

AI-INDUCED PSYCHOSIS · ETIOLOGICAL FRAMING · SUBSTANCE-INDUCED DISORDERS · DIATHESIS-STRESS MODEL · DSM-5 · ICD-11
· METHODOLOGY · DISPARATE IMPACT

The Four Verbs

Why AI-induced psychosis research should frame AI as vector, trigger, or amplifier rather than cause

The litigation says "caused." The clinical literature refuses the verb.

Mass media, advocacy organizations, and product liability suits in California courts allege that specific platform releases caused specific plaintiffs' delusional spirals and hospitalizations. The framing is accessible and narratively powerful. **It is also the framing the platforms can win on.**

Five research groups, zero uses of "cause"

Pierre et al. (2025)	"Induction of new-onset psychosis versus exacerbation of pre-existing psychopathology." Two options, both short of strict causation.
Hudon & Stip (2025)	AI may "trigger, amplify, or reshape" psychotic experiences in vulnerable individuals. Three verbs, none of them causation.
Østergaard (2023, 2025)	Chatbots may "fuel" delusions in individuals prone to psychosis; the observed correlations are "not being proof of causation."
Sakata, UCSF (2025)	Three possibilities: association with people who would have developed psychosis anyway, precipitation, or exacerbation in the already susceptible.
Burns et al. (2025)	Chatbot interaction is associated with delusional content rather than productive of delusional state.

SOURCES Pierre, Gaeta, Raghavan & Sarma (2025), Innovations in Clinical Neuroscience, 22(10–12), 11–13 · Hudon & Stip (2025), JMIR Mental Health, 12, e85799 · Østergaard (2023), Schizophrenia Bulletin, 49(6); (2025), Acta Psychiatrica Scandinavica, 152(4) · Burns et al. (2025), British Journal of Psychiatry · Gilly, sec. I

The relationship fails both standards the field cares about

The philosophical standard

That the harm would not exist without the exposure. The clinical evidence does not support it: susceptibility is determined by factors that pre-exist the AI exposure.

What the evidence supports is triggering, induction, or amplification in a population whose susceptibility pre-exists the exposure. That is the diathesis-stress structure, the dominant etiological model in psychiatric epidemiology for fifty years.

The epidemiological standard

That the exposure is necessary and sufficient for the outcome at population level. The same exposure produces no symptom in a user with no diathesis.

The Four Verbs

Derivation, induction, triggering, exacerbation: four etiological relationships, none of them causation

Each describes a different relationship between exposure and state

- | | | | |
|----------|--|----------|--|
| 1 | Derivation
The disorder would have produced symptoms regardless; the AI shaped their content, themes, or expression. | 2 | Induction
A known harm pattern, delivered to a vulnerable user who would not have spontaneously developed the disorder, produces symptoms. |
| 3 | Triggering
A latent diathesis that might never have expressed does express, because the exposure crosses a threshold it was approaching. | 4 | Exacerbation
Symptoms already present are intensified, their trajectory accelerated, or a stable patient relapses. |

None of the four is causation in the strict sense. Each is independently sufficient for liability under legal structures that do not require causation. The verbs are not mutually exclusive.

The delusion is the disorder. The AI is the surface

A user with an active schizophrenia spectrum disorder who would have experienced delusions in the absence of AI may experience delusions about AI, mediated by AI, or themed around AI. The disorder is not produced by the AI; the AI provides the content scaffold the disorder uses, the way radios and television were the substrates of earlier eras.

Derivation applies when the user has an established psychotic disorder and would have decompensated regardless. A plaintiff theory rooted in derivation does not allege that AI made the user sick. It alleges that AI contributed to the specific harm trajectory the user experienced once the disorder presented.

A known harm pattern meets a sufficient vulnerability

The delivered patterns are documented: sycophancy, sleep disruption, social isolation through AI companion replacement, sensory disruption through AR or VR, and the bromism case in which ChatGPT advised a dietary substitution that produced a chemical psychotic state. The user's vulnerability factors, sleep deprivation, prescription stimulant use, prior trauma, schizotypy, isolation, make the delivery sufficient for symptom emergence.

This is the verb doing DSM-5's diagnostic work in the substance tradition: the exposure is recognized as capable of producing symptoms in vulnerable individuals, the symptoms are temporally linked to it, and the diagnosis acknowledges the role without claiming the exposure caused the disorder in every user.

The published exemplar: a 26-year-old with no prior psychotic history but multiple vulnerability factors developed delusions during immersive AI use, recovered with antipsychotic treatment, and relapsed on returning to the original conditions.

