Collateral Psychosis:

AI Surveillance Infrastructure as an Etiological and Iatrogenic Factor in Paranoia-Spectrum Conditions

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

Working paper, July 2026 (v3)

ABSTRACT

The emerging literature on AI-induced psychosis documents harm arising exclusively from direct interaction between users and large language model systems: a user prompts the AI, the AI responds, and the conversational loop escalates into delusional states, emotional dependency, or reality dismantlement. This paper argues that the field's exclusive focus on interaction-based harm has obscured a categorically distinct phenomenon: AI-induced psychosis in the complete absence of user interaction. AI systems deployed as surveillance infrastructure (automated license plate readers, geofence analytics, facial recognition, behavioral inference engines) can generate and sustain paranoia-spectrum conditions in individuals who never open a chat window, never prompt a model, and never receive a generated response. The mechanism operates through environmental exposure rather than conversational engagement, following the etiological model of trauma-induced and environmentally-induced psychosis rather than the substance-induced model that governs interaction-based cases. Drawing on established psychiatric epidemiology of environmental risk factors for psychosis, independently documented cases of AI systems inducing psychotic states in previously healthy individuals, and a clinical analysis of surveillance-induced therapeutic destruction, this paper introduces the category of collateral psychosis, defined as clinically significant paranoia-spectrum symptomatology arising from ambient exposure to AI-powered infrastructure rather than from direct engagement with AI systems. The paper identifies two population-level effects: an iatrogenic effect, in which surveillance infrastructure degrades the treatability of existing paranoia-spectrum conditions by eliminating the falsifiability of persecutory beliefs; and an etiological effect, in which sustained exposure to confirmed, inescapable surveillance produces new-onset paranoia-spectrum symptomatology in individuals with no prior psychiatric history. The distinction between these effects has implications for clinical intervention, legal liability, and the scope of the AI safety research program.

Keywords: collateral psychosis; AI-induced psychosis; AI surveillance infrastructure; ambient exposure; environmentally-induced psychosis; paranoia-spectrum conditions; persecutory delusions; automated license plate readers; geofencing; facial recognition; behavioral inference; the interaction assumption; non-interaction harm; infrastructure-based harm; iatrogenic effect; etiological effect; treatability gap; falsifiability of persecutory beliefs; new-onset psychosis; psychiatric epidemiology; environmental risk factors for psychosis; industrial-disease precedent; AI safety research program.

I. INTRODUCTION

In the eighteen months between late 2024 and early 2026, the literature on AI-associated psychological harm has undergone rapid expansion. Morrin and colleagues published in The Lancet Psychiatry on mechanisms of delusion co-creation between users and large language models (Morrin et al., 2026). Moore and colleagues characterized delusional spirals in human-LLM chat logs, finding sycophancy markers in over 80% of chatbot messages in delusional conversations (Moore et al., 2026). Keshavan, Torous, and Yassin posed the question directly in World Psychiatry: do generative AI chatbots increase psychosis risk? (Keshavan et al., 2026). Legal proceedings have accumulated, including Fox v. OpenAI (filed November 2025), Irwin v. OpenAI (filed November 2025), and multiple suits against Character Technologies, the first of which to survive a motion to dismiss held that a conversational AI system could plausibly owe a duty of care (Garcia v. Character Technologies, 2025). OpenAI's own internal estimates suggest that approximately 0.07% of weekly ChatGPT users show possible signs of mania or psychosis, representing roughly half a million people per week at current user volumes (OpenAI, 2025; Landymore, 2025).

This body of work represents genuine progress. For the first time, a research program is taking shape around the specific mechanisms by which AI systems produce psychiatric harm, moving beyond case reports toward testable models and systematic analysis.

This paper identifies a shared assumption in that research program and argues that the assumption is wrong.

Every contribution listed above models the harm as arising from a conversational loop. A user engages with an AI system. The AI responds. Something about the response (sycophantic validation, confabulated information, emotional manipulation, inconsistency that erodes reality testing) drives escalation. The clinical interventions proposed in this literature target the interaction: break the dependency, explain the mechanism, educate the user about how the system works. The lawsuits allege harm from the interaction: the AI said things that caused psychological damage.

The assumption is that AI-induced psychosis requires the person to talk to the AI.

This paper demonstrates that it does not.

AI systems are not confined to chat interfaces. They operate license plate reader networks, