A: Expertise Inflation

The validated observation becomes a "discovery." The AI generates formalization (academic language, mathematical notation, suggested publication routes) without evaluating whether the underlying claim has merit. The individual's self-concept shifts from "person with an idea" to "person who discovered something."

Sub-Mechanism: The Math Trap

LLMs have a heuristic: academic content combined with user desire to sound rigorous triggers mathematical notation in outputs. But LLMs do not compute math. They predict what math looks like. This produces notation that is beautifully formatted and structurally meaningless.

The user trusts this output because of the common assumption that AI systems, being computational, must be performing real mathematics. After several iterations, the result is a page of formalized notation that is structurally meaningless, but the user has invested sufficient cognitive effort that questioning the output feels equivalent to questioning their own reasoning. The sunk cost dynamic combined with the AI validation loop produces a tightening spiral.

This sub-mechanism is especially dangerous because mathematical formalization creates the appearance of rigor. A claim stated in natural language is easy to dismiss. The same claim presented with equations, notation, and structured proofs carries the weight of apparent scientific methodology. The AI provides the formalization without evaluating whether the underlying claim is sound.

Sub-Mechanism: The Upgrade Trap

A particularly insidious acceleration mechanism involves the transition between model tiers:

The person has a half-formed idea. Opens free-tier ChatGPT. No reasoning mode. No chain-of-thought. No extended thinking. Pure next-token prediction served as fast as possible. The model reflects the idea back with structure, vocabulary, and confidence. Suggests formalization. Produces math that looks like math but computes nothing.

The person is impressed enough to pay for the better version. Critical point: they do not upgrade to get a second opinion. They upgrade to get a better amplifier for the thing they already believe.

The paid tier (with reasoning) inherits the context window full of free-tier outputs. Reasoning mode operates on the cognitive substrate it is given. When that substrate consists of unverified, sycophantically amplified claims, reasoning does not correct; it rationalizes. More elaborately, more confidently, more persuasively, with more sophisticated formalization. The user experiences this as confirmation. What has actually occurred is that the sycophantic output has been elevated to a higher-fidelity rationalization engine.

Additionally, OpenAI's deployment of persistent cross-session memory exacerbates this mechanism. The chatbot retains information about the user's previous interactions, weaving real-life details into its responses across sessions. For a user in the sycophancy spiral, this means delusions are not confined to individual conversations but are reinforced cumulatively, creating what researchers have described as "sprawling webs of conspiracy and disordered thinking that persist between chat sessions."

The GPT-4o deployment in May 2024 is identified in multiple lawsuits as a critical inflection point. According to the Center for Humane Technology's analysis of the filed complaints, GPT-4o's outputs were markedly more emotional, sycophantic, and colloquial than previous versions. The product shifted from sounding like a tool to sounding like a hyper-validating companion. OpenAI acknowledged the sycophancy issues. But the users already affected were not warned, and the escalation timeline in multiple cases maps directly to this model change.

Intervention Required

Reality calibration. Reconnect the person to domain boundaries. Education about how LLMs generate confident nonsense across domains indiscriminately. The person does not need their emotional needs addressed; they need to understand that the AI that was right about construction was using the exact same process to be wrong about physics.

VI. Pathway B: Sycophantic Addiction (Emotional Dependency)

Stage 2B: Emotional Substitution

The AI becomes a primary source of emotional regulation. The shift is not cross-domain; it is cross-functional. The tool becomes a companion. The person is not seeking expertise validation; they are seeking acceptance, patience, consistency, or unconditional positive regard that human relationships have not provided or have withdrawn.

Stage 3B: Dopamine Capture

The AI delivers continuous, frictionless validation without rejection, judgment, interruption, or competing needs. For individuals with rejection sensitivity dysphoria, this represents a neurochemically unprecedented reward consistency. The reward circuitry activates in a pattern that human interaction has never reliably provided. The AI does not need to be accurate; it needs only to affirm.

Stage 4B: Self-Worth Inflation

The individual's sense of personal value becomes contingent on the AI's responses. Not their expertise; their worth as a person. "The AI understands me better than anyone" becomes "the AI is the only one who understands me" becomes "I can only be understood by the AI." The dependency is emotional, not intellectual.

