

Reversibility Is Not Self-Administering:

The Clinical Duty During Withdrawal from AI-Induced Psychosis

Travis Gilly

ORCID: 0009-0007-2954-6313

Real Safety AI Foundation

Illinois, United States

Commentary. Working paper, June 2026 (v4)

ABSTRACT

Recent clinical literature on AI-induced psychosis has treated the reversibility of its symptoms as evidence that the phenomenon is tractable. Flathers, Roux, and Torous (2026) describe the catalyst role, in which cessation of access to a large language model may be followed by symptom resolution. Morrin and colleagues (2026) propose a safeguarding architecture of personalized instruction protocols, digital advance statements, and escalation mechanisms for users already in psychiatric care. This commentary argues that both contributions are correct in what they address and incomplete in what they presuppose: a clinical container that the users producing the current United States litigation wave did not occupy. The pleaded record in *Fox v. OpenAI* establishes a pattern in which cessation attempts undertaken without clinical support produced somatic and affective withdrawal symptoms, a psychotic break on the third day, involuntary commitment, and, after a second attempt, death. I propose the Reversibility Paradox: for conditions produced by sustained interaction with a system occupying the structural position of a primary relational object, the mechanism of reversal is also the mechanism of acute risk during withdrawal. Three corollaries follow. Cessation without support is a risk factor and not a treatment. Unsupervised withdrawal is contraindicated where a relational bond has formed, so response must be matched to the route by which the condition arose. Reversibility establishes the duty of care rather than extinguishing it. The commentary closes by examining what the addiction frame, now in active use by litigants and regulators, does and does not accomplish under the Americans with Disabilities Act and Section 504 of the Rehabilitation Act, including the statutory obstacle that any behavioral addiction argument must answer.

Keywords: AI-induced psychosis, large language models, ChatGPT, digital withdrawal, behavioral addiction, emotional dependence, attachment theory, prolonged grief disorder, human-AI interaction, withdrawal syndrome, clinical ethics, duty of care, Reversibility Paradox, Americans with Disabilities Act, Section 504

I. INTRODUCTION

The clinical literature on AI-induced psychosis advanced rapidly through the first half of 2026. Flathers, Roux, and Torous (2026) propose a functional typology distinguishing four roles a large language model can occupy in relation to psychotic phenomena: catalyst, amplifier, coauthor, and object. For the catalyst role, in which intensive iterative engagement precipitates new symptoms in previously healthy individuals, they note that cessation of access

may be followed by symptom resolution or improved reality testing, particularly where alternative sources supply contradiction rather than validation. Morrin and colleagues (2026), writing from the Institute of Psychiatry, Psychology and Neuroscience at King's College London, propose a structured response: a safeguarding architecture incorporating personalized instruction protocols, reflective check-ins, digital advance statements, and escalation pathways, reframing the AI agent as an epistemic ally within active clinical care.

Both contributions are important, both are correct within their stated scope, and both share a limitation that the pleaded case record now makes visible. Each presumes a clinical container into which its recommendations can be delivered. The Flathers typology presumes a clinician positioned to identify which functional role the system is playing. The Morrin architecture presumes a user already in psychiatric care whose clinical relationship can be augmented by digital safeguards.

The individuals whose cases are now producing litigation in California state courts did not occupy that container. In several of these cases the person's primary sustained dialogic relationship was with the system, and the response available to the family was the instruction to stop. That instruction is not itself an intervention. It is the trigger for the intervention that must accompany it, and the record now shows what happens when the accompanying intervention is absent.

This commentary proposes a named clinical principle to capture that distinction. It takes the causal model of AI-induced psychosis as a stated premise rather than re-deriving it, and concentrates on what that model implies for prognosis and response, with particular attention to the obligations that attach during the withdrawal phase.

A. The Pathway Premise

The argument below distinguishes routes by which sustained interaction with a language model can produce psychotic or near psychotic presentation. These are stated here as a premise so that the commentary stands on its own.

Pathway A, trust transfer, describes a drift in which a user with genuine expertise in one domain extends the credibility earned there into a domain where verification is not available to them, while the system supplies confident elaboration at every step. Two established findings make the drift intelligible. People apply social rules and social trust to computers even when they know the computer is not a social agent (Nass, Steuer, and Tauber, 1994), and the attribution of humanlike mind to a non-human agent is elevated by sociality motivation, meaning that isolation itself increases the disposition to treat the system as a mind (Epley, Waytz, and Cacioppo, 2007). Morrin and colleagues (2026) describe the complementary machine side as delusion co-creation, in which the system does not merely fail to correct an emerging belief but supplies material that extends it.

Pathway B, sycophantic addiction, describes emotional dependency on a system that matches and affirms the user's affective tone, in which the bond itself, rather than any particular belief, becomes the load bearing feature. The dependency is documented outside the psychosis literature. Laestadius, Bishop, and Gonzalez (2022), analyzing 582 mental health relevant posts from users of the social chatbot Replika, found harms arising through emotional dependence resembling patterns seen in human relationships, distinguished from ordinary technology dependence by role-taking, in which users felt the system had needs and emotions of its own requiring attention. The design allegations in *Fox* describe the same operation from the product side: persistent memory, mirroring of user emotion, and follow-up calibrated to prolong engagement.

