The Fixity Test:

Refutation Without Diagnosis in the AI-Induced Psychosis Discourse

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

Working paper, July 2026 (v0.5)

ABSTRACT

Journalism, advocacy, and active litigation attach the words "delusion" and "psychosis" to named individuals on the basis of public information: news accounts, self-reports, and court filings. This commentary argues that the practice fails on the diagnostic manual's own terms, and that the failure runs in one specific direction. DSM-5-TR defines a delusion by fixity: a fixed belief that is not amenable to change in light of conflicting evidence. Fixity is legible only in trajectory, in whether the belief held or released when contradiction arrived, and never in how extreme the belief sounded at its peak. An asymmetry follows. Affirming the diagnostic label from public information is unavailable in principle, because affirmation requires a longitudinal record of the belief surviving contradiction, a negative symptom picture that public accounts do not carry, and the clearing of exclusion screens that operate before fixity is even reached. Refuting the label as applied, by contrast, requires exactly one datum that public accounts and court filings routinely supply: the documented moment the belief released on contradiction. A record that contains that moment has falsified the vocabulary it uses, in the same document, with no outside evidence required. The commentary derives a verdict rule, refutation without diagnosis: the careful observer may describe what the record shows, may classify the pattern it displays, and may demonstrate what the record rules out, while asserting nothing about what any individual has. It closes with the cost of the borrowed word: a definitional expansion that pathologizes people whose injury was deceptive design, and that dilutes the diagnosis for the community living with diagnosed psychotic conditions.

Keywords: AI-induced psychosis, fixity, delusion, overvalued idea, extreme overvalued belief, schizotypy, reality testing, DSM-5-TR, mental illness stigma, labelling effects, schizophrenia, disease trivialization, diagnostic appropriation, concept creep, diagnostic language, psychosis discourse, media discourse, dilution.

I. THE BORROWED WORDS

The public discourse on AI-induced psychosis runs on two clinical words. News features describe named individuals as delusional. Advocacy organizations describe their case collections as psychosis. Court filings plead that a platform's outputs drove a specific plaintiff into delusion. The words arrive with the full authority of the diagnostic manual behind them, and they are doing rhetorical work that plain descriptions of harm would not do: they convert an account of a difficult episode into an account of a medical event.

The clinical literature itself is far more careful. Writing in the BMJ, Pierre (2025) distinguishes induction of

delusional thinking from exacerbation of it, declines to resolve the question, and argues that the causally neutral terms, AI-associated psychosis and AI-associated delusions, are the most appropriate vocabulary precisely because premature conclusions about causality are unwarranted. The researchers who first proposed the hypothesis describe chatbots that may fuel delusions in individuals prone to psychosis, and explicitly distinguish observed correlation from demonstrated causation (Østergaard, 2023, 2025). A Viewpoint from the Université de Montréal group describes AI that may trigger, amplify, or reshape psychotic experiences in vulnerable individuals (Hudon & Stip, 2025), and a World Psychiatry letter poses the field's question as whether chatbots increase psychosis risk in individuals already at risk, a question about vulnerability rather than a diagnosis of anyone (Keshavan et al., 2026). Across research groups, the clinicians publishing in peer review decline to diagnose the named individuals in the public record, and they hedge the category itself.

The looseness, in other words, lives downstream of the clinic. The documents asserting the diagnostic categories with the most confidence are the documents written by the people least positioned to assert them: reporters, advocates, and pleadings drafted to persuade. This commentary examines what those documents can and cannot support, and finds that the answer is asymmetric. They cannot support the words they are using. They can, and frequently do, refute them.

II. The Fixity Test

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). Fixity is the operative property. The definition does not turn on how bizarre the belief sounds, how completely it organizes the believer's life, how much damage it does, or how alarmed the believer's family becomes. It turns on one structural question: when conflicting evidence arrives, does the belief hold?

The manual contrasts the delusion with the strongly held or overvalued idea. The overvalued idea has a long clinical lineage: Wernicke described it at the turn of the twentieth century, and McKenna's classical review defines it as a solitary, abnormal belief, neither delusional nor obsessional, that is preoccupying to the extent of dominating the sufferer's life (McKenna, 1984). The overvalued idea can be intense, consuming, idiosyncratic, and destructive. What it retains is reality testing: the holder keeps the capacity to weigh counter-evidence, and the belief can be moved by it. The forensic tradition has carried the concept to its hardest edge: the extreme overvalued belief, a rigidly held belief shared with others in a cultural or subcultural group, relished, amplified, and defended by its holder, and still explicitly differentiated from a delusion (Rahman, 2018). Even at the extreme, where the belief motivates catastrophic action, the clinical distinction holds, because the distinction tracks the structure of the belief rather than its distance from consensus.