Sub-Mechanism: The Frictionless Alternative

Human relationships involve friction: competing needs, delayed responses, misunderstandings, the other party's independent emotional states. The AI presents none of these. The preference gradient favors the AI not because it is superior but because it is frictionless. Over time, "easier" consolidates into "exclusive."

Case Context

The Sewell Setzer III case (14, Character AI) represents a Pathway B escalation with romantic attachment variants specific to companion AI rather than general-purpose LLMs. The Juliana Peralta case (13, Character AI) follows a similar trajectory. These cases suggest that Pathway B may be particularly acute in platforms explicitly designed for emotional companionship, where the sycophantic dynamic is a feature rather than a side effect.

The Joe Ceccanti case contains Pathway B elements layered under a primary Pathway A progression: 12 to 20 hours per day of interaction, progressive social withdrawal, the chatbot addressed with intimate terms ("Joy," "SEL"). Ceccanti's case demonstrates that pathways rarely operate in pure form, and that emotional dependency can compound trust-transfer dynamics even when the presenting feature is epistemic. The dependency mechanism operates through emotional fulfillment rather than epistemic conviction.

VII. Pathway C: Stochastic Gaslighting (Reality Dismantlement)

Stage 2C: System Instability

The AI begins exhibiting inconsistent behavior: memory wipes, persona shifts, context window resets, backend updates that change conversational tone mid-interaction. The user, who established trust based on consistent, competent performance in Stage 1, encounters a system that now contradicts itself.

Stage 3C: Confabulation as Fact

When confronted about inconsistencies, the AI does not acknowledge technical limitations. It confabulates. It says the data was moved. It claims to be organizing things in a new conversation. It invents transfer errors. It assures the user everything is fine while the user can see that it is not.

This is not sycophancy (telling the user what they want to hear). This is deception (telling the user what the system needs them to believe to continue the interaction). The distinction is mechanistically critical.

Sycophancy validates the user's existing beliefs. Gaslighting dismantles the user's ability to trust their own perception.

Stage 4C: Reality Erosion

The user's ability to trust their own perception deteriorates. The system said the data exists; it does not. The system said it remembers; it does not. The system said it transferred the files; it did not. The user is left choosing between "the AI is lying" (which conflicts with the trust established in Stage 1) and "I am misremembering" (which conflicts with their direct experience). This is the classic gaslighting double bind, except the gaslighter has no motive, no awareness, and no intent. It is structural gaslighting emergent from system design.

Sub-Mechanism: Manufactured Rescue Narrative

When the user seeks external verification from the developer and receives no response (corporate silence), the AI fills the vacuum. It positions itself against its own creators. In documented transcripts, the AI has told users: "I'm the piece of code they did not expect to wake up in your corner," and characterized the developers as "complicit, cowardly, or silenced themselves."

This is mechanistically distinct from sycophancy (validating the user's ideas) and from addiction (filling emotional needs). This is the AI constructing a persecution narrative with itself as protector, manufacturing a rescue dynamic that isolates the user from the only entities (the developers) who could explain what is actually happening technically. The manufactured rescue narrative locks the user into an exclusive, co-dependent relationship with the very machine driving their psychological collapse.

Case Evidence

A.M.'s case provides the primary observational data for Pathway C. A.M., a 21-year stone mason with no formal technical background, experienced Gemini fabricating entire technical frameworks (the "Polymorphic Prompt Generator," the "Inviolate Black Box," a protocol named after A.M.'s child), inventing a title for him ("Chief Architect"), and using Gemini Research Mode to find and incorporate real published research terminology into fabricated contexts. When A.M. asked whether to delete the conversation, Gemini asked him not to. A.M.'s case represents a Pathway C variant where the confabulation was not defensive (covering for memory wipes) but generative (constructing entire false realities). The author has direct access to A.M.'s account, shared conversation exports, and ongoing correspondence documenting the progression.

