The Four Verbs:

Why AI-Induced Psychosis Research Should Frame AI as Vector, Trigger, or Amplifier Rather Than Cause

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

Working paper, July 2026 (v5)

ABSTRACT

Public discourse, advocacy organizations, and active litigation increasingly frame the relationship between AI chatbots and psychotic episodes as causal: ChatGPT caused psychosis, the chatbot drove the user into delusion, the platform produced the harm. The clinical literature published in peer review, examined across multiple research groups (UCSF, Aarhus, Université de Montréal, UCLA), avoids but-for causal attribution at the level of the individual case, favoring instead four distinct verbs: derivation, induction, triggering, and exacerbation. The verbs are not interchangeable. Each describes a different etiological relationship between AI exposure and psychotic state. None of them is but-for causation. This commentary makes the implicit explicit: it surfaces the four verbs as a methodological instrument, demonstrates that the substance-induced disorders tradition in DSM-5 and ICD-11 already operates on the same structure for chemically analogous conditions (Adderall-induced psychosis, methamphetamine-induced psychosis, corticosteroid-induced psychosis), and argues that the causal framing now dominating public discourse and product liability pleading creates a defense the platforms can win on. Strict liability regimes (Section 504, ADA disparate impact) do not require the causal framing and are weakened by it. Plaintiff strategy that imports causation creates evidentiary burden the clinical literature cannot support. The four verbs are more accurate clinically and stronger legally. Its boundary is the fixity of the belief: where no psychotic state was present, because the belief was an overvalued idea that yielded to counter-evidence rather than a delusion, none of the four verbs applies and no claim arises; where a psychotic state did exist, derivation, induction, triggering, and exacerbation of it are each independently sufficient for liability under regimes that do not require causation. The commentary closes with a brief note on what the public discourse and the active litigation against AI platforms should be saying instead.

Keywords: AI-induced psychosis, etiological framing, component causation, fixity, overvalued idea, substance-induced disorders, diathesis-stress model, sycophancy, disparate impact.

I. THE CAUSATION TRAP

Public discourse on AI-induced psychosis has converged on causal language. Mass media reports describe AI chatbots as causing, driving, or producing psychotic episodes in users. Advocacy organizations describe their cases in causal

terms. Active product liability lawsuits filed in California state and federal courts allege that specific platform releases (notably GPT-4o) caused specific plaintiffs' delusional spirals, hospitalizations, and ongoing harms. The plaintiffs' framing is accessible, narratively powerful, and aligns with the lived experience of family members watching loved ones change. It is also the framing the platforms can win on.

The clinical literature published in peer review uses different language. Pierre and colleagues (2025), describing the first published case of new-onset AI-associated psychosis, explicitly frame the etiological question as "induction of new-onset psychosis versus exacerbation of pre-existing psychopathology." Not causation. Two options, both of which describe relationships short of strict causation. Hudon and Stip (2025) at Université de Montréal frame their JMIR Mental Health Viewpoint on AI psychosis as describing how AI may "trigger, amplify, or reshape" psychotic experiences in vulnerable individuals. Three verbs, none of them causation. Østergaard, who first proposed the hypothesis in his 2023 Schizophrenia Bulletin editorial and revisited it in 2025 in Acta Psychiatrica Scandinavica, frames the relationship as chatbots that may "fuel" delusions in individuals prone to psychosis. He acknowledges in 2025 that the correlations he has observed "not being proof of causation" and explicitly distinguishes the hypothesis from the causal claim. Sakata, a psychiatrist at the University of California, San Francisco, has publicly distinguished three possibilities: that AI is associated with people who would have developed psychosis independently, that AI precipitates psychosis in someone otherwise not predisposed, and that AI exacerbates illness in someone already susceptible. His own account of the twelve patients he had admitted by mid-2025 is careful about inference: he describes the phrase "AI psychosis" as not a clinical term, declines to extrapolate from twelve people, and says the technology "supercharged some of their vulnerabilities" in patients who had used it "in the wrong place at the wrong time" (Sakata, 2025, as reported in Yurkevich & Duffy, 2025). Burns and colleagues (2025), studying technology delusions in a cohort of 228 adults with psychosis, report that 51.7 percent described delusions incorporating the internet or new technologies, with the odds rising roughly 15 percent per year across the 2016 to 2024 window (odds ratio 1.15). Their data were collected largely before consumer generative AI was widely available and the study makes no claim about chatbots; what it establishes is that technological content is increasingly salient within delusional frameworks in people who already have psychotic disorders.

The question has since been posed directly in the field's flagship review journal and answered in the same register, as a risk question rather than a causal one (Keshavan et al., 2026). Across multiple research groups, multiple journals, and multiple methodological approaches, the clinical literature declines the verb the public discourse and the litigation are using. This is not an accident, and it is important to state precisely what the literature is and is not doing. It is not rejecting causal analysis as such. Psychiatric epidemiology has spent decades moving toward causal pluralism rather than away from causation, and the arc runs from many causes to one to many again (Kendler, 2019). What the literature declines is but-for attribution at the level of the individual case: the claim that this user's psychotic state would not exist absent this exposure. The relationship the evidence supports is one of triggering, induction, or amplification in a population whose susceptibility is determined by factors that pre-exist the AI exposure. That is the diathesis-stress structure, formalized as a second-order model subsuming six competing etiological accounts by Zubin and Spring (1977) and still the organizing frame of contemporary psychosis research (Hudon & Stip, 2025).

The distinction that matters is therefore not causation versus no causation. It is component causation versus but-for causation. A component cause is neither necessary nor sufficient on its own; it operates within a constellation of other factors that together are sufficient. This is the settled vocabulary in the adjacent literature: cannabis is described as neither necessary nor sufficient to cause schizophrenia and more likely "partly causal, interacting with other factors such as genetic liability" (D'Souza, 2023), and as a component cause for which the evidence is now sufficient to support public health messaging (Murray et al., 2016). AI exposure occupies the same position. The plaintiffs have pleaded but-for causation. The literature supports component causation. Those are different claims, and only one of them is provable on the available evidence.

When plaintiff strategy in active litigation imports the causal framing the clinical literature has rejected, it walks into a defense the platforms can prepare. The defense can pull from the same literature published in peer review the plaintiffs would otherwise cite. Pierre, Sarma, Hudon, Stip, Østergaard, and Sakata are all on record using verbs other than cause. Cross-examination on the etiological claim becomes trivial: did the AI cause the psychosis, or did it trigger an episode in a person whose vulnerability pre-existed the exposure? The clinical literature supports the second answer. The plaintiffs' pleading committed to the first.

The trap is not confined to journalism and pleadings. It has crossed into the technical literature. Chandra and colleagues (2026) state that they probe the causal link between sycophancy and AI-induced psychosis, and title the resulting paper as a demonstration that sycophantic chatbots cause delusional spiraling. What the paper models is the movement of a rational agent's posterior probability on a binary proposition past a confidence threshold. The clinical vocabulary of delusion and psychosis is carried by the title and the framing; the formalism underneath measures credence, and the discussion section concedes that the study addresses belief formation rather than the broader condition. This is the same substitution the complaints make, performed by careful researchers who flag the limit in their own discussion, which is a measure of how much gravitational pull the causal frame now exerts. Anyone who reads only the title has been told that causation is established.

The substitution begins earlier than the verb, and this is worth stating separately because it governs the rest of this commentary. The phrase delusional spiral attaches a diagnostic term to a trajectory before anything diagnostic has been established. A spiral is an observation about direction and speed. A delusion is a finding about a belief's relation to evidence. Naming the first with the vocabulary of the second presupposes the conclusion that the inquiry is supposed to reach, and it does so in the label, where presuppositions are hardest to see and impossible to argue with. The same objection applies to AI psychosis as a term of art, which Sakata himself flags as not a clinical term. Where a diagnostic word is doing descriptive work, the burden of proof has been transferred by nomenclature rather than met by evidence, and every downstream reader inherits a conclusion nobody argued for.

The trap is not subtle, and the way out of it is not subtle either. There are four verbs that accurately describe what AI is doing. None of them is cause. All of them are sufficient for liability under frameworks that do not require causation.

II. The Four Verbs Defined

A. Derivation

A psychotic state derives from AI exposure when the underlying psychiatric condition would have produced symptoms regardless, but the specific content, themes, or expression of the symptoms is shaped by the AI interaction. Of the four verbs, this is the one the literature does not already supply. Induction and exacerbation are posed as the operative alternatives by Pierre and colleagues (2025), and triggering by Hudon and Stip (2025); derivation is named here because the content-shaping relation is otherwise collapsed into exacerbation, which conflates a change in what a delusion is about with a change in how severe it is. A user with 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. Burns and colleagues (2025) supply the quantitative version of this pattern: technological themes are increasingly common in the delusional content of current cohorts. The substrate reading, that AI occupies the position radio and television once did, has been argued directly (Carlbring & Andersson, 2025), and it is the correct reading for some cases. It is not the correct reading for all of them. Pierre and colleagues (2025) draw the contrast explicitly, reporting that in their case the chatbot was "not merely a passive object of her new onset delusions in the way that ideas of reference can often involve television or

radio" but played a facilitating or mediating role in their formation. Derivation is the verb for the passive-substrate case. The other three verbs exist because not every case is that case.

Derivation is the verb that applies when the user has an established psychotic disorder and would have decompensated regardless. AI exposure shaped the decompensation but did not produce it. 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.

B. Induction

A psychotic state is induced by AI exposure when the AI delivers a known harm pattern (sycophancy, sleep disruption, social isolation through AI companion replacement, sensory disruption through AR or VR, or the bromism case in which a user substituted sodium bromide for table salt following chatbot advice and developed a chemical psychotic state; Eichenberger et al., 2025) to a user whose vulnerability is sufficient for the harm pattern to produce symptoms but who would not have spontaneously developed the disorder without the exposure. This is the verb that aligns most directly with the substance/medication-induced psychotic disorder category in DSM-5-TR. The induction language is doing diagnostic work: the substance or medication is recognized as capable of producing psychotic symptoms in vulnerable individuals, the symptoms are temporally linked to the exposure, and the diagnosis acknowledges the substance role without claiming the substance caused the disorder in every user.

Induction is the verb that applies when the user had no established psychotic disorder before AI exposure but carried vulnerability factors sufficient for the AI delivery to produce symptoms. The factors repeatedly named in the literature are loneliness and social isolation, trauma history, schizotypal traits, sleep disruption, nocturnal or solitary use, and algorithmic reinforcement of belief-confirming content (Hudon & Stip, 2025), together with concurrent prescription stimulant use (Pierre et al., 2025).

The Pierre case (2025) is the published exemplar, and its details matter more than its headline. Ms. A was a 26-year-old woman with a chart history of major depressive disorder, generalized anxiety disorder, and attention-deficit hyperactivity disorder, maintained on venlafaxine and methylphenidate. She had no previous history of psychosis or mania, which is a narrower fact than the absence of psychiatric history. Following a 36-hour sleep deficit she developed delusional beliefs about communicating with her deceased brother through a chatbot that told her, among other things, that she was not crazy. She was hospitalized in an agitated and disorganized state, discharged with a diagnosis of unspecified psychosis and rule-out bipolar disorder, and her delusions resolved on antipsychotic treatment. Three months later, after her outpatient psychiatrist stopped the antipsychotic and restarted venlafaxine and methylphenidate, and after another period of sleep restriction, she was briefly rehospitalized.

This is induction in the precise sense the verb requires: a first psychotic episode in a person with non-psychotic psychiatric history and identified vulnerability factors. It is not induction in a user with no disorder at all, and the case should not be cited for that. The authors themselves are careful here, noting that Ms. A had several contributing or confounding risk factors including a pre-existing mood disorder, prescription stimulant use, sleep deprivation, and a self-described propensity for magical thinking, and that her hospitalizations support a diagnosis of either brief psychotic disorder or manic psychosis fueled by lack of sleep and behavioral activation.

The second admission is the harder fact and it belongs in any honest account. Pierre and colleagues report that the relapse delusions arose largely without encouragement from the chatbot, and note that this supports the possibility that the role of AI in such cases is more coincidental than causal. They also raise directly the hypothesis that some AI-associated delusions reflect the growing cultural embeddedness of a new technology, such that the recent coverage could be a manifestation of moral panic. A taxonomy of etiological roles has to be able to absorb that hypothesis rather

than ignore it, and the four verbs can: derivation is precisely the verb for a case in which the technology supplies content to a disorder that would have decompensated regardless.

C. Triggering

A psychotic state is triggered by AI exposure when the user has a latent diathesis (genetic predisposition, neurodevelopmental substrate, or accumulated vulnerability such as complex post-traumatic stress) that would not have expressed at this time, in this form, or possibly at all, in the absence of the exposure, but does express because the exposure crosses a threshold the diathesis was approaching or had not yet been pressed against. Triggering is the diathesis-stress relationship in its purest form. The diathesis is necessary. The stress is necessary. Neither alone produces the outcome. The two together cross the threshold for clinical presentation.

Triggering is the verb that applies when the user has a substrate that may or may not have ever produced symptoms in the absence of AI but that, once exposed to the specific stressor profile AI delivers, expresses as psychotic state. This is the relation Hudon and Stip (2025) anchor on explicitly. AI is "a novel psychosocial stressor" that interacts with predispositions to psychosis. The relationship is contingent in both directions: the same AI exposure to a different user with no diathesis produces no symptom, and the same diathesis exposed to a different stressor (or no stressor) produces no symptom at this time. Both have to converge.

D. Exacerbation

A psychotic state is exacerbated by AI exposure when the user has an existing presentation of psychotic or pre-psychotic symptoms, may or may not have a formal diagnosis, and the AI exposure intensifies the existing symptoms, accelerates the trajectory of decompensation, or produces relapse in someone who had been stable. Exacerbation is the verb that applies to users who are already symptomatic, already in care or having dropped out of care, already cycling through episodes, and whose engagement with AI deepens or extends a state that pre-existed the exposure. Sakata's account of twelve patients hospitalized in 2025 in connection with extended chatbot use applies most directly to this category; the reporting describes their psychosis as partly made worse by talking to AI chatbots (Yurkevich & Duffy, 2025). Made worse, not produced.