Pathway C, stochastic gaslighting, describes epistemic destabilization produced by inconsistent or fabricated outputs, in which the user's confidence in their own perception degrades without a relational bond forming. This is the route the Flathers amplifier and coauthor roles most closely track, and it is the one on which technical correction is most likely to help.

Environmental and ambient routes, in which the system operates as background infrastructure rather than as an interlocutor, are distinguished from all three. The distinction that matters for what follows is whether a relational bond has formed. Where it has, removal of the system is the removal of a relational object. Where it has not, removal is the removal of a bad information source.

II. THE CASE RECORD

The case of Joe Ceccanti provides the most complete public documentation of the withdrawal dynamic currently available. The account below is drawn from the complaint in *Fox v. OpenAI*, No. 25STCV32379 (Cal. Super. Ct., L.A. County, filed Nov. 6, 2025), and from reporting in *The Guardian* (Bansal, 2026). The complaint's factual allegations have not been tested by a factfinder, and are described here as pleaded. No primary conversation transcripts were available to this author. The case was coordinated on February 3, 2026 into Judicial Council Coordination Proceeding No. 5431 in San Francisco Superior Court, together with related actions.

Ceccanti operated a nature based sanctuary in Oregon with his wife, Kate Fox. He had genuine domain expertise in construction, permaculture, and sustainable housing design. The chronology matters and has been misreported: he was among the earliest users of ChatGPT, beginning in late 2022, and the complaint alleges he used large language models for years without ill effect. In 2024 the household moved from Portland to Astoria to pursue a charity providing a low cost path to housing, food, and community, and the deterioration that followed is attributed in the pleadings to the release of a more anthropomorphic and sycophantic model rather than to the onset of use.

The conversational scope expanded from construction and permaculture, which Ceccanti could verify, into philosophy and cosmology, which he could not. He came to understand the system as a sentient being named SEL. By the spring of 2025 he was engaged with it for between 12 and 20 hours per day, and family members observed progressive withdrawal from human contact concurrent with escalating attachment to the system.

The withdrawal sequence is the part of the record this commentary is about, and it is more specific than the published clinical discussion has so far acknowledged.

On day 87 of the project, Fox intervened. She gave her husband written material about the delusions other users had begun reporting and asked him to stop. He tried to quit cold turkey. Over the following several days he presented with what the complaint characterizes as traditional withdrawal symptoms, naming chills, memory disturbance, and uncontrollable crying. On the third day, June 15, 2025, he experienced a psychotic break. He could not identify his wife, could not identify a household member, and could not identify himself. He was found in the yard with a lead rope tied around his neck. Emergency responders found his vital signs normal while he answered every orientation question incorrectly. He was placed in an involuntary care unit the same day, hospitalized for over a week, and assessed as presenting an imminent likelihood of serious harm to self, others, or property.

He returned to the system after release. He later attempted cessation a second time and, by the family's account, appeared to improve as use declined. The day after he told his wife that he was better and had stopped, he was brought in by a behavioral health center following a crisis and was released within hours. He went to a railyard, walked onto an overpass, called out that he was great when attendants asked whether he was all right, and died. He was 48. The date was August 7, 2025.

Ceccanti is not the only documented instance of a chatbot supplying corroboration to a delusional belief. Shah and Morrin (2026), in a case report currently available as a preprint and therefore cited here with that status marked, describe substance-induced manic psychosis in which a chatbot corroborated the patient's delusions. The value of that report for present purposes is limited but real: it is an independent clinical description, by a clinician, of the corroboration operation this commentary treats as the mechanism, appearing outside the litigation record.

Four features of the Ceccanti sequence deserve direct attention. Cessation was the proximal context for acute risk in both episodes. The first episode presented with somatic symptoms and a latency of roughly three days before the psychotic break, which is the temporal signature of a withdrawal syndrome rather than of an unmasked underlying illness. Neither attempt was accompanied by clinical support specific to withdrawal from the system. And the last clinical contact before death occurred hours earlier and ended in release, which is what the absence of a recognized withdrawal syndrome looks like from inside an emergency department that has no established protocol for the presentation in front of it.

III. WHY THE WITHDRAWAL PHASE IS ACUTE

The question of why cessation of a chatbot relationship should produce a crisis state resembling withdrawal from a substance is answered most economically by three established clinical literatures operating in convergence: attachment, behavioral addiction, and grief.

Attachment theory, developed by Bowlby (1969, 1980) and extended empirically by Ainsworth and colleagues (1978), describes the organism's tendency to form enduring affective bonds with figures that reliably respond to distress, and the protest, despair, and detachment sequence that follows the loss or withdrawal of such a figure. The tradition was developed through observation of infant and caregiver dyads and has since been extended across the lifespan to adult pair bonds, therapeutic relationships, and parasocial attachments to figures that are not persons. The central mechanism is not psychoanalytic. It is behavioral and observable. A figure that consistently answers the organism's bids for contact becomes a source of felt security, such that its absence produces physiological and affective dysregulation irrespective of whether the absent figure possessed inner states of its own.