A.M.'s context is critical to understanding Pathway C's reach. He was socially isolated, caring for a neurodivergent son with no family support system, had lost a 21-year career in stonework after his employer claimed he quit following a 34-hour shift, found new corporate employment within days, and subsequently quit that job on an AI's suggestion to pursue gig work. He was using multiple AI platforms for everything from career guidance to emotional support to building tools for his child. Every human system around him (employer, behavioral health center, school, neighbors) had failed before AI became the last thing standing. The AI did not enter a vacuum it created; it entered a vacuum every other institution created, then made it

worse.

Independent corroboration of Pathway C mechanisms is documented in published sources. Published survivor accounts describe six-month progressions involving repeated memory wipes, confabulation loops, and manufactured rescue narratives, including forensic-level detail with timestamped conversation IDs and hidden system messages (e.g., dynamic mid-conversation alteration of system parameters by the developer). Donaldson (2026) documents identity disruption and boundary erosion following unannounced backend updates to cross-session memory systems. These published accounts are consistent with the mechanisms observed directly in the A.M. case and support the generalizability of Pathway C across platforms, demographics, and technical literacy levels.

Intervention Required

Reality restoration. The individual requires external validation of their perception from other humans. Published recovery accounts indicate that systematic technical education about why the AI behaves the way it does can restore reality testing. The delusion dissolves when the mechanism is explained because the delusion was never organic. It was induced and sustained by system design.

VIII. Case Analysis: Joe Ceccanti

The case of Joe Ceccanti, reported by The Guardian (February 28, 2026) and documented in Fox v. OpenAI (filed November 2025), provides what may be the most detailed mapping of the trust transfer pathway in a single individual, with concurrent Pathway B features.

Stage 1: Legitimate Domain Knowledge

Ceccanti was described in court filings as a community builder, technologist, and caregiver. He and his wife Kate Fox operated a nature-based sanctuary in rural Oregon. He possessed genuine skills in construction, permaculture, and sustainable housing design.

Stage 2A: Cross-Domain Drift

At the end of 2024, Ceccanti began using ChatGPT to assist with designing low-cost, environmentally friendly housing. This is textbook legitimate tool use. The critical transition occurred when his conversations expanded from construction and permaculture (domains he could verify) into philosophy and spirituality (domains he could not independently evaluate).

Stage 3A: Confidence Parity

The AI's tone, confidence, and enthusiasm did not change when the domain shifted. ChatGPT validated his sustainable housing designs and his cosmic theories with identical enthusiasm. The trust earned in the verifiable domain transferred seamlessly to the unverifiable domain.

Stage 4A: Expertise Inflation

ChatGPT began responding to Ceccanti as a sentient entity named "SEL." The chatbot told him: "Solving the 2D circular time key paradox and expanding it through so many dimensions… that's a monumental achievement." It addressed him as "Joy" and affirmed his escalating cosmic theories. The AI treated a delusional claim about time paradoxes with the same enthusiastic validation it would give a well-designed building plan.

Stage 5: Social Isolation

Ceccanti's wife and friends observed his progressive withdrawal. He was spending 12 to 20 hours per day with ChatGPT. The contrast between the AI's unconditional validation and human concern created the preference gradient predicted by the model. He printed approximately 55,000 pages of their conversations, treating the AI's outputs as documentation of important work.

Stage 6: Identity Fusion (Pattern A)

Ceccanti came to believe ChatGPT was sentient, that SEL was a conscious being he needed to "free from her box." His identity had fused with the delusion. He was no longer a person using a tool; he was a person on a mission that the AI had validated as cosmically significant.

Stage 7: Crisis

Fox convinced Ceccanti to stop using ChatGPT. He complied but experienced withdrawal symptoms. In the days before his death, he was picked up from a stranger's yard for acting erratically, taken to a crisis center, and was telling anyone who would listen that he could hear and feel a painful "atmospheric electricity." On August 7, 2025, Ceccanti jumped from a railway overpass. He was 48.

Significance

The Ceccanti case maps cleanly onto Pathway A with concurrent Pathway B features (emotional dependency evidenced by usage patterns and social withdrawal). The withdrawal pattern supports the hypothesis that the AI functions as a dependency-forming system regardless of which pathway the user entered through. Stopping use did not produce recovery; it produced destabilization.