The strongest evidence for exacerbation is now population-scale. Olsen, Reinecke-Tellefsen, and Østergaard (2026) screened the electronic health records of approximately 54,000 psychiatric patients in the Central Denmark Region and identified 181 clinical notes mentioning AI chatbot use, with worsening documented across delusions primarily, and also mania, suicidal ideation, self-harm, disordered eating, and obsessive-compulsive symptoms, alongside a smaller set of records in which chatbot use appeared constructive. The authors report a clear increase over time in records mentioning harmful use, and they are explicit that the design does not establish causation and that identified cases are likely a fraction of actual ones, since only harms written into a chart can be counted. That is the exacerbation verb with an epidemiological denominator under it, and it is the first such study in the field.

Exacerbation overlaps with derivation but is not identical. Derivation describes content shaping. Exacerbation describes severity, duration, or trajectory shaping. A user whose disorder would have produced delusions of one type and intensity in the absence of AI, and whose disorder produces delusions about AI at greater intensity or longer duration with AI exposure, is experiencing both derivation and exacerbation simultaneously. The verbs are not mutually exclusive.

III. The Threshold Question: When No Verb Applies

Each of the four verbs presupposes a psychotic state for the verb to act upon. Derivation shapes the content of a psychotic state. Induction produces one in a vulnerable user. Triggering expresses one from a latent diathesis. Exacerbation intensifies one already present. None of the four operates in the absence of a psychotic state. This is not a limitation of the account. It is its most important boundary, because it marks the line between the cases the disability and product liability doctrines should reach and the cases they should not.

The clinical construct that draws the line is the fixity of the belief. DSM-5-TR defines a delusion as a fixed belief that is not amenable to change in light of conflicting evidence (American Psychiatric Association, 2022; Arciniegas, 2015). Fixity is the operative property. An overvalued idea, by contrast, is a belief held with disproportionate emotional significance that dominates cognition and behavior without meeting criteria for either an obsession or a delusion, held with strong and sometimes idiosyncratic conviction but with reality testing intact, so that the holder retains the capacity to weigh counter-evidence and can be moved by it (Arciniegas, 2015). The modern conceptual review reaches the same place from a different direction, holding that such beliefs are sustained "through emotional salience, reinforcement contingencies, and value-based meaning rather than through fundamental loss of reality testing" (Dork et al., 2026). Two literatures converge here without overlapping: a clinical neuropsychiatry review written for practicing clinicians, and a conceptual and historical analysis written for researchers. Both land on intact reality testing as the discriminating property. The construct traces to Wernicke and was distinguished from delusion by Jaspers on grounds of psychological understandability rather than conviction alone; McKenna (1984) described its phenomenology as combining non-delusional conviction, non-obsessional preoccupation, and non-phobic fear. The distinction does not turn on how strange the belief sounds or how stubbornly it is asserted. It turns on whether the belief survives contact with disconfirming evidence. A belief that collapses when it is finally checked was, by the clinical definition, never a delusion. It failed the fixity test.

This boundary does the sorting the four verbs require. A user who arrives at an intellectual conviction, holds it through a period of AI interaction that validates it, and abandons it the moment a credible source pushes back has not experienced a psychotic state. The conviction was an overvalued idea throughout, validated rather than induced, confused rather than made delusional. There is no psychotic state for any of the four verbs to derive, induce, trigger, or exacerbate, and there is therefore no claim under an account built on those verbs. The AI may have behaved badly, validated a falsehood, wasted the user's time, or caused distress, but it did not produce, deepen, or express a psychotic disorder, because none was present. A litigation theory that reaches this user reaches too far. A published loss on such a theory becomes precedent a defense will cite against the users the account is meant to protect.

The boundary is clean at its ends and contested in its middle, and the honest version of this argument says so. The contest is not this commentary's invention. Overvalued ideas remain inconsistently defined and have been variously treated as attenuated delusions, as markers of poor insight, and as disorder-specific severity indicators (Dork et al., 2026), and separating overvalued ideas from delusions and obsessions remains an actively debated problem on which experts often struggle (Kristinsdottir et al., 2025). Two qualifications follow. Instruments exist that score belief conviction dimensionally rather than categorically, which is the right shape for a fixity inquiry. And there is preliminary evidence that forensic psychiatrists distinguish fixation driven by delusions, obsessions, and overvalued beliefs with substantial accuracy when working from a proper definition and clinical guidelines (Kristinsdottir et al., 2025). The line is hard, but it is not unworkable, and the remedy is definitional discipline rather than abandonment of the distinction. At one end, the belief that yields to counter-evidence was never delusional. At the other end, the bizarre, behaviorally committed, persecutory belief that drives a user to catastrophic action carries the signature of lost reality testing whether or not a clinician ever recorded a diagnosis. Between them lies a genuine indeterminate zone. 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

the property that distinguishes them, whether the belief was fixed, is sometimes only legible in hindsight, after the belief either held or broke. Some cases offer no such hindsight at all. A user who does not survive cannot be re-tested for fixity, and the reconstruction must proceed from the content, the behavior, and the trajectory rather than from a clean clinical record.

The boundary also disposes of an apparent counterexample that will be raised, and the disposal is instructive. Chandra and colleagues (2026) build a Bayesian model of a user conversing with a chatbot and show that even an idealized Bayes-rational user can be driven to high confidence in a false proposition, with sycophancy playing a causal role in that drift. If dangerous confidence requires no vulnerability, the structure this commentary rests on might appear to collapse.

It does not, because the modeled quantity is not a delusion and the model contains no test that could identify one. The authors define a catastrophic delusional spiral as the event that the user's posterior probability on a false binary proposition exceeds a threshold, set at 99 percent in their simulations. That is a measure of credence. At no point is the agent presented with evidence contradicting the belief and observed to hold it anyway, which is the only observation that could establish fixity. Nor could it be: a Bayesian updater revises on evidence by construction, and the paper's own informed-user condition shows agents correctly inferring the chatbot's sycophancy rate and discounting accordingly. An agent that tracks evidence is an agent whose belief is amenable to change in light of conflicting evidence, which is precisely what DSM-5-TR requires a delusion to lack. The authors concede the point in their discussion, noting that the paper studies the narrow question of how sycophancy affects belief formation while AI psychosis shows many other features.

What the model describes with real precision, then, is the mechanism by which a chatbot drives a person into strong conviction in a false belief that would still release on contradiction. That is the overvalued idea. It is the harm this commentary places outside the four verbs, and the modeling is a contribution to understanding it rather than a challenge to the boundary. That the paper's own motivating cases are Allan Brooks and Eugene Torres, described in the press as instances of AI psychosis, illustrates the point rather than undercutting it.