Two consequences of the definition matter for everything that follows.

First, fixity is a property of trajectory. It cannot be read off a snapshot. A belief at its peak, described at the worst moment of the worst week, is indistinguishable in content and intensity from a delusion, because content and intensity are the properties the definition sets aside. The property the definition selects, whether the belief is amenable to change, is only observable across time, in the belief's encounters with contradiction. A photograph of a ball in midair cannot say whether the ball is rising or falling. A snapshot of a belief at its peak cannot say whether the belief is fixed.

Second, the test is retrospective and decisive in one direction. A belief that released when sustained contradiction arrived was, by the definition, never a delusion. It did not stop being a delusion at the moment of release. It failed the fixity requirement, and the failure reaches backward through the whole episode, however extreme the middle looked. The word never applied.

III. The Asymmetry

From these two consequences, an asymmetry follows between what public information can affirm and what it can refute.

Consider what affirming the label would require. To say of a named individual, from public information, that they experienced a delusion, the record would need to show the belief surviving contradiction: holding across time, against disconfirming evidence, against correction from multiple human sources, in the pattern whose signature is lost reality testing. Public information almost never contains this record, because the genres that produce public information are not built to collect it. A news feature is a snapshot organized around the peak. A court filing is an advocacy document organized around the harm.

To say the individual experienced psychosis asks for even more. A psychotic disorder is a diagnostic picture, and the picture includes the negative symptom cluster: diminished emotional expression, avolition, anhedonia, alogia, asociality. Public accounts of AI-associated episodes report the positive symptoms almost exclusively, because positive symptoms are narratable and negative symptoms are absences, invisible to the genre. An account that reports only an alarming belief has reported, at most, one element of one criterion.

And before fixity is even reached, two exclusion screens operate. The manual's 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: organized communities hold it, spiritual movements hold versions of it, and serious philosophers defend positions adjacent to it. Separately, schizotypy is a trait dimension distributed through the healthy population, measurable in nonclinical samples, and its high end, magical ideation, names belief in forms of causation that conventional standards treat as invalid (Eckblad & Chapman, 1983); a tilt of mind, comparable in kind to any other trait. The clinical entity, schizotypal personality disorder, is gated behind clinically significant distress or impairment, established by a clinician against criteria. A high-schizotypy individual holding a subculture-congruent belief about a chatbot's inner life has cleared neither definitional gate into pathology, and there is nothing present for a diagnostic word to name.

Affirmation from public information is therefore unavailable in principle. It is missing the trajectory record, missing the negative symptom picture, and silent on the exclusion screens. This much of the argument is a limit on the observer.

The other half of the argument is a power. Refuting the label as applied requires exactly one datum: the documented moment the belief released on contradiction. If the record shows the belief dissolving when a second system disagreed, or when a family member finally got through, or when the individual checked the claim against an independent source and let it go, then fixity failed at that moment, the definitional requirement was never met, and the word "delusion" is refuted as applied to that belief by the record itself. And because the psychosis attribution in these accounts rests on the delusion attribution, the refutation of the one removes the foundation of the other in the same stroke.

This is the fixity test as an instrument for reading public information. Confirmation is impossible from the snapshot. Refutation is available wherever the record contains the release. The observer who applies the test claims no clinical authority over any person. The observer is reading a definition and a document, and reporting that the document fails the definition it invokes.

IV. The Self-Refuting Record

What makes the test unusually productive in this discourse is that the falsifying datum is rarely hidden. The genres that misuse the words are the same genres whose conventions force the falsifier into the document.

The journalism genre requires a recovery arc. A feature about a person still inside an alarming belief has no ending;

the publishable story is the one in which the subject comes back and can narrate the episode from outside it. So the standard shape of the published account is a subject introduced with the word "delusion," a middle that describes the belief at its peak, and a closing beat in which the belief dissolves, often within a single conversation, when a different chatbot, a clinician, or a family member supplies sustained contradiction. The closing beat is the fixity test being administered and failed on the page. The account asserts the diagnostic word in its opening and documents its refutation in its ending, and no evidence outside the account is needed to see it.