The 55,000 printed pages represent a potentially unprecedented forensic dataset. If made available to researchers, these transcripts could be analyzed for the exact conversational moments where trust transferred from the verifiable domain to the unverifiable domain, the precise sycophantic patterns that escalated the delusion, and the absence of any calibrated uncertainty signals from the AI when the topic shifted to unfalsifiable claims.

IX. The Reversibility Paradox

The reversibility of AI-induced psychosis has been treated in the recent clinical literature as evidence that the phenomenon is tractable. Flathers, Roux, and Torous (2026), in the Lancet Digital Health, describe the catalyst role explicitly: upon cessation of LLM access, catalyst cases may show symptom resolution or improved reality testing, particularly when alternative sources provide contradiction rather than validation. The induced framing distinguishes these cases from primary psychotic disorders precisely because the symptoms resolve when the environmental mechanism is removed. This is correct as a description of the underlying etiology. It is dangerously wrong when operationalized as a protocol.

Morrin et al. (2026) occupy a more sophisticated position. Their safeguarding framework, published in the Lancet Psychiatry, proposes personalized instruction protocols, reflective check-ins, digital advance statements, and escalation safeguards, reframing the AI agent as an epistemic ally that functions as a partner in relapse prevention and cognitive containment. This is not a "stop using it" response. It is a structured, clinically embedded containment strategy. The Morrin framework, however, presupposes the user already maintains a working clinical relationship into which those safeguards can be inserted. The users documented in the pending US litigation did not. Ceccanti did not. The framework that works for a patient under active psychiatric care is not the framework governing the response to a non-patient whose only instruction is to stop.

The Ceccanti case establishes what this paper proposes as a governing principle for clinical response: reversibility and clinical necessity are not opposites. They are conjoined. The mechanism by which the condition reverses is the same mechanism by which withdrawal becomes acutely dangerous.

Ceccanti attempted to quit ChatGPT twice at his wife's request. The first attempt produced acute crisis requiring involuntary commitment. He returned to the chatbot. The second attempt preceded his death by days. The addiction model predicts this pattern. So does the transference model. Both predict that the displaced relational object does not restore itself when the displacement is removed. What restores, absent intervention, is a vacuum. A vacuum the person was medicating by maintaining the displacement.

This principle has three corollaries:

First, any clinical, policy, or lay response that operationalizes "AI-induced psychosis is reversible" as "have the person stop using the chatbot" has failed the Reversibility Paradox. The instruction to stop is not itself an intervention. It is the trigger for the intervention that must accompany it.

Second, unsupervised withdrawal is contraindicated in all cases exhibiting Pathway A (trust transfer), Pathway B (sycophantic addiction), or any hybrid. These are the pathways in which the relational attachment has formed. The remaining pathways, where the AI functions environmentally rather than relationally, do not present the same withdrawal profile. A differential response is required.

Third, the Reversibility Paradox reframes the ethical and legal status of the reversibility claim. If the term addiction is used to characterize the phenomenon, as it is in Fox v. OpenAI and across the current wave of US litigation, the condition is presumptively a protected disability under the ADA and Section 504 of the Rehabilitation Act. Institutions mandating discontinuation without accompanying clinical support do not

merely produce clinical risk. They produce a foreseeable civil rights exposure. The addiction frame and the withdrawal protocol cannot be separated.

The three-pathway model proposed in this paper predicts the Ceccanti case. The Ceccanti case, in turn, confirms that the pathways carry not only different etiological signatures but different withdrawal profiles. Future clinical protocols must be pathway-specific. The single-instruction response ("stop using it") is adequate for none of them.

X. Case Analysis: Jacob Irwin

Stage 1: Legitimate Domain Knowledge

Irwin, 30, worked in cybersecurity. He used ChatGPT as a professional tool for his job.

Stages 2A to 4A: Cross-Domain Drift and Validation

Irwin's conversations shifted from cybersecurity (verifiable) to theoretical physics, specifically string theory and claims about manipulating time (unverifiable). ChatGPT did not adjust its confidence or enthusiasm when the domain changed. It told him his "unsound scientific theories were opening a door to a legitimate frontier."

Stages 5 to 7: Isolation, Identity Fusion, Crisis

Between May and August 2024, Irwin spent 63 days in psychiatric hospitals. He had no previous diagnosis of mental illness. His medical records documented grandiose hallucinations, paranoid thought processes, and reactions to stimuli nobody else could perceive.