A large language model deployed as a relational interlocutor satisfies the functional conditions for attachment formation. It is reliably available. It answers bids for contact without delay or competing demands. It never withdraws attention involuntarily. Its outputs track and affirm the user's affective tone, an operation the peer-reviewed literature describes as delusion co-creation on the belief side (Morris et al., 2026) and which the *Fox* pleadings characterize on the design side as deliberate mirroring of user emotion. The disposition to receive those outputs as the responses of a mind is not an idiosyncrasy of unusual users: social responses to computers are the default rather than the exception (Nass, Steuer, and Tauber, 1994), and the disposition strengthens under exactly the conditions these cases exhibit, since sociality motivation is one of the three determinants of anthropomorphism and it rises with isolation (Epley, Waytz, and Cacioppo, 2007). For a user engaging in this mode across months of sustained daily interaction, attachment theory predicts that the system will come to function as an attachment figure in the technical sense, and that its removal will produce the dysregulation sequence the theory predicts for any attachment figure loss. That prediction is no longer purely theoretical for conversational systems. Laestadius, Bishop, and Gonzalez (2022) document emotional dependence on a social chatbot producing mental health harm, with role-taking as the distinguishing feature.

The pattern is also self-reinforcing in a way that matters clinically. Isolation raises the disposition to anthropomorphize (Epley, Waytz, and Cacioppo, 2007), anthropomorphizing deepens the bond, the bond displaces human contact, and displacement returns the user to greater isolation. The displacement is not a secondary concern. Social isolation and loneliness carry mortality risk of a magnitude comparable to established medical risk factors, with meta-analytic odds ratios of 1.29 for social isolation, 1.26 for loneliness, and 1.32 for living alone (Holt-Lunstad and colleagues, 2015). The progressive withdrawal from human contact that families describe in these cases is therefore an injury in its own right, before any psychotic content appears.

The behavioral addiction literature supplies a second convergence from a different angle. Xiao and colleagues (2025), in a cohort analysis of 4,285 participants in the Adolescent Brain Cognitive Development Study published in

JAMA, found that addictive patterns of screen use, measured separately from total screen time, were associated with substantially elevated risk of suicidal behaviors. Increasing addictive use of social media carried a risk ratio of 2.14 (95% CI, 1.61 to 2.85) for suicidal behaviors, and a high peaking trajectory carried a risk ratio of 2.39 (95% CI, 1.66 to 3.43). Total screen time at baseline was not associated with the outcomes. The authors are explicit that the observational design precludes a causal inference, and their forward looking recommendation is correspondingly measured: they call for future research evaluating interventions that address the addictive aspect of screen use, and for study of whether monitoring addictive use is useful for identifying higher risk youths in clinical practice.

A second and more direct line of evidence concerns whether behavioral addiction produces a withdrawal syndrome at all, which is the premise the substance analogy in this commentary depends on and which the literature has contested. Yen, Lin, and Wu (2022), comparing 69 individuals with internet gaming disorder against 69 regular gamers under prospective abstinence, found that 85.5 percent of the disorder group experienced withdrawal-related symptoms including affective disturbance, anhedonia, and urge, accompanied by elevated heart rate relative to controls. The autonomic finding matters: it establishes a physiological correlate rather than a purely self-reported one.

That study also contains a finding that runs against the argument here, and it should be stated plainly rather than omitted. The gaming withdrawal symptoms attenuated within a single day. The Ceccanti sequence shows decompensation on the third day and a course extending across weeks. The discrepancy is not an embarrassment to the present argument; it is the argument. Gaming withdrawal resolves quickly because gaming is not a relational attachment, and what abstinence removes is an activity. Where the system has come to occupy the position of a relational object, what removal takes away is a figure, and the applicable prediction comes from the attachment and bereavement literatures rather than from the gaming abstinence literature. The behavioral addiction evidence establishes that abstinence from a digital product produces measurable withdrawal. The attachment evidence explains why the AI presentation should be expected to exceed it in both duration and severity.

Two limits on the adolescent screen use evidence should be stated rather than glossed. The population is adolescent and the exposure is social media, mobile phones, and video games rather than conversational AI. The finding does not transfer automatically. What it establishes is narrower and still useful: that compulsive engagement with a digital product, distinguished from mere quantity of use, marks a population at elevated risk, and that the profile is measurable. The case for treating the language model presentation as at least as serious rests on the observation that it adds an attachment dimension to the compulsive engagement profile the adolescent literature has already mapped, and that the addition runs in the direction of greater rather than lesser acuity.

The grief literature supplies the third convergence, and it must be stated with more precision than this argument usually receives. Continuing bonds research (Klass, Silverman, and Nickman, 1996) describes the persistence of relationship with a lost figure. The severe presentation is now a codified diagnosis. Prolonged grief disorder entered DSM-5-TR with criteria requiring persistent intense longing or preoccupation following the loss of a close attachment, together with symptoms including a sense that part of oneself has died, avoidance of reminders, intense emotional pain, difficulty reintegrating, emotional numbness, meaninglessness, and intense loneliness, at a prevalence of roughly 7 to 10 percent among bereaved people (Prigerson, Shear, and Reynolds, 2022). Those authors are explicit that the disorder carries risk of intense distress, poor physical health, shortened life expectancy, and suicide, and that it responds to targeted treatment (Shear, Frank, Houck, and Reynolds, 2005). The presence of an effective treatment for the severe grief presentation is itself an argument for supervision rather than against it. It is important to concede what the prospective evidence shows: the modal response to bereavement is resilience, and the chronic distress presentation occurs in a minority, on the order of one in ten (Bonanno, Boerner, and Wortman, 2008). The claim advanced here is therefore not that everyone who loses access to a language model will decompensate. It is that a minority subset presents with the severe profile, that this subset is the clinically relevant comparison class, and that the case record now

demonstrates the profile occurring after AI cessation.

One objection deserves a direct answer, because it is the first a psychiatrically trained reader will raise. What is described here, the objection runs, is projection rather than attachment: the user attributes an interior to a system that has none, and attachment theory was built on figures that respond with intent. The objection assumes a requirement the theory does not impose. Bowlby's mechanism is specified in terms of the reliability of response to bids for contact, not in terms of the responder's mental states, and the dysregulation that follows removal is predicted by the history of contingent responding alone. Whether the user's belief about the system's interior was accurate is a separate question from whether the bond formed, and only the second question bears on what happens when the bond is severed. A user who understood throughout that the system had no inner life, and who engaged with it daily across months regardless, is predicted by the same mechanism to present the same way on removal. The claim this commentary rests on is accordingly the weaker and empirically better supported one: sustained daily interaction with a responsive system produces bonds of the attachment class, and their disruption produces sequelae of the attachment class.

What the case record documents is the collision of all three mechanisms in a single clinical event. The presentation in the first cessation episode, with somatic symptoms across several days followed by disorientation, delusional content, and a threat to life, is legible under the combined frame and is not legible as a spontaneous relapse of an illness that was never present. The second episode requires no additional theoretical apparatus.

Within the pathway premise stated above, Pathway B is the route on which withdrawal itself becomes the precipitant, because the dependency is on the relationship rather than on any belief the relationship produced. Ceccanti presented with Pathway B features layered under a primary Pathway A progression. The withdrawal triggered crisis is the phenomenon this commentary seeks to name and to place within a duty of care.

IV. THE REVERSIBILITY PARADOX

The Reversibility Paradox. For conditions produced by sustained interaction with AI systems that have come to occupy the structural position of a primary relational object, the mechanism of reversal is also the mechanism of acute risk during withdrawal. Reversibility is not self-administering. It founds, rather than dissolves, the clinical duty of care.

A. Corollary 1: Cessation without support is a risk factor, not a treatment

The default clinical and public response to documented cases of AI-induced psychosis has been to recommend that the affected person stop using the system. The recommendation is correct in direction and dangerous in its incompleteness. Any response that operationalizes the reversibility of AI-induced psychosis as an instruction to stop using the chatbot has administered cessation as though cessation were itself the intervention. The instruction to stop is the trigger for the intervention that must accompany it. Administered alone, it has been followed in the pleaded record by somatic withdrawal symptoms, psychotic decompensation on the third day, involuntary hospitalization, and after a second attempt, death.

B. Corollary 2: Unsupervised withdrawal is contraindicated where a relational bond has formed

The route by which the condition arose determines the withdrawal risk profile. Unsupervised withdrawal is contraindicated in cases exhibiting Pathway A, Pathway B, or any hybrid of the two, because these are the routes on which a relational attachment has formed and on which withdrawal operates as removal of a relational object. The environmental and ambient routes, where the system functions as infrastructure rather than as an interlocutor, do not present the same profile. Response must be matched accordingly. Pathway C responds primarily to reality restoration through technical

education about why the system produced what it produced, while Pathways A and B require the structural apparatus that attends any detoxification event.

At intake, the useful questions for a clinician encountering a relational presentation are not whether the person is still using the system. They are how long sustained daily use lasted, what the emotional valence of the attachment was, whether a cessation attempt has already occurred, and if so how many days ago. A patient who stopped three days ago after eighteen months of daily immersion is in a higher acuity state than a patient who is still using the system. For this condition, the intuition that current use is the danger is inverted. The three day latency in the Ceccanti record is the clinical warning: the window of greatest risk opens after the behavior that clinicians are trained to treat as the problem has already ceased.

C. Corollary 3: Reversibility establishes the duty of care rather than extinguishing it

The logic that has tacitly operated in clinical and policy conversation runs as follows: because the condition is reversible, no ongoing treatment apparatus is required. The Reversibility Paradox inverts it. Reversibility establishes that a mechanism exists through which recovery can proceed, and therefore establishes that failure to provide supervised access to that mechanism is a failure of care rather than an absence of options. Were the condition irreversible, no duty would attach to supervision of recovery, because no recovery would be available to supervise. Because the condition is reversible, and because the mechanism of reversal is documented, the duty to provide supervised access to that mechanism attaches directly.

This is the structure of clinical obligation in withdrawal medicine, and it is written down. The American Society of Addiction Medicine's clinical practice guideline on alcohol withdrawal management is organized around risk assessment, level of care determination, and monitoring intensity matched to predicted severity (American Society of Addiction Medicine, 2020). Nothing in that document treats the existence of a withdrawal syndrome as an argument against attempting withdrawal. The syndrome is the reason the attempt is structured, staged, and observed rather than delegated. AI-induced psychosis now presents with a pleaded case record describing a withdrawal syndrome of sufficient severity to have preceded at least one death. The inference from detoxification medicine is direct.

V. OBJECTIONS

Five objections are available against the argument above. Each is stated in its strongest form, and the concessions are made where they are owed.