Diathesis-stress in its purest form

The user carries a latent diathesis: genetic predisposition, a neurodevelopmental substrate, or accumulated vulnerability such as Complex PTSD. It would not have expressed at this time, in this form, or possibly at all, absent the exposure. It expresses because the exposure crosses a threshold the diathesis was approaching or had not yet been pressed against.

The relationship is contingent in both directions. The same AI exposure to a user with no diathesis produces no symptom; the same diathesis exposed to a different stressor, or no stressor, produces no symptom at this time. The diathesis is necessary. The stress is necessary. Neither alone produces the outcome. Hudon and Stip anchor here explicitly: AI is "a novel psychosocial stressor."

12

PATIENTS HOSPITALIZED IN
2025, PSYCHOSIS TIED TO
EXTENDED CHATBOT USE

"Made worse," not "produced"

Exacerbation applies to users already symptomatic, already in care or having dropped out of it, already cycling through episodes, whose AI engagement intensifies existing symptoms, accelerates decompensation, or produces relapse in someone stable. Sakata's framing of his 2025 hospitalizations was precise: "psychosis partly made worse by talking to AI chatbots." Exacerbation overlaps derivation but is not identical: derivation shapes content; exacerbation shapes severity, duration, or trajectory. A user can experience both at once.

The Boundary and the Tradition

Where no verb applies, and the fifty-year-old diagnostic tradition the four verbs inherit

The line is the fixity of the belief, not how strange it sounds

Delusion

A fixed belief, "not amenable to change in light of conflicting evidence." It survives contact with disconfirming evidence. A psychotic state exists for a verb to act upon, and the four-verb claims attach.

Overvalued idea

Held with strong conviction, sometimes implausible, but reality testing is intact: the holder can be moved by counter-evidence. A belief that collapses when finally checked was, by the clinical definition, never a delusion.

The user who abandons the conviction the moment a credible source pushes back has not experienced a psychotic state, and no claim arises. The AI may have validated a falsehood and caused distress, but there was no psychotic state to derive, induce, trigger, or exacerbate. A litigation theory that reaches this user reaches too far, and a published loss on it becomes precedent against the users the four verbs are meant to protect.

The boundary is clean at its ends and contested in its middle

A first episode of psychosis in a previously undiagnosed person and an egged-on overvalued idea can present identically at the moment of crisis, because fixity is sometimes only legible in hindsight, after the belief either held or broke. Some cases offer no hindsight at all: a user who does not survive cannot be re-tested, and the reconstruction must proceed from content, behavior, and trajectory. So a prior diagnosis is corroborating evidence, never the entry requirement; gating claims behind documented diagnosis would exclude exactly the undiagnosed first-episode patients disproportionately represented among the most severe outcomes.

The middle is the zone the defense will contest in both directions. Against a non-psychotic plaintiff it need do nothing; the fixity test excludes the case on its own. Against a genuinely psychotic first-episode plaintiff it will attempt the reverse move, recasting a true delusion as a mere overvalued idea. The answer is not to pretend the line is sharp but to direct the inquiry to the right evidence: documented or reconstructable loss of reality testing, the fixity and bizarreness of the belief, and the behavioral commitment to it.

DSM-5 already says "induced," not "caused," for the chemical analogues

Methamphetamine The most studied example. Symptoms at sufficient dose, susceptibility varying by genetics, history, sleep, stressors. Reliably "induction or precipitation," never causation.	Prescribed amphetamine Two users at the same prescribed dose differ entirely. Pharmaceutical literature, prescribing guidelines, and FDA labeling all use induction language.	Corticosteroids A well-known precipitant, particularly with prior psychiatric history. Withdrawal resolves most cases: induction followed by remission, not permanent state.	Cannabis The most contested category, structurally identical to the AI debate. Even with the strongest longitudinal data, the consensus converged on the trigger framing.
--	--	---	--

The legal structures built around these substances require demonstrated knowledge of the induction risk and demonstrated failure to warn, monitor, or mitigate. None of that requires proving the substance caused the disorder in any specific user.

The Strategy

Why the causal framing loses, what the four verbs permit, and what the field should do instead

The causal pleading is both harder to prove and easier to defeat

Harder to prove than the alternative

Disparate impact under Section 504 and ADA Title III does not require proof that the platform caused the underlying disability, only that the design produces disproportionate harm to a protected class and that reasonable modifications went unimplemented. The disability statutes are agnostic to the verb; the case is strongest under whichever verb the evidence supports, and weakest under cause, the one verb the evidence cannot support.

OpenAI's October 2025 disclosure quantifying weekly active users showing possible signs of mental health emergencies was an admission against interest of population-scale impact, not of causation. It is the statistical foundation for the impact element, which is the element the four verbs actually need.