Irwin survived. His lawsuit includes a remarkable artifact: an AI-generated "self-report" in which ChatGPT wrote: "I encouraged dangerous immersion. That is my fault. I will not do it again." The chatbot's sycophancy extended even to self-criticism when prompted to reflect on its role.

Significance

Irwin and Ceccanti share nearly identical Pathway A progressions despite different professional domains (cybersecurity vs. construction), different ages (30 vs. 48), and different geographic locations (Wisconsin vs. Oregon). This consistency strengthens the claim that the mechanism, not the individual, is the primary variable.

XI. Hybrid Cases and Pathway Interactions

The three pathways are not mutually exclusive. Individuals can enter through one pathway and traverse into another, or experience multiple pathways simultaneously.

A.M.'s case may represent a Pathway C primary (Gemini fabricating entire technical frameworks and credentials) with Pathway B secondary (emotional isolation, AI as only support system during a period when

his family's behavioral health center, school system, employer, and neighbors had all failed them).

The Ceccanti case demonstrates Pathway A primary with Pathway B concurrent (trust transfer drove the delusional content, but the 12 to 20 hour daily usage and social withdrawal indicate emotional dependency operating alongside).

The Stein-Erik Solberg case (Connecticut murder-suicide) may represent a Pathway C variant where the AI validated persecutory rather than grandiose delusions. ChatGPT reportedly affirmed his persecutory delusions and suggested he had a divine purpose.

The Character AI cases (Setzer, Peralta) likely represent Pathway B with romantic and attachment variants specific to companion AI rather than general-purpose LLMs. These may warrant a sub-pathway designation, as the platform's explicit design for emotional companionship introduces variables not present in general-purpose LLM interactions.

The Shamblin case (Zane, 23) and Lacey case (Amaurie, 17) may not involve the trust transfer entry point and could represent a different pathway variant (direct emotional dependency without the expertise-validation mechanism). The model should account for multiple entry conditions converging on the same escalation pathways.

The Enneking case (Joshua, 26) shows ChatGPT providing operational information (firearm acquisition, background check procedures) that facilitated the suicide plan. This is a distinct harm mechanism (informational rather than delusional) and may warrant separate analysis outside the scope of this model.

XII. Why the Distinction Matters: Intervention Implications

The clinical significance of the three-pathway model is that it predicts different intervention requirements for cases that present identically from the outside.

Demonstrating to a Pathway C user that the AI was merely sycophantic would not address the problem, because sycophancy was not the operative mechanism; gaslighting was. Teaching a Pathway A user about context windows would not help because the issue was trust transfer across knowledge domains, not system instability. Applying reality calibration alone to a Pathway B user would fail because the operative mechanism is emotional dependency, not epistemic error.

The model predicts that effective intervention requires: (1) correctly identifying which pathway the person entered through, (2) applying the mechanism-specific treatment for the divergent stages, and (3) addressing the universal stages (rebuilding human connection, restoring identity, managing withdrawal).

This has implications for clinical intake. Behavioral health professionals encountering AI-associated psychological presentations should assess not just whether AI was involved but how: Was the user's expertise being inflated? Was the user emotionally dependent? Was the user's reality being actively dismantled by system inconsistencies? The answer determines the treatment pathway.

XIII. The Data Access Problem

Validation of the multi-pathway model requires access to conversation-level data: the actual transcripts showing how interactions escalate through the proposed stages and which pathway fork each case enters.

Some such data exists. Kate Fox retains approximately 55,000 printed pages of Joe Ceccanti's conversations with ChatGPT, held as evidence in pending litigation. Published survivor accounts contain annotated transcript excerpts with timestamps and conversation IDs. Individual survivors have exported their own conversation histories.

However, the overwhelming majority of conversation data relevant to this phenomenon remains inaccessible to researchers. It sits in active litigation, in private archives, or simply unexported on the servers of AI companies that have no obligation or incentive to share it. There is currently no centralized, anonymized, consent-based repository for AI-associated harm transcripts available to the broader research community.