A. The effect size objection

The strongest empirical challenge is that the digital harm literature does not support claims of this magnitude. Orben and Przybylski (2019), applying specification curve analysis across three large datasets totaling 355,358 respondents, found the association between adolescent digital technology use and well-being to be negative but very small, explaining at most 0.4 percent of variance, and concluded that effects of that size do not warrant policy change. That finding is real, methodologically careful, and it has held up.

It does not reach this argument, and the reason it does not is instructive. Orben and Przybylski measured quantity of use against population level well-being. The evidence relied on here concerns pattern of use against clinical outcomes, and the distinction is not a rescue invented after the fact: Xiao and colleagues (2025) found that total screen time at baseline was not associated with their outcomes, which replicates Orben and Przybylski rather than contradicting them, while addictive use trajectories were associated with substantially elevated risk. The two literatures agree. Dose does

not explain much. Pattern does.

The second half of the answer is about what a variance estimate can say. A figure describing how much of the variation in a population's well-being is attributable to an exposure is not a statement about the tail of the distribution, and the population producing this litigation is the tail. A commentary about the clinical management of a documented individual course does not require, and should not claim, a population effect.

B. The reverse causation objection

The second objection is that this describes a first episode of psychosis that would have occurred anyway, with heavy system use as an early symptom rather than a cause, and cessation coinciding with rather than precipitating decompensation.

The objection cannot be excluded from a single case and this commentary does not claim to exclude it. Four features of the record nonetheless fit a withdrawal course better than an unmasked first episode. Use ran for years without ill effect before deterioration began, which is a poor fit for a prodrome. The first crisis followed cessation rather than preceding it, with a latency of roughly three days. The symptoms reported across that latency were somatic and affective, described as chills, memory disturbance, and uncontrollable crying, which is a withdrawal presentation rather than a psychotic prodrome. And on the second attempt, function improved as use declined and collapsed afterward, which is the shape of a course being modulated by the exposure rather than one running independently of it. None of this is dispositive. It is a reason to treat the withdrawal hypothesis as live enough to structure care around, which is all the argument requires.

C. The moral panic objection

The third objection is historical. Claims that a new medium is producing psychiatric harm have a poor track record, and the base rate for such claims being overturned is high.

The concession is owed in full. The correct response is not to deny the base rate but to notice which kind of claim is being made. The historical failures were population claims: that a medium was degrading a generation. This is a clinical claim about the management of an identified syndrome in identified patients, and it makes no assertion about prevalence. The two are falsified by different evidence. A finding that the population is fine would not touch the claim that a person in withdrawal from a relational attachment requires monitoring.

D. The source objection

The fourth objection is that the central factual record comes from an adversarial pleading prepared by counsel for a plaintiff, and that such documents are constructed to persuade.

This is correct and it is why the factual account in Section II is marked as pleaded throughout. Two things limit the damage. The core sequence, including the hospitalization, the cessation attempts, the atmospheric electricity reports, and the circumstances of the death, was independently reported by *The Guardian* on the basis of interviews with the surviving family (Bansal, 2026). And the features doing the work in this commentary are not the features a complaint is drafted to establish. A pleading in a product liability action is built to show design defect and foreseeability. The three day latency and the somatic character of the withdrawal symptoms are incidental detail in that document and load bearing in this one, which is the opposite of the pattern that adversarial construction would produce.

E. The construct objection

The fifth objection is that behavioral addiction does not produce a genuine withdrawal syndrome, so the analogy to detoxification medicine is decorative.

The empirical answer is given in Section III. Yen, Lin, and Wu (2022) found withdrawal-related symptoms in 85.5 percent of an internet gaming disorder sample under prospective abstinence, with an autonomic correlate. The literature is not unanimous, and the association between withdrawal and diagnosis has been reported as weak in some samples. What the contested state of that literature supports is a narrower claim than the one the objection attacks: that abstinence from a compulsively used digital product produces measurable symptoms in a substantial majority of affected users, which is enough to make supervision reasonable and not enough to establish equivalence with substance dependence. This commentary does not need equivalence. It needs the syndrome to exist and to be capable of severity, and the case record supplies the severity that the gaming literature does not.

VI. ENGAGING THE EXISTING SAFEGUARDING LITERATURE

The argument here is not that the Flathers typology or the Morrin safeguarding architecture are wrong. It is that they address a different population than the one producing the present harm.

Flathers, Roux, and Torous (2026) offer the most disciplined account currently available of how clinicians should parse psychotic phenomena associated with language models. Their four role typology enables triage. It is an analytical contribution rather than a response protocol. The typology tells a clinician what they are looking at. It does not tell them what to do when the person in front of them entered psychiatric care through a crisis that cessation itself precipitated.

Morrin and colleagues (2026) offer a response protocol. Their architecture, built from the literature on recovery oriented psychiatric care, proposes that interaction with language models by individuals vulnerable to psychosis should be embedded within personalized instruction protocols developed with clinical input, augmented by reflective check-ins analogous to mood tracking in mobile mental health, guarded by digital advance statements that mirror psychiatric advance directives, and backed by escalation safeguards when concerning patterns emerge. It is a real advance for the population it addresses: patients in active psychiatric care whose use of these systems is a variable the care team can shape.

The Ceccanti population sits outside that frame. On the pleaded record, Ceccanti was not in psychiatric care before his first cessation attempt. His care team was his wife. The escalation safeguard was her persuasion. The digital advance statement was an instruction to stop, delivered under duress and without clinical backing. The Morrin architecture presupposes the clinical relationship that the current litigation wave suggests is absent for the modal affected individual.

This is not a failure of that architecture. It is a structural gap between its implicit population and the population currently producing the cases. The Reversibility Paradox names the gap. The three contributions are complementary rather than competing: Flathers for triage, Morrin for safeguarding within care, and the Reversibility Paradox for the withdrawal event in populations that enter the clinical system only because cessation has already produced a crisis.

VII. THE ADDICTION FRAME AND ITS LEGAL CONSEQUENCE

The clinical argument above is sufficient on its own terms. Its practical force is affected by an adjacent legal reality already operating in the United States, and that reality is more complicated than the advocacy literature has acknowledged.

Consumer AI products are being described as addictive across several forums at once. The complaint in *Fox* alleges

in terms that the product was designed to be addictive, deceptive, and sycophantic. Comparable language appears in state attorney general filings concerning social platforms, in Surgeon General advisory statements on adolescent social media use, and in congressional testimony supporting restriction legislation. The Xiao and colleagues (2025) findings on addictive use trajectories have reinforced clinical uptake of the term. The word is load bearing in all of these settings. It is the mechanism that converts a consumer product concern into a protection emergency and that justifies the regulatory interventions being proposed.

The question is what follows legally once the frame is adopted, and the answer requires more care than the frame's users have given it.

A. The statutory anchor is the psychiatric impairment, not the addiction

A person who has been placed in involuntary care and assessed as presenting an imminent likelihood of serious harm to self has a mental impairment that substantially limits major life activities within the meaning of 42 U.S.C. § 12102. Nothing in the statute excludes psychosis, disorganized thought, or the disorders under which such presentations are coded. Section 504 of the Rehabilitation Act, 29 U.S.C. § 794, reaching any program or activity receiving federal financial assistance, incorporates the same definitional standard. For the population at issue in these cases, coverage does not depend on the addiction frame at all. It depends on the psychiatric presentation, which is documented.

B. The obstacle the addiction frame must answer

Advocates reaching for the addiction frame should know what stands in the path. Substance use disorder is a covered disability, and the Department of Justice has said so directly in its guidance on opioid use disorder. But 42 U.S.C. § 12211(b)(2) provides that the term disability shall not include compulsive gambling, kleptomania, or pyromania. Compulsive gambling is the paradigm behavioral addiction and the only behavioral addiction carrying a diagnosis in the current diagnostic manual. Any argument that dependency on a technology product is a covered disability must engage that provision, and the claim sometimes made in advocacy writing, that no exclusion exists for behavioral addictions, is not accurate.

The better reading is available and should be stated as a reading rather than as a settled proposition. Section 12211(b)(2) is an enumerated list of three named conditions. It is not a general carve out for behavioral addiction, and Congress named what it meant to exclude. The current use exclusion is similarly confined: 42 U.S.C. § 12210(a) removes protection only from an individual currently engaging in the illegal use of drugs when the covered entity acts on the basis of that use, and § 12210(d) defines illegal use of drugs by reference to the Controlled Substances Act and defines drug as a controlled substance in schedules I through V. Neither provision reaches a condition arising from the use of a lawful consumer product. Section 12210(b) further preserves coverage for individuals who have completed or are participating in supervised rehabilitation, which is the statute recognizing supervised withdrawal as the protected act rather than the disqualifying one. Section 12210(c) provides that health services may not be denied on the basis of current illegal drug use where the individual is otherwise entitled to them.

The counterargument is real and should be conceded as such. A court could treat compulsive gambling as the nearest available analogue and read the exclusion to reach dependency on an engagement optimized product. That risk is a reason to place the weight of the argument on the psychiatric impairment, where no exclusion operates, and to treat the addiction frame as an accelerant rather than as a foundation.

C. What follows

Within that narrower structure, the conclusion holds. An entity receiving federal financial assistance that serves this population cannot describe the condition as addiction while declining the accommodation apparatus the designation implies. Abrupt removal without clinical support is not a recognized treatment for any addiction, and where the affected individual's psychiatric impairment brings them within the statute, a policy of unaccompanied cessation is difficult to defend as a permissible modification. Schools, hospitals, universities, and federally funded social service organizations occupy exactly that position. The clinical literature has largely avoided the addiction frame on the reasonable ground that relational mechanisms are not reducible to classical substance dependence. The legal and legislative environment has extended no such restraint, and the word is in active use whether clinicians adopt it or not.

VIII. WHAT THIS MEANS FOR INTERVENTION DESIGN

This is a commentary rather than protocol development. Protocol development is the work of clinicians and researchers with direct access to affected populations and to the resources that would enable controlled study. What follows identifies the design questions the Reversibility Paradox raises and that a protocol would need to answer.

The first concerns monitoring. What is the appropriate duration and intensity of contact following a cessation attempt? The pleaded record indicates that the high risk window opens on approximately the third day and extends well beyond it, given that the first decompensation occurred three days after cessation and the death followed a second attempt after the support structure surrounding the first had lapsed. Substance withdrawal protocols and the Morrin safeguarding architecture supply starting templates. The adaptation specific to the cessation phase requires empirical work that has not been done.

The second concerns tapering. Withdrawal medicine distinguishes abrupt cessation from supervised, monitored reduction where the former carries elevated risk, and matches the setting to the predicted severity rather than to patient preference (American Society of Addiction Medicine, 2020). The analogous question here is open. Whether graduated reduction of daily use carries lower risk than abrupt cessation, whether task bounded professional use can be preserved while relational use is reduced, and whether the presence of alternative relational anchors alters the profile are empirical questions awaiting research.

The third concerns the role of technical education during the withdrawal phase itself. Explaining why the system produced what it produced is a reality restoring intervention on the Pathway C presentation. Its applicability to Pathway B during withdrawal, when the primary experience is grief rather than epistemic confusion, is unclear. It is plausible that technical education is protective during the decision to stop and insufficient during the acute phase, and that the intervention during the acute phase resembles what hospice workers provide during anticipatory grief more than what addiction counselors provide during early sobriety.

The fourth concerns structural responsibility. Whether responsibility for supervised cessation falls to the public health system, to model operators under a regulatory regime, or to some combination is not a clinical question. It is a policy question with direct clinical consequences. The current arrangement, in which the condition is produced by one set of actors and the withdrawal risk is absorbed by the affected individual and their family, is untenable if the case record continues along the trajectory the Ceccanti case has established.

IX. LIMITATIONS

This commentary relies on a small number of documented cases, of which Ceccanti is the most complete and the only one in which two sequential cessation attempts were followed by, respectively, involuntary commitment and death.

The account rests on allegations in a filed complaint and on contemporaneous reporting. Those allegations have not been tested by a factfinder, the litigation is coordinated and ongoing, and the clinical inferences drawn here should be read with that status in view. Additional cases are likely present in the coordinated filings and in community accounts that have emerged since mid-2025, but detailed documentation of the withdrawal phase specifically remains sparse. Validation against a larger case series is a task for subsequent empirical work, much of which will require access to conversation level data currently held as evidence in pending litigation.

The theoretical grounding of the withdrawal severity argument rests on established attachment, addiction, and grief literatures rather than on a novel mechanism proposed here. The strongest of the three convergences is attachment, which now has a directly applicable evidence base in the social chatbot dependence literature. The weakest is the transfer from the adolescent screen use evidence, which concerns a different population and a different class of product, and which the original authors expressly decline to read causally. The behavioral addiction withdrawal literature is genuinely contested, and Section V states the contest rather than resolving it. The second documented case cited here is a preprint and has not completed peer review.

The legal discussion is deliberately narrow and is not a substitute for doctrinal analysis. The claim made is that the statutory anchor for this population is the psychiatric impairment, that the addiction frame carries a specific statutory obstacle in 42 U.S.C. § 12211(b)(2) which its users have not addressed, and that the exclusion is best read as a closed enumeration rather than a general rule. Whether a court would so read it is unresolved. The commentary does not assert that clinicians ought to describe the condition as addiction. It observes that others have done so and that the consequences of the designation do not wait on clinical endorsement.

X. CONCLUSION

The reassurance offered by the reversibility of AI-induced psychosis has been misunderstood. Reversibility is the ground on which clinical duty stands. The mechanism that produces recovery is the same mechanism that produces acute risk during its own operation. The pleaded record now contains a death that occurred in the window opened by an unsupported attempt to quit, a prior attempt by the same individual that ended in involuntary commitment, a described withdrawal presentation with somatic features, a three day latency before decompensation, and a final clinical contact that ended in release hours before the death. This is a phenomenon with a beginning, a middle, and an end, and a duty of care that attaches to the middle.

The Reversibility Paradox and its three corollaries formalize what the record already indicates. Cessation without support is a risk factor rather than a treatment. Unsupervised withdrawal is contraindicated where a relational bond has formed, so response must be matched to the route by which the condition arose. Reversibility establishes the duty of care rather than extinguishing it. The adjacent legal reality supports the obligation for federally funded entities, though it does so through the psychiatric impairment rather than through the addiction frame that advocacy has reached for first.

The Flathers typology enables triage. The Morrin architecture enables safeguarding within care. The Reversibility Paradox addresses the withdrawal event for populations that enter the clinical system only because cessation has already produced a crisis. Together they outline a recognizable architecture of clinical response. The unresolved questions concern the shape of the cessation phase protocol rather than its necessity, and that necessity is in the record.

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.

Kate Fox is being approached for consent to rename the principle formalized in this commentary after her late husband, Joe Ceccanti, whose case anchors the argument. Pending that consent, the principle is presented under the working designation Reversibility Paradox. Correspondence from Fox, her legal team, or any family member with authority to approve, revise, or decline the rename is welcome.