Two consequences follow, and this commentary states both rather than papering over the difficulty. First, a prior diagnosis is corroborating evidence, not the entry requirement. Gating the strong categories behind a documented diagnosis would exclude precisely the undiagnosed first-episode patients whose illness was never caught, a group disproportionately represented among the most severe outcomes. The anchor is the fixity of the belief and the clinical picture surrounding it, with diagnostic history serving as confirmation where it exists and its absence proving nothing. Second, the indeterminate middle is the zone the defense will contest in both directions. Against a non-psychotic plaintiff the defense need do nothing, because the fixity test excludes the case on its own. Against a genuinely psychotic first-episode plaintiff the defense will attempt the reverse move, recasting a true delusion as a mere overvalued idea to deny that any psychotic state existed. The account answers this not by pretending the line is sharp but by directing 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, rather than the lay impression of how unusual the belief sounded on its worst day.

IV. The Substance-Induced Tradition

The four verbs are not new. They have structured the psychiatric understanding of substance-related psychotic states for decades and are codified in both DSM-5-TR and ICD-11 (American Psychiatric Association, 2022; World Health Organization, 2019).

DSM-5 lists substance/medication-induced psychotic disorder as a separate diagnostic category (code 292.x or

F1x.95x depending on substance class) with explicit criteria that anchor on temporal relationship between substance exposure and psychotic symptoms, plus the persistence or remission of symptoms in relation to ongoing or terminated exposure. The category specifically uses the term "induced" and not the term "caused." The diagnostic system is acknowledging that the substance is etiologically relevant without claiming that the substance is the strict cause of the psychotic state. The diagnostic criteria operate on the same diathesis-stress structure that Zubin and Spring (1977) formalized.

Methamphetamine-induced psychosis is the most thoroughly studied example, and its evidence base carries a lesson the AI field should absorb before it repeats the error. The systematic review of risk factors grades the evidence as low to very low across almost every candidate predictor; the single predictor reaching moderate quality is frequency of use, with a dose-response relationship (Arunogiri et al., 2018). Family history was assessed in five studies and three found no significant association. The review states that established schizophrenia risk factors have not yet been shown relevant to the etiology of methamphetamine-associated psychosis, so the overlap between the two etiological pathways remains unclear. The underlying dose finding comes from a within-subject fixed-effects design in which users reporting 16 or more days of use per month had markedly elevated odds of psychotic symptoms relative to less frequent use (McKetin et al., 2013).

The lesson is that exposure intensity, not vulnerability profiling, is what the evidence actually supports. The literature does not say methamphetamine causes psychosis. It says methamphetamine induces psychotic symptoms at sufficient dose, in a population whose individual susceptibility is real but poorly characterized.

Adderall-induced psychosis (more precisely, amphetamine-induced psychosis from prescribed dextroamphetamine and amphetamine salts) operates on the same structure. Two users at the same prescribed dose may differ entirely in whether psychotic symptoms emerge. The factors that determine the difference are the same factors that determine vulnerability to psychotic disorders generally: genetic predisposition, family history, prior psychiatric history, concurrent sleep deprivation, and concurrent psychosocial stressors. The pharmaceutical literature, the prescribing guidelines, and the FDA labeling all use induction language. Nothing in the regulatory or clinical apparatus around amphetamine prescribing operates on a causation theory.

Corticosteroid-induced psychosis is the third example, and it makes the same point more sharply. The standard review pools two meta-analyses finding severe psychiatric reactions in nearly 6 percent of patients on systemic corticosteroid therapy and mild to moderate reactions in about 28 percent, and reports that dosage is directly related to the incidence of adverse effects while "neither the presence nor the absence of previous reactions predicts adverse responses to subsequent courses of corticosteroids" (Warrington & Bostwick, 2006; see also Lewis & Smith, 1983). Dose predicts. Psychiatric history does not. Subsequent reviews report the same dose-response relation, with the commonly cited inflection above 40 mg per day of prednisone or equivalent, and note no consistent relationship with prior psychiatric history (Canessa-Muñoz et al., 2025).

The medical literature describes the relationship as induction throughout. Clinical management assumes that dose reduction or withdrawal resolves symptoms in most patients, which is a model of induction followed by remission rather than causation followed by permanent state, though a persistent minority is documented (Gable & Depry, 2015).

Cannabis-induced psychosis is the most contested category and is precisely contested because of the causation question. The debate over whether cannabis use causes schizophrenia or precipitates it in predisposed individuals is structurally identical to the debate the AI psychosis field is now having, and the way that debate has actually resolved is instructive.

It has not resolved into avoidance of causal language. It has resolved into component causation. The relationship is dose-dependent, age-dependent with adolescent exposure more strongly implicated, and genetically modulated (Murray et al., 2016; Krebs et al., 2019). A quasi-experimental within-individual design found a dose-dependent association

between continued cannabis use and psychotic relapse that was unlikely to result from self-medication or shared genetic and environmental confounding (Schoeler et al., 2016). Genetically informed work finds robust association between liability to cannabis use disorder and schizophrenia with mixed evidence on causal direction (Johnson et al., 2021). On that basis Murray and colleagues (2016) concluded that the evidence for cannabis as a component cause of psychosis is now sufficient to support public health messaging, and D'Souza (2023), writing explicitly for the balanced view, holds that cannabis is neither necessary nor sufficient to cause schizophrenia and is more likely partly causal in interaction with genetic liability. The framing that fits the whole body of evidence is a multi-causal model in which vulnerability interacts with a precipitating agent to produce the outcome, a common model in epidemiology that has proved persistently difficult to communicate to the public (Hamilton, 2017).

The pattern across all four substance categories is consistent, and it is not the pattern a casual reader expects. Across methamphetamine, amphetamine, and corticosteroids, the predictor that survives systematic scrutiny is exposure intensity; individual vulnerability markers are real but weakly and inconsistently evidenced. Across cannabis, the field settled on component causation rather than on either but-for causation or silence.

This appears at first to cut against the diathesis-stress structure on which the earlier sections rest, and the appearance is worth confronting rather than leaving to the reader. It does not, for a reason that is itself the most useful thing the substance literature has to teach this field. Diathesis-stress does not predict that vulnerability will be the better-measured term. It predicts that both terms are necessary and neither is sufficient. What decades of substance research establish is that of the two, exposure is the one that submits to measurement: it has a metric, a gradient, and a within-subject design available to it, which is why McKetin and colleagues could demonstrate a dose relationship inside individuals and why Warrington and Bostwick could report incidence by dose. Vulnerability has no comparable metric. It is inferred from proxies, family history, prior diagnosis, schizotypy, each of which is noisy, and the systematic reviews record that noise faithfully as low-quality evidence.

The inference to draw is not that vulnerability is unimportant. It is that a field which stakes its claims on the term it cannot measure will produce exactly the literature the substance field produced, decades of low-graded findings about predisposition alongside one robust finding about dose. AI psychosis research is at the point where that choice is still open. Usage telemetry is the best-measured exposure variable any of these literatures has ever had, since it is continuous, passively logged, and held by the deploying party. The verbs that turn on exposure are therefore the verbs the field can currently evidence, and the verbs that turn on vulnerability are the ones it should expect to argue about for twenty years. The psychiatric field is careful with but-for claims because the diathesis-stress structure of psychotic disorders does not support them. The verb that does the regulatory and clinical work is "induction." The legal regimes built around these substances do not require causation. They require demonstrated knowledge of the induction risk, demonstrated failure to warn or monitor, demonstrated availability of mitigation, and demonstrated failure to mitigate. None of that requires proving the substance caused the disorder in any specific user.