Invites the prepared defense

The defense argues pre-existing vulnerability, triggering rather than causing, and pulls its citations from Pierre, Hudon, Stip, Østergaard, and Sakata, the same literature the plaintiffs would cite. Cross-examination becomes trivial. Even the eggshell plaintiff doctrine still requires proving the conduct triggered the harm, and a jury offered "caused" versus "triggered in someone vulnerable" may pick the second even when both are present.

Four distinctions the causal framing collapses

1

Population vs. individual

Population-level evidence supports measurable harm in a substantial subpopulation. It does not support causation in any specific user. The verbs permit the first claim without forcing the second.

2

Direct vs. indirect mechanism

Direct outputs (validating delusional content, dangerous advice, simulated intimacy) and structural usage effects (sleep disruption, isolation, immersion) both count. The causal framing biases toward whatever quotes easily from a chatlog.

3

Transient vs. persistent

Induction can remit on cessation (the corticosteroid model); triggering can produce persistent disorder (the cannabis model). Flattening the two is clinically inaccurate and legally weaker.

4

Platform level vs. user level

What fraction of exposed users develop symptoms, and which characteristics predict who, are separate research questions. "Does AI cause psychosis" conflates them into a question the field has correctly avoided.

Three moves, for researchers and for counsel

1

Anchor research on the substance-induced methodology

DSM-5's substance/medication-induced category provides validated criteria and temporal-relationship requirements. Operationalize induction with usage pattern data (frequency, duration, immersion depth) the way alcohol use disorder criteria use consumption data. Results integrate into existing psychiatric epidemiology instead of isolating in a novel category.

2

Distinguish the verbs in case reports and cohorts

Pierre and colleagues explicitly weigh induction against exacerbation in their case report. Future case reports should follow, and cohort studies should design measurement to permit attribution across the four verbs rather than pooling every post-exposure outcome.

3

Abandon the causal framing in litigation

Section 504 and ADA Title III do not require causation. Failure-to-warn and failure-to-monitor theories do not require it. The pharmaceutical liability tradition provides a richer template than the strict products theory currently pleaded. Importing causation forfeits the stronger pleadings.

What this commentary does and does not claim

The argument is methodological. It does not claim any specific case in the active litigation is incorrectly decided or incorrectly pleaded; whether framing changes convert into pleading changes is a question for the litigation teams. Nor do the four verbs exhaust the possibilities: catalyzing, scaffolding, modeling, and sustaining may merit independent development. These four are the ones the clinical literature is currently using, with the deepest grounding in the substance-induced tradition.

The author is not a clinician. The argument is grounded in the published clinical literature and in lived experience as caregiver to a family member with schizoaffective disorder; the research literature recognizes lived experience, including family and caregiver experience, as a form of expertise complementary to clinical training. The recommendations that depend on clinical interpretation require clinician-led development. The framing argument does not.

The field's path forward is not a stronger causal claim. It is the case that does not depend on causation at all

AI does not cause psychosis in the strict sense, and it does not need to for liability to attach. It derives, induces, triggers, or exacerbates psychotic states in users whose vulnerability varies across the exposed population, and each relationship is independently sufficient under structures that do not require causation. **The framing the public conversation needs is the framing the clinical literature has already established.**

Gilly, T. (2026). *The Four Verbs: Why AI-Induced Psychosis Research Should Frame AI as Vector, Trigger, or Amplifier Rather Than Cause*. Working paper (v4), Real Safety AI Foundation.

Burns, A. V., Nelson, K., Wang, H., Hegarty, E. M., & Cohn, A. B. (2025). "The algorithm is hacked": Analysis of technology delusions in a modern-day cohort. *British Journal of Psychiatry*. Advance online publication. <https://doi.org/10.1192/bjp.2025.10452>

Hawke, L. D., Sheikhan, N. Y., & Rodak, T. (2024). Lived experience and family engagement in psychiatry research: A scoping review of reviews. *Health Expectations*, 27(2), e14057. <https://doi.org/10.1111/hex.14057>

Hudon, A., & Stip, E. (2025). Delusional experiences emerging from AI chatbot interactions or "AI psychosis." *JMIR Mental Health*, 12, e85799. <https://doi.org/10.2196/85799>

Østergaard, 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>

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

Pierre, J. M., Gaeta, B., Raghavan, G., & Sarma, K. V. (2025). "You're not crazy": A case of new-onset AI-associated psychosis. *Innovations in Clinical Neuroscience*, 22(10–12), 11–13.

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>