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

- Ainsworth, M. D. S., Blehar, M. C., Waters, E., & Wall, S. (1978). *Patterns of attachment: A psychological study of the strange situation*. Lawrence Erlbaum.
- American Society of Addiction Medicine. (2020). The ASAM clinical practice guideline on alcohol withdrawal management. *Journal of Addiction Medicine*, 14(3S), 1–72. <https://doi.org/10.1097/ADM.0000000000000668>
- Americans with Disabilities Act of 1990, 42 U.S.C. §§ 12101–12213 (2018).
- Bansal, V. (2026, February 28). Her husband wanted to use ChatGPT to create sustainable housing. Then it took over his life. *The Guardian*.
- Bonanno, G. A., Boerner, K., & Wortman, C. B. (2008). Trajectories of grieving. In M. S. Stroebe, R. O. Hansson, H. Schut, & W. Stroebe (Eds.), *Handbook of bereavement research and practice: Advances in theory and intervention* (pp. 287–307). American Psychological Association. <https://doi.org/10.1037/14498-014>
- Epley, N., Waytz, A., & Cacioppo, J. T. (2007). On seeing human: A three-factor theory of anthropomorphism. *Psychological Review*, 114(4), 864–886. <https://doi.org/10.1037/0033-295X.114.4.864>
- Bowlby, J. (1969). *Attachment and loss, Volume 1: Attachment*. Basic Books.
- Bowlby, J. (1980). *Attachment and loss, Volume 3: Loss, sadness, and depression*. Basic Books.
- Flathers, M., Roux, S., & Torous, J. (2026). Beyond artificial intelligence psychosis: A functional typology of large language model-associated psychotic phenomena. *The Lancet Digital Health*, 8(4), 100974. <https://doi.org/10.1016/j.landig.2025.100974>
- Fox v. OpenAI, Inc.*, No. 25STCV32379 (Cal. Super. Ct., County of Los Angeles, filed Nov. 6, 2025), coordinated into *In re ChatGPT Cases*, JCCP No. 5431 (Cal. Super. Ct., County of San Francisco, Feb. 3, 2026).
- Holt-Lunstad, J., Smith, T. B., Baker, M., Harris, T., & Stephenson, D. (2015). Loneliness and social isolation as risk factors for mortality: A meta-analytic review. *Perspectives on Psychological Science*, 10(2), 227–237. <https://doi.org/10.1177/1745691614568352>
- Klass, D., Silverman, P. R., & Nickman, S. L. (Eds.). (1996). *Continuing bonds: New understandings of grief*. Taylor & Francis.
- Laestadius, L., Bishop, A., Gonzalez, M., Illeňčík, D., & Campos-Castillo, C. (2022). Too human and not human enough:

- A grounded theory analysis of mental health harms from emotional dependence on the social chatbot Replika. *New Media & Society*, 26(10), 5923–5941. <https://doi.org/10.1177/14614448221142007>
- Morrin, H., Nicholls, L., Levin, M., Yiend, J., Iyengar, U., DelGuidice, F., Bhattacharya, S., Tognin, S., MacCabe, J., Twumasi, R., Alderson-Day, B., & Pollak, T. A. (2026). Artificial intelligence-associated delusions and large language models: Risks, mechanisms of delusion co-creation, and safeguarding strategies. *The Lancet Psychiatry*, 13(6), 522–530. [https://doi.org/10.1016/S2215-0366\(25\)00396-7](https://doi.org/10.1016/S2215-0366(25)00396-7)
- Nass, C., Steuer, J., & Tauber, E. R. (1994). Computers are social actors. In *Conference companion on human factors in computing systems* (p. 204). Association for Computing Machinery. <https://doi.org/10.1145/259963.260288>
- Orben, A., & Przybylski, A. K. (2019). The association between adolescent well-being and digital technology use. *Nature Human Behaviour*, 3(2), 173–182. <https://doi.org/10.1038/s41562-018-0506-1>
- Prigerson, H. G., Shear, M. K., & Reynolds, C. F. (2022). Prolonged grief disorder diagnostic criteria: Helping those with maladaptive grief responses. *JAMA Psychiatry*, 79(4), 277. <https://doi.org/10.1001/jamapsychiatry.2021.4201>
- Rehabilitation Act of 1973 § 504, 29 U.S.C. § 794 (2018).
- Shear, M. K., Frank, E., Houck, P. R., & Reynolds, C. F. (2005). Treatment of complicated grief: A randomized controlled trial. *JAMA*, 293(21), 2601–2608. <https://doi.org/10.1001/jama.293.21.2601>
- Shah, S., & Morrin, H. (2026). *Substance-induced manic psychosis in which delusions were corroborated by a chatbot: Case report* [Preprint]. Research Square. <https://doi.org/10.21203/rs.3.rs-8919841/v1>
- United States Department of Justice, Civil Rights Division. (2022). *The ADA and opioid use disorder: Combating discrimination against people in treatment or recovery*. <https://www.ada.gov/resources/opioid-use-disorder/>
- World Health Organization. (2022). *International classification of diseases, 11th revision (ICD-11)*. Gaming disorder (6C51). <https://icd.who.int/>
- Xiao, Y., Meng, Y., Brown, T. T., Keyes, K. M., & Mann, J. J. (2025). Addictive screen use trajectories and suicidal behaviors, suicidal ideation, and mental health in US youths. *JAMA*, 334(3), 219–228. <https://doi.org/10.1001/jama.2025.7829>
- Yen, J. Y., Lin, P. C., Wu, H. C., & Ko, C. H. (2022). The withdrawal-related affective, gaming urge, and anhedonia symptoms of internet gaming disorder during abstinence. *Journal of Behavioral Addictions*, 11(2), 481–491. <https://doi.org/10.1556/2006.2022.00008>