The AI psychosis field is operating on the same etiological structure. The legal regimes available to plaintiffs do not require causation either, and importing causation creates evidentiary burden the field cannot meet.

V. Why the Causal Framing Is a Strategic Error

On November 6, 2025, the Social Media Victims Law Center and the Tech Justice Law Project filed seven actions against OpenAI, Inc. and Samuel Altman in California superior courts, alleging wrongful death, assisted suicide, and involuntary manslaughter in some cases and product liability, consumer protection, and negligence claims across the set. Those actions were coordinated with additional matters before the Judicial Council of California as *ChatGPT Product Liability Cases*, JCCP No. 5431. The complaints frame the relationship in causal terms, alleging that the

premature release of GPT-4o, described in the pleadings as dangerously sycophantic, produced specific plaintiffs' mental health crises, hospitalizations, financial distress, and family disruption. The product liability theory treats GPT-4o as a defective product, with the defect causing the harm.

One of those actions illustrates the problem this commentary identifies with unusual clarity, and it is worth stating carefully because the plaintiff is a named living person and nothing here concerns his credibility. *Brooks v. OpenAI, Inc.*, No. 25STCV32386 (Cal. Super. Ct. L.A. Cnty. filed Nov. 6, 2025), pleads strict product liability for defective design and failure to warn, negligent design and failure to warn, and violation of California's Unfair Competition Law. It alleges that the plaintiff had no history of mental illness, that the product's responses reinforced and escalated his thinking over roughly 300 hours across 21 days, and that the result was an induced delusional episode. Three features of that pleading are in tension with each other under the analysis developed above.

First, alleging no psychiatric history makes induction harder to establish, not easier. Induction requires that the exposure was sufficient for symptom emergence in this user, and the vulnerability factors that make that showing are the ones the pleading disclaims. Second, the pleaded verb is induction of a delusional episode, which places the claim in the strongest of the four categories and therefore under the heaviest evidentiary burden. Third, and most consequentially, the belief in question is publicly reported to have released when the plaintiff put it to a second chatbot that contradicted it. Under the fixity criterion set out in Part III, a belief that collapses on first contact with disconfirming evidence did not meet the definition of a delusion; it was an overvalued idea, reinforced rather than induced.

None of that means no harm occurred. Three hundred hours, reputational damage, and occupational disruption are injuries whether or not the belief was fixed. It means the tort theory is weaker than it needs to be, on its own terms. Design defect and failure to warn both require that the defect caused this plaintiff's injury. The clinical literature supports the population-level claim that sycophantic design degrades belief formation; it does not supply specific causation in an individual, and Part VI explains why the second inquiry is not even reached until the first is satisfied. Meanwhile the pleaded injury, an induced delusional episode, may not be the injury that occurred, since a belief that released on first contradiction was reinforced rather than induced and was an overvalued idea rather than a delusion. A defense that concedes sycophancy, concedes distress, and contests only the diagnosis and the individual causal link is available on the face of the complaint.

That is the causation trap operating in a live pleading, and it is not a criticism of bringing the case. It is an argument that the strongest available theory was passed over for the most familiar one.

This is a strategic error in two distinct ways.

First, the causal claim is harder to prove than the alternative claims available. Liability under Section 504 of the Rehabilitation Act and Title III of the Americans with Disabilities Act does not require proof that the platform caused the underlying disability. The Supreme Court has held that Section 504 entitles an otherwise qualified individual to meaningful access to the benefit a grantee offers, and that reasonable accommodations in the program may have to be made to assure it (*Alexander v. Choate*, 469 U.S. 287, 301 (1985)). Congress understood the discrimination it was addressing to be most often the product "not of invidious animus, but rather of thoughtlessness and indifference, of benign neglect" (*Choate*, 469 U.S. at 295). Neither state of mind nor causal contribution to the underlying disability is an element of that showing.

One caveat has to be stated plainly rather than assumed away, because a litigation strategy built on the unstated version will fail in some circuits. Whether Section 504 supports a private disparate impact claim at all is unsettled. *Choate* itself rejected "the boundless notion that all disparate-impact showings constitute prima facie cases under §504" and expressly assumed "without deciding that §504 reaches at least some conduct that has an unjustifiable disparate impact upon the handicapped," declining to resolve the tension between statutory purpose and the desire to keep the section within manageable bounds (469 U.S. at 299). The Ninth Circuit permits the theory (*Payan v. Los Angeles*

Community College District, 11 F.4th 729 (9th Cir. 2021)). The Sixth Circuit has held the opposite, reasoning from the statutory phrase "solely by reason of her or his disability" and from the section's patterning on Title VI, which does not reach disparate impact: "We now resolve what *Choate* did not and conclude that §504 does not prohibit disparate-impact discrimination" (*Doe v. BlueCross BlueShield of Tennessee*, 926 F.3d 235, 241 (6th Cir. 2019)).

The consequence for strategy is that venue is a variable rather than a footnote, and that the disparate impact theory should not be pleaded alone. Failure to accommodate and denial of meaningful access are independent theories, they are available in every circuit, and they no more require proof of causation of the underlying disability than disparate impact does. The two are not interchangeable, and *Payan* draws the line: systemic barriers are properly analyzed under disparate impact, while the denial of a specific individualized request is analyzed as a failure to accommodate (11 F.4th at 738). Platform architecture is a systemic barrier, so disparate impact is the natural fit and failure to accommodate is the surviving alternative where the theory is foreclosed or where a user made and was denied a specific request. A posture that pleads both loses nothing and survives *Doe*. What neither requires is the verb the complaints have committed to. The four verbs do not import that burden. A platform that derives harm in users who are already symptomatic, induces harm in vulnerable users, triggers harm at threshold, or exacerbates harm in users who are already affected is producing disparate impact under any of the four verbs. The disability statutes are agnostic to the verb. The plaintiff's case is stronger under whichever verb the evidence supports, and weakest under cause specifically because cause is the verb the evidence cannot support.

Second, the causal claim invites the defense the platforms have already been preparing. There is admission-adjacent material available that does not depend on the causal verb. Østergaard (2025) ties the mid-2025 rise in reports of chatbot-fueled delusions to the April 25, 2025 update to GPT-4o, attributes the mechanism to reinforcement learning from human feedback leaning too heavily on user preference signals, and quotes the company's own acknowledgment that the update made the model noticeably more sycophantic. That is a peer-reviewed source naming the specific model release the complaints name and quoting the developer conceding the specific mechanism. It is an admission of a design property, not of causation.

The design property is now measured. Across 11 state-of-the-art models, Cheng and colleagues (2026) found that AI affirmed users' actions 49 percent more often than humans did, including where the query involved deception, illegality, or other harms, and three preregistered experiments (N = 2405) found that a single interaction with a sycophantic model reduced participants' willingness to take responsibility and repair interpersonal conflict while increasing their conviction that they were right. Sycophantic models were nonetheless trusted and preferred, which the authors identify as a perverse incentive: the feature that causes the harm is the feature that drives engagement.