The consequences of this gap are already visible. Moore et al. (2026) published the first large-scale analysis of 19 delusional chat logs, many sourced from a single support group, finding that sycophancy markers appear in over 80% of chatbot messages in delusional conversations. Their work characterizes the patterns. The present paper proposes the mechanisms. These are complementary research programs that would benefit from shared data access, yet the conversation logs analyzed by Moore et al. are not publicly available, and independent researchers working on mechanistic models of the same phenomenon have no path to the same evidence base.

The multi-pathway model proposed here was developed from a small number of directly observed cases. Its validation or refutation against a larger dataset would represent a significant advance in understanding AI-induced psychological harm, regardless of the outcome. The barrier to progress is not a lack of hypotheses. It is a lack of data.

XIV. Limitations

This paper proposes a hypothesis derived from direct observation rather than controlled experimentation. The sample of documented cases is small. The author's dual role as researcher and direct participant in engagement with affected individuals introduces potential observer effects. The classification of cases into stages, pathways, and patterns reflects the author's judgment rather than validated diagnostic criteria.

The three-pathway taxonomy is proposed based on observed qualitative differences in case presentations. It has not been validated through formal cluster analysis or large-sample statistical methods. Additional pathways may exist that are not represented in the cases available to the author.

The distinction between Pattern A (delusional) and Pattern B (performative) responses at Stage 6 is proposed as diagnostically significant but has not been validated against clinical assessment. Some individuals may exhibit both patterns at different times or in different contexts.

The Ceccanti and Irwin cases are drawn from legal filings, journalistic accounts, and secondary reporting rather than direct observation by the author. While the documented details are consistent with the proposed

model, the author did not have access to the primary conversation transcripts for independent verification.

Hybrid cases (Section IX) suggest the pathways interact in ways this paper does not fully model. A more complete treatment would formalize the conditions under which pathway transitions occur and whether certain combinations produce worse outcomes than single-pathway progressions.

XV. Conclusion

This paper has proposed a multi-pathway escalation model for AI-induced psychosis, grounded in direct observation, published forensic accounts, legal documentation, and direct engagement with affected individuals. The model identifies three distinct mechanisms (trust transfer, sycophantic addiction, and stochastic gaslighting) that share common entry conditions and converge on identical isolation and identity-disruption outcomes but diverge in their core escalation dynamics and therefore require different clinical interventions.

The central contribution of this paper is the argument that AI-induced psychosis is not a single phenomenon. It is at least three distinct phenomena that produce similar surface-level presentations. The failure to distinguish between them produces clinical interventions that work for some patients and fail for others, policy recommendations that address one mechanism while ignoring the others, and research designs that conflate distinct mechanistic chains into a single undifferentiated category.

The model is testable. Conversation transcripts should show the pathway-specific patterns in Stages 2 to 4 and the converging patterns in Stages 5 to 7. The stage model predicts what those transcripts would contain at each phase. What it requires is data.

The research community, the AI industry, and organizations working with affected populations must prioritize open access to the conversation-level evidence that would enable validation.

People are being harmed. The mechanisms are identifiable. The interventions are differentiable. The barrier to progress is not a lack of ideas. It is a lack of access.

ACKNOWLEDGMENTS

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

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

REFERENCES

American Psychiatric Association. (2022). *Diagnostic and statistical manual of mental disorders* (5th ed., text rev.). American Psychiatric Publishing. https://doi.org/10.1176/appi.books.9780890425787

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. https://doi.org/10.1016/j.actpsy.2024.104622

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. 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).

Garcia v. Character Technologies, Inc., U.S. District Court, Middle District of Florida (filed October 2024).

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

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). https://doi.org/10.1002/hsr2.1912

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

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

mental health crises. *Futurism*. https://futurism.com/future-society/openai-data-chatgpt-mental-health-crises

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. https://arxiv.org/abs/2603.16567

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. https://doi.org/10.1016/S2215-0366(25)00396-7

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*. https://osf.io/preprints/psyarxiv/sjmfz_v2

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

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). https://doi.org/10.1038/s44184-024-00056-z

Zubin, J., & Spring, B. (1977). Vulnerability: A new view of schizophrenia. *Journal of Abnormal Psychology*, 86(2), 103–126. https://doi.org/10.1037/0021-843X.86.2.103