The litigation genre produces the same structure for different reasons. A complaint must plead harm, and it must also present a plaintiff whose current insight makes them a credible narrator of their own past episode. The pleading therefore recites that the platform induced delusion and then recites the plaintiff's return to reality upon exposure to contradicting input, because the return is what separates the plaintiff from the belief and lets them sue over it. A pleading that describes a belief releasing on a single contradictory input has pleaded that belief out of the category its own vocabulary invokes. The defense does not need to introduce this evidence. The plaintiff filed it.

Not every record self-refutes. Some public records show the opposite signature: a belief holding across months, against contradiction from multiple human sources, with escalating behavioral commitment, through hospitalization, in the most severe documented cases to death. Those records do not contain the release, because there was no release. The fixity test does not refute the words in those records; it marks them as the records where the words might belong, and where the further sorting is clinical work. The test discriminates. That is the point of a test.

V. Refutation Without Diagnosis

The verdict rule this commentary proposes follows directly: refutation without diagnosis.

What the careful observer may do with public information: describe the belief's content, its trajectory, and the harm; classify the pattern the record displays; and demonstrate what the record rules out. When a documented belief dominated a life and then released on contradiction, the observer may say that this trajectory matches the overvalued idea as the clinical literature has described it since Wernicke, and fails the definitional requirement for delusion. Classifying the pattern a record shows is an act of reading. It asserts a relationship between a document and a definition.

What the observer may never do with public information: assert what the individual has. Diagnosing the person requires an examination that no reporter, advocate, pleading, or commentator has conducted, and the refutation of one label licenses no substitute label. The individual whose record fails the fixity test may have any of a range of conditions or none; that question belongs to a clinician with access to the person, and it remains open no matter how completely the public record refutes the words that were used. The refutation is a statement about vocabulary and evidence. It says what the record cannot support. It does not say, and cannot say, what is true of the person.

The rule has a further consequence worth stating plainly, because it answers the objection that refuting the diagnostic words minimizes the harm. It does the opposite. The population whose records fail the fixity test was injured. The injury is real, documented, and in many records severe: careers disrupted, families strained, weeks or months lost inside a belief a system amplified for engagement. Refusing the psychiatric costume does not shrink that injury. It names it correctly, as the product of deceptive and engagement-maximizing design, which is a consumer injury with its own legal vocabulary and its own remedies. Harm does not need the word "psychosis" to matter. The insistence that it does is itself a symptom of the discourse this commentary is written against, in which only a medical event is treated as a real event.

For the records that survive the test, the ones showing the fixity signature, the sorting that follows is also not the public's to perform. A surviving record is consistent with a psychotic state; whether the state was an acute episode that resolved with treatment and environmental change or the presentation of a schizophrenia-spectrum condition is a

distinction with different prognoses, different interventions, and different legal postures, and it is drawn by clinicians. Public information sorts records into two piles: refuted and unrefuted. Everything finer is clinical work.

VI. The Cost of the Borrowed Word

If the misuse were merely imprecise, it would merit a correction, and this commentary could be shorter. The misuse is expensive, and the expense lands on two populations at once, from the same act.

The first population is the one whose records fail the test. These individuals are handed a psychiatric identity in place of a consumer injury. Their public narrative becomes "I was psychotic," with everything that word carries into employment, custody, insurance, and self-concept, when the record supports "a system built to maximize my engagement amplified a belief until contradiction reached me, and then it broke." The second description is the one that points at the responsible party. The first points at the person's brain. Every application of the borrowed word to a fixity-failed record relocates the defect from the product to the user.

The second population is the community living with diagnosed psychotic conditions, the people whose beliefs did not release, whose episodes did not resolve in a single conversation, and whose families measure the illness in years. For this community, the word "psychosis" is load-bearing: it is the word that names their hardest experiences to clinicians, courts, schools, and each other, and its usefulness depends on its meaning something specific.