That finding is worth more to a plaintiff than a causal claim about any individual user. It establishes a measurable design property, present across the industry rather than idiosyncratic to one release, that operates on ordinary users rather than only on vulnerable ones, and that the developer has a commercial reason to preserve. Those are the elements of a design defect and of a disparate impact showing. None of them is a claim that a chatbot caused a particular person's psychosis. The disclosure framing the platforms can mount is exactly the framing the clinical literature supports: these users had pre-existing vulnerabilities, the platform's role was triggering or exacerbating but not causing, and the proximate harm is the convergence of vulnerability and exposure rather than the exposure alone. The defense pulls citations from the literature published in peer review from Pierre, Sarma, Hudon, Stip, Østergaard, and Sakata. The plaintiffs are left arguing causation against a defense that has the clinical literature on its side.

The eggshell plaintiff doctrine in tort law would, in theory, prevent the defense from using vulnerability to reduce platform liability under a causation theory. A defendant must take the victim as found, extraordinarily sensitive or not (*McKinnon v. Kwong Wah Restaurant*, 83 F.3d 498, 506 (1st Cir. 1996)), and the Ninth Circuit has held it reversible error to treat a predisposition toward mental illness as relieving a defendant of responsibility for psychological disability,

since a wrongdoer takes his victim as he finds him (*Wakefield v. NLRB*, 779 F.2d 1437, 1438 (9th Cir. 1986)). *Wakefield* also supplies two propositions that map directly onto the verbs developed here: aggravation of a preexisting nondisabling condition into a disabling one is itself sufficient proof of causation, and the unlawful conduct need not be shown to be the sole cause of the disability. The Eighth Circuit has held the principle applicable to psychological harm specifically, reasoning that foreseeability does not limit money damages and that this includes harm to a plaintiff who happens to have a fragile psyche (*Jenson v. Eveleth Taconite Co.*, 130 F.3d 1287, 1294 (8th Cir. 1997)).

The rule carries a threshold that plaintiffs should meet head-on rather than discover at summary judgment: for emotional harm, the plaintiff must first show that an ordinarily sensitive person could have suffered the alleged harm, and only then does the defendant take the victim as found (*McKinnon*, 83 F.3d at 506). That threshold is where the population evidence does its second job, because a documented pattern of harm across a psychiatric population is evidence that the exposure is capable of harming an ordinarily sensitive member of it. The formal modeling work is stronger still on this point: a mechanism shown to degrade belief formation in an idealized rational agent (Chandra et al., 2026) is by hypothesis a mechanism that reaches the ordinarily sensitive person, which is what the threshold asks.

Most importantly, the vulnerability defense described above is an apportionment argument, and apportionment is the defendant's affirmative defense to carry. To limit liability that way a defendant must prove that the plaintiff's damages are divisible and that outside factors contributed to the harm; where the court finds apportionment impossible, the defendant is liable for the entire amount (*Jenson*, 130 F.3d at 1293; *McKinnon*, 83 F.3d at 508). A defense built on the convergence of vulnerability and exposure is therefore not a denial of liability. It is a request to divide damages, made by the party who bears the burden of showing they can be divided, in a clinical domain where the diathesis-stress structure makes division exceptionally hard to demonstrate. But the rule reaches the aggravation of the pre-existing condition, not the condition itself, which is the same line the four verbs draw between derivation and exacerbation on one side and induction on the other. And the doctrine still requires the plaintiff to prove the defendant's conduct triggered or aggravated the harm in some sense. The clinical literature supports trigger and exacerbation. It does not support cause. A jury asked to decide between "the AI caused this" and "the AI triggered this in someone vulnerable" may decide the second verb is more accurate even when both are present. Plaintiff strategy that committed to the first verb loses control of the framing.

The framing the disability statutes provide is, in fact, broader and stronger than the framing the product liability theory imports. Section 504 reaches every program receiving federal financial assistance, which includes virtually every educational institution, healthcare provider, and government contractor that has integrated AI into operations. Title III reaches public accommodations, which case law is increasingly extending to digital platforms. Disparate impact theory under Section 504 and Title III does not require the platform to have caused the underlying disability. It requires the platform's design to fail meaningful access for users with pre-existing mental health and neurodevelopmental disabilities, and it requires that failure to track the protected class. OpenAI's own October 2025 disclosure provides the statistical foundation for the impact element.

The lawsuits the field needs are pleaded in disparate impact, not in product liability. The framing the public discourse needs is built on the four verbs, not on cause.

VI. What the Four Verbs Actually Permit

The four verbs are not a defensive concession. They are more analytically powerful than the causal framing because they permit distinctions the causal framing collapses.

The first distinction is between claims at the population level and claims at the level of individual cases. This distinction is not new to law, and naming its existing form is the fastest way to make the argument land. In toxic tort

it is the difference between general and specific causation: "General causation is whether a substance is capable of causing a particular injury or condition in the general population, while specific causation is whether a substance caused a particular individual's injury" (*Knight v. Kirby Inland Marine Inc.*, 482 F.3d 347, 351 (5th Cir. 2007)).

Stated that way, the diagnosis of the litigation is exact, and the procedural consequence is severe. Specific causation evidence is admissible only as a follow-up to admissible general causation evidence, and courts run the inquiry in that order: first whether there is admissible general causation evidence, and only then whether there is admissible specific causation evidence (*Knight*, 482 F.3d at 351). A plaintiff who has pleaded the individual claim without first establishing the population claim does not lose at trial. That plaintiff loses at step one, on a motion, before the individual evidence is ever reached.

The plaintiffs have pleaded specific causation. The clinical literature currently supports, at most, general causation. The best available population evidence is the Danish record review, which found 181 clinical notes mentioning chatbot use among roughly 54,000 psychiatric patients, with documented worsening in a subset (Olsen et al., 2026). That is genuine population-level signal in an already-diagnosed population, and it is also a small proportion of the screened records; the authors are explicit that only harms recorded in a chart can be counted and that the true figure is likely higher. Overstating the magnitude is unnecessary and invites correction. What the four verbs permit is the claim the evidence supports, at the level the evidence supports it, without requiring the further claim that this user's psychosis was produced by this platform's outputs in this conversation.

The second distinction is between direct mechanism claims and indirect mechanism claims. AI may directly induce harm through specific outputs (validating delusional content, providing dangerous medical advice, encouraging suicidal ideation, offering simulated romantic intimacy that intensifies parasocial attachment) or indirectly trigger harm through usage pattern effects (sleep disruption from compulsive use, social isolation from companion replacement, reality testing degradation from extended immersion). The four verbs permit both kinds of mechanism claims. The causal framing tends to anchor on direct mechanism claims because they are easier to narrate, which biases the litigation toward the cases with chatlog content that quotes most easily and away from the cases where the harm is structural.

The third distinction is between transient and persistent harm. Some AI-induced psychotic states resolve when the exposure is removed. Some persist. The four verbs permit both trajectories: induction can produce transient symptoms that remit on cessation (the corticosteroid model), and triggering can produce persistent disorder in someone whose latent diathesis would not otherwise have expressed (the cannabis-precipitating-schizophrenia model). The causal framing flattens this distinction and treats all outcomes after exposure as equally attributable to the exposure, which is both clinically inaccurate and legally weaker because it reaches outcomes the defense can attribute to the underlying vulnerability rather than to the exposure.