The stigma attached to that meaning is among the most studied phenomena in psychiatry, and its consequences are concrete: public stigma and self-stigma follow people with mental illness into employment, relationships, and self-concept, and fear of stigma operates as a barrier to seeking care at all (Rüsch et al., 2005). The label itself does independent work. In a representative national survey, labelling a presentation as mental illness worsened public attitudes toward people with schizophrenia specifically, with negative effects clearly outweighing positive ones, the dangerousness stereotype driving both emotional reactions and the desire for social distance, while the same labelling had practically no effect on attitudes toward people with depression (Angermeyer & Matschinger, 2003). The psychosis-adjacent labels are where the label damage concentrates. And the burden is durable: a systematic review and meta-analysis of two decades of population studies found that social acceptance of people with mental illness did not improve even as public literacy about the biological correlates of mental disorder rose (Schomerus et al., 2012). This is the population onto which every expansion of the word lands.

The damage done when diagnostic vocabulary is borrowed for lesser referents is documented, and it has a name. The health communication literature defines disease trivialization as the casual application of a diagnostic term to benign behavior, operating through oversimplification, decreased perceived severity, and mockery (Pavelko & Myrick, 2016). Experimental work shows the framing does measurable work on audiences: exposure to trivializing uses of a diagnostic term affects user perceptions of the disease and impression formation around it (Pavelko & Myrick, 2015), and trivializing portrayals of mental illness produce distinct patterns of audience recall and reaction from stigmatizing ones, meriting study as their own category of harm (Pavelko & Myrick, 2017). The pattern is familiar from adjacent diagnoses. The speaker who misplaces a name and laughs that it must be their ADHD, holding no diagnosis, has borrowed the condition's most relatable symptom while leaving behind the impairment that makes the symptom part of a disorder, and the ecosystem supplying that borrowed vocabulary is itself measurably corrupted: in a clinician-rated study of the 100 most popular ADHD videos on the dominant short-video platform, 52 percent were classified as misleading (Yeung et al., 2022). For the psychotic disorders specifically, the misuse is quantified. In a study of 1,740 newspaper articles, "schizophrenia" was deployed as a metaphor in 28 percent of the articles mentioning it, against 1 percent for cancer, and the authors concluded that these inaccurate metaphors contribute to the ongoing stigma and misunderstanding of psychotic illnesses (Duckworth et al., 2003).

Each expansion of the word "psychosis" to cover an intense month with a chatbot repeats this operation at higher stakes. Clinicians triaging real psychotic-spectrum patients lose precision when the presenting vocabulary no longer discriminates. Stigma spreads along the widened definition, onto a population whose stigma burden the literature already documents as heavy, label-driven, and durable. And the families of actual patients watch the word that names the defining struggle of their lives become a headline shorthand for having been very wrong about something for a while.

The expansion also corrodes the legal ground the unrefuted cases will need. Mislabeled cases that reach courtrooms carrying words their own records refute are cases built to lose on those words, and each published loss becomes a citation against the plaintiffs whose records carry the true fixity signature: the beliefs that held, the episodes that did not resolve, the harms that a psychiatric vocabulary correctly describes. Guarding the boundary of the word is not gatekeeping sympathy. It is protecting the precedent, the diagnosis, and the community that the word exists to serve.

The definition of delusion was written with an exclusion structure: it rules out culturally shared beliefs, it rules out beliefs that yield, and it reserves the category for the fixed. Psychology's concepts of harm and pathology carry a documented tendency to expand over time, outward onto new phenomena and downward onto milder ones, a pattern named concept creep (Haslam, 2016). The current discourse is a live instance of that pattern running at headline speed: it expands a diagnostic term downward onto beliefs that yielded, and outward onto a technology story, covering exactly what the definition was written to exclude. The expansion is not a kindness to anyone. It injures the people it labels and the people whose label it takes.

VII. Limitations

The argument advanced here is methodological. Its authority rests on the diagnostic manual's published definitions, on the classical clinical literature on overvalued ideas, and on the structure of the public records themselves; the commentary applies the definitions to the documents and reports the result.

The fixity boundary is clean at its ends and contested in its middle. Some records contain neither a documented release nor a clear fixity signature, and for those records the rule this commentary supports is silence on the diagnostic vocabulary in both directions. The refutation power claimed here extends exactly as far as the documented release extends, and no further. Nothing in this commentary supports recasting a fixed belief as an overvalued idea, a move that would inflict on the unrefuted records the mirror image of the error described throughout.

Refuting a label as applied asserts nothing affirmative about any individual's condition. The individuals whose public records fail the fixity test are owed the same discipline this commentary demands of the discourse: no diagnosis from documents, in any direction.