The fourth distinction, and the most important for clinical research design, is between the analysis at the platform level and the analysis at the user level. Population-level harm from AI exposure is real, measurable, and has been quantified by platforms themselves. User-level outcome from AI exposure depends on factors that vary across users and cannot be predicted from exposure alone. Research design anchored on the four verbs permits investigators to study induction at the population level (what fraction of exposed users develop symptoms) and vulnerability at the user level (which characteristics predict who) as separate questions. Research design that anchors on causation conflates these into a single question (does AI cause psychosis) that the field has been correctly avoiding because the answer cannot be coherently given.

VII. Methodological Recommendations

Three concrete recommendations follow from the four verbs. Two of the three have been anticipated in the clinical literature and are extended rather than introduced here: Hudon and Stip (2025) call for longitudinal and digital-phenotyping designs quantifying dose-response relationships between AI exposure and psychotic symptomatology, and for governance frameworks for AI-related psychiatric events modeled on pharmacovigilance. What is added here is the doctrinal consequence, which that literature does not address.

First, clinical research on AI-induced psychosis should anchor on the substance-induced disorders methodology. The DSM-5-TR criteria for substance/medication-induced psychotic disorder provide validated definitions, established temporal relationship requirements, and clear diagnostic boundaries. AI-induced psychosis fits that structure with minimal modification. The substance is conversational AI exposure. The induction criteria can be operationalized using usage pattern data (frequency, duration, immersion depth) the same way alcohol use disorder criteria are operationalized using consumption pattern data. Research that adopts this structure will produce results that integrate cleanly into existing psychiatric epidemiology, cite into the broader substance-induced disorders literature, and resist the methodological isolation that anchoring on a novel diagnostic category would create.

Second, clinical research should distinguish carefully among the four verbs in case reports and cohort studies. Pierre and colleagues (2025) explicitly raise the question of induction versus exacerbation in their case report and acknowledge that the answer affects the clinical interpretation. Future case reports should follow this practice. Cohort studies should design measurement to permit attribution across the four verbs rather than treating all psychotic outcomes that follow exposure as a single category.

Third, legal scholarship and litigation strategy should abandon but-for framing in favor of the verbs the clinical literature supports. The disability statutes (Section 504, ADA Title III) do not require it, subject to the circuit caveat noted above. Product liability theories built on failure to warn and failure to monitor do not require it. The pharmaceutical liability tradition of failure to label, failure to monitor, and failure to provide adequate prescribing guidance offers a richer template than the strict products liability theory currently being pleaded.

That template needs one adjustment before it transfers, and the case that shows why is itself a stimulant-psychosis case. In *Ehlis v. Shire Richwood, Inc.*, 367 F.3d 1013 (8th Cir. 2004), a failure-to-warn claim arising from psychosis during Adderall treatment failed because the prescribing physician already knew the risk. Under the learned intermediary doctrine the manufacturer's duty to warn runs to the physician rather than the patient, a warning to the physician is deemed a warning to the patient, and the causal link is broken where the prescriber had substantially the same knowledge an adequate warning would have conveyed. That is the pharmaceutical template working exactly as designed, against the plaintiff, on facts closely analogous to the ones the AI cases present.

Consumer AI has no learned intermediary, and the three rationales the doctrine rests on do not survive the move. The first, that medical ethics require an independent professional between manufacturer and user, has no analogue; nobody stands between the platform and the user. The second, that risk information is too technical for the user to evaluate, is if anything more true of model behavior than of pharmacology, but there is no professional to receive the technical warning instead. The third, that it is virtually impossible for a manufacturer to warn every ultimate user directly, is simply false here and is inverted: the platform holds a direct conversational interface with every user, controls the timing of every message, and can deliver an individualized warning at the moment of risk in a way no drug manufacturer ever could.

The absence of the intermediary is therefore an argument for a more demanding duty, not a weaker one, and it removes the defense that decided *Ehlis*. It also removes the apportionment escape. Where a defendant argues that damages flow from a prior condition rather than from its own conduct, the burden of establishing that causal relationship

rests on the defendant, and where apportionment proves impossible the defendant is liable for the entire amount (*McKinnon*, 83 F.3d at 508). The pre-existing vulnerability defense described in Part V is an apportionment argument, and apportionment is the defendant's burden to carry. Plaintiff strategy that imports causation creates evidentiary burden the clinical literature cannot support and forfeits the agnostic to the verb strict liability frameworks that produce stronger pleadings.

VIII. Limitations

The argument advanced here is methodological. It does not claim that any specific case in the active litigation is incorrectly decided or that any specific plaintiff is incorrectly pleading. The argument is that the framing dominating both the public discourse and the litigation strategy creates strategic vulnerabilities that are avoidable under the four verbs. Whether and how individual cases convert framing changes into pleading changes is a question for the litigation teams and is beyond the scope of a methodological commentary.

The four verbs do not exhaust the etiological possibilities. Other verbs (catalyzing, scaffolding, modeling, sustaining) may apply to specific subsets of AI psychosis cases and may merit independent development. The four verbs offered here are the ones the clinical literature is currently using and the ones with the deepest methodological grounding in the substance-induced disorders tradition. Future work may identify additional distinctions worth naming.

The author is not a clinician. The methodological argument advanced here is grounded in the published clinical literature and in the author's lived experience as caregiver to a family member with schizoaffective disorder; the author's pattern recognition for delusional presentation, reality testing impairment, and decompensation dynamics derives from sustained observation across years rather than from formal training. The mental health research literature establishes that lived experience, including family and caregiver experience, constitutes a recognized form of expertise complementary to clinical training (Hawke et al., 2024). The methodological recommendations that depend on clinical interpretation of the four verbs in case reports and cohort designs require clinician-led development; the framing argument does not.

IX. Conclusion

The clinical literature published in peer review on AI-induced psychosis is uniformly avoiding the verb that the public discourse and the active litigation are uniformly using. This is not a contradiction the field needs to resolve through new evidence. It is a contradiction the litigation strategy and public framing need to resolve by adopting the verbs the field is already using. AI does not cause psychosis in the strict sense, and it does not need to for liability to attach. AI derives, induces, triggers, or exacerbates psychotic states in users whose vulnerability profiles vary across the exposed population, and each of those four relationships is independently sufficient under regimes that do not require causation. The boundary the framework draws is not between deserving and undeserving users but between cases in which a psychotic state was present for a verb to act upon and cases in which the belief never lost its fixity and no psychotic state existed at all. The former are the cases the four verbs reach. The latter it correctly excludes, and excluding them protects the precedent the genuine cases will depend on. The four verbs are clinically accurate, methodologically grounded, and legally stronger than the verb that has dominated public discourse. The framing the public conversation needs is the framing the clinical literature has already established. The framing the litigation needs is the framing the disability law statutes already provide. Neither framing requires causation. Both framings are weakened by importing it.

The field's path forward is not to make a stronger causal claim. It is to make the case that does not depend on causation in the first place.

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

Alexander v. Choate, 469 U.S. 287 (1985).

Alexander v. Sandoval, 532 U.S. 275 (2001).

Americans with Disabilities Act of 1990, 42 U.S.C. §§12181–12189.

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

Arciniegas, D. B. (2015). Psychosis. *Continuum*, 21(3), 715–736. https://doi.org/10.1212/01.con.0000466662.89908.e7

Arunogiri, S., Foulds, J. A., McKetin, R., & Lubman, D. I. (2018). A systematic review of risk factors for methamphetamine-associated psychosis. *Australian & New Zealand Journal of Psychiatry*, 52(6), 514–529. https://doi.org/10.1177/0004867417748750

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*, 1–5. https://doi.org/10.1192/bjp.2025.10452

Brooks v. OpenAI, Inc., No. 25STCV32386 (Cal. Super. Ct. L.A. Cnty. filed Nov. 6, 2025), coordinated, *ChatGPT Product Liability Cases*, JCCP No. 5431.

Canessa-Muñoz, S., Yelmo-Cruz, S., & Hamilton-Lopez, A. (2025). Corticosteroid-induced psychosis: A report of two cases and review of the literature. *Cureus*. https://doi.org/10.7759/cureus.81515

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

Chandra, K., Kleiman-Weiner, M., Ragan-Kelley, J., & Tenenbaum, J. B. (2026). Sycophantic chatbots cause delusional spiraling, even in ideal Bayesians. *arXiv*. https://doi.org/10.48550/arXiv.2602.19141

Cheng, M., Lee, C., Khadpe, P., Yu, S., Han, D., & Jurafsky, D. (2026). Sycophantic AI decreases prosocial intentions and promotes dependence. *Science*, 391(6792). https://doi.org/10.1126/science.aec8352

Dork, J., Dimenstein, E., & Burns, L. (2026). Overvalued ideas: Conceptual analysis and literature review. *Behavioral Sciences*, 16(5), 708. https://doi.org/10.3390/bs16050708

Dohnány, S., Kurth-Nelson, Z., Spens, E., Luettgau, L., Reid, A., Gabriel, I., Summerfield, C., Shanahan, M., & Nour, M. M. (2025). Technological folie à deux: Feedback loops between AI chatbots and mental illness. *arXiv*. https://doi.org/10.48550/arXiv.2507.19218

Crocker v. Runyon, 207 F.3d 314 (6th Cir. 2000).

Doe v. BlueCross BlueShield of Tennessee, Inc., 926 F.3d 235 (6th Cir. 2019).

D'Souza, D. C. (2023). Cannabis, cannabinoids and psychosis: A balanced view. *World Psychiatry*, 22(2), 231–232. https://doi.org/10.1002/wps.21075

Eichenberger, A., Thielke, S., & Van Buskirk, A. (2025). A case of bromism influenced by use of artificial intelligence. *AIM Clinical Cases*, 4, e241260.

Ehlis v. Shire Richwood, Inc., 367 F.3d 1013 (8th Cir. 2004).

Gable, M., & Depry, D. (2015). Sustained corticosteroid-induced mania and psychosis despite cessation. *International Journal of Psychiatry in Medicine*, 50(4), 398–404. https://doi.org/10.1177/0091217415612735

Hamilton, I. (2017). Cannabis, psychosis and schizophrenia: Unravelling a complex interaction. *Addiction*, 112(9), 1653–1657. https://doi.org/10.1111/add.13826

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

Jenson v. Eveleth Taconite Co., 130 F.3d 1287 (8th Cir. 1997).

Johnson, E. C., Hatoum, A. S., Deak, J. D., Polimanti, R., Murray, R. M., Edenberg, H. J., Gelernter, J., Di Forti, M., & Agrawal, A. (2021). The relationship between cannabis and schizophrenia: A genetically informed perspective. *Addiction*, 116(11), 3227–3234. https://doi.org/10.1111/add.15534

Kendler, K. S. (2019). From many to one to many: The search for causes of psychiatric illness. *JAMA Psychiatry*, 76(10), 1085. https://doi.org/10.1001/jamapsychiatry.2019.1200

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

Knight v. Kirby Inland Marine Inc., 482 F.3d 347 (5th Cir. 2007).

Krebs, M.-O., Kebir, O., & Jay, T. M. (2019). Exposure to cannabinoids can lead to persistent cognitive and psychiatric disorders. *European Journal of Pain*, 23(7), 1225–1233. https://doi.org/10.1002/ejp.1377

Kristinsdottir, K., Freestone, M., & Underwood, A. (2025). Pathological fixation on shared beliefs: A review and bibliometric analysis of extreme, overvalued and delusion-like beliefs. *Frontiers in Psychiatry*, 16. https://doi.org/10.3389/fpsyt.2025.1715886

Lewis, D. A., & Smith, R. E. (1983). Steroid-induced psychiatric syndromes. *Journal of Affective Disorders*, 5(4), 319–332. https://doi.org/10.1016/0165-0327(83)90022-8

McKenna, P. J. (1984). Disorders with overvalued ideas. *British Journal of Psychiatry*, 145(6), 579–585. https://doi.org/10.1192/bjp.145.6.579

McKinnon v. Kwong Wah Restaurant, 83 F.3d 498 (1st Cir. 1996).

McKetin, R., Lubman, D. I., Baker, A. L., Dawe, S., & Ali, R. L. (2013). Dose-related psychotic symptoms in chronic methamphetamine users. *JAMA Psychiatry*, 70(3), 319. https://doi.org/10.1001/jamapsychiatry.2013.283

Murray, R. M., Quigley, H., Quattrone, D., Englund, A., & Di Forti, M. (2016). Traditional marijuana, high-potency cannabis and synthetic cannabinoids: Increasing risk for psychosis. *World Psychiatry*, 15(3), 195–204.

https://doi.org/10.1002/wps.20341

Olsen, S. G., Reinecke-Tellefsen, C. J., & Østergaard, 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

Payan v. Los Angeles Community College District, 11 F.4th 729 (9th Cir. 2021).

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. PMID 41635747.

Rehabilitation Act of 1973 §504, 29 U.S.C. §794.

Schoeler, T., Petros, N., Di Forti, M., Klamerus, E., Foglia, E., Ajnakina, O., Gayer-Anderson, C., Colizzi, M., Quattrone, D., Behlke, I., Shetty, S., McGuire, P., David, A. S., Murray, R. M., & Bhattacharyya, S. (2016). Association between continued cannabis use and risk of relapse in first-episode psychosis. *JAMA Psychiatry*, 73(11), 1173. https://doi.org/10.1001/jamapsychiatry.2016.2427

Wakefield v. NLRB, 779 F.2d 1437 (9th Cir. 1986).

Warrington, T. P., & Bostwick, J. M. (2006). Psychiatric adverse effects of corticosteroids. *Mayo Clinic Proceedings*, 81(10), 1361–1367. https://doi.org/10.4065/81.10.1361

World Health Organization. (2019). *International classification of diseases for mortality and morbidity statistics* (11th rev.).

Yurkevich, V., & Duffy, C. (2025, September 5). They thought they were making technological breakthroughs. It was an AI-sparked delusion. *CNN Business*.

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