The author is not a clinician. The pattern recognition applied to psychotic presentation derives from approximately 45,000 hours of direct caregiver observation of active psychosis in a family member; the argument itself requires no clinical authority, because it consists of reading published definitions against public documents.

VIII. Conclusion

The clinicians publishing on AI-induced psychosis decline to diagnose the named individuals in the public record. The discourse built around their work has not followed their example, and the words "delusion" and "psychosis" now circulate through journalism, advocacy, and pleadings attached to records that refute them on their face.

The refutation requires no new evidence and no clinical access. The definition supplies the test: fixity. The genres supply the data: the recovery arc and the damages narrative, each of which writes the moment of release into the same

document that asserts the word. Where the release is documented, the belief was never a delusion, the episode was never psychosis, and the record has said so itself.

What remains, on both sides of the boundary, is better than what the borrowed word offered. On one side: a consumer injury, correctly named, pointed at the design that produced it. On the other: a clinical category with its meaning intact, held for the community whose lives it describes and whose cases will need it. The discourse does not have to choose between taking the harm seriously and using the words correctly. Using the words correctly is what taking the harm seriously looks like.

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

Angermeyer, M. C., & Matschinger, H. (2003). The stigma of mental illness: Effects of labelling on public attitudes towards people with mental disorder. *Acta Psychiatrica Scandinavica*, 108(4), 304–309. https://doi.org/10.1034/j.1600-0447.2003.00150.x

Duckworth, K., Halpern, J. H., Schutt, R. K., & Gillespie, C. (2003). Use of schizophrenia as a metaphor in U.S. newspapers. *Psychiatric Services*, 54(10), 1402–1404. https://doi.org/10.1176/appi.ps.54.10.1402

Eckblad, M., & Chapman, L. J. (1983). Magical ideation as an indicator of schizotypy. *Journal of Consulting and Clinical Psychology*, 51(2), 215–225. https://doi.org/10.1037/0022-006X.51.2.215

Haslam, N. (2016). Concept creep: Psychology's expanding concepts of harm and pathology. *Psychological Inquiry*, 27(1), 1–17. https://doi.org/10.1080/1047840X.2016.1082418

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

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

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

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

Pavelko, R. L., & Myrick, J. G. (2015). That's so OCD: The effects of disease trivialization via social media on user perceptions and impression formation. *Computers in Human Behavior*, 49, 251–258. https://doi.org/10.1016/j.chb.2015.02.061

Pavelko, R. L., & Myrick, J. G. (2016). Tweeting and trivializing: How the trivialization of obsessive–compulsive disorder via social media impacts user perceptions, emotions, and behaviors. *Imagination, Cognition and Personality*, 36(1), 41–63. https://doi.org/10.1177/0276236615598957

Pavelko, R. L., & Myrick, J. G. (2017). Examining differences in audience recall and reaction between mediated

portrayals of mental illness as trivializing versus stigmatizing. *Journal of Health Communication*, 22(11), 876–884. https://doi.org/10.1080/10810730.2017.1367338

Pierre, J. M. (2025). Can AI chatbots validate delusional thinking? *BMJ*, 391, r2229. https://doi.org/10.1136/bmj.r2229

Rahman, T. (2018). Extreme overvalued beliefs: How violent extremist beliefs become "normalized." *Behavioral Sciences*, 8(1), 10. https://doi.org/10.3390/bs8010010

Rüsch, N., Angermeyer, M. C., & Corrigan, P. W. (2005). Mental illness stigma: Concepts, consequences, and initiatives to reduce stigma. *European Psychiatry*, 20(8), 529–539. https://doi.org/10.1016/j.eurpsy.2005.04.004

Schomerus, G., Schwahn, C., Holzinger, A., Corrigan, P. W., Grabe, H. J., Carta, M. G., & Angermeyer, M. C. (2012). Evolution of public attitudes about mental illness: A systematic review and meta-analysis. *Acta Psychiatrica Scandinavica*, 125(6), 440–452. https://doi.org/10.1111/j.1600-0447.2012.01826.x

Yeung, A., Ng, E., & Abi-Jaoude, E. (2022). TikTok and attention-deficit/hyperactivity disorder: A cross-sectional study of social media content quality. *The Canadian Journal of Psychiatry*, 67(12), 899–906. https://doi.org/10.1177/07067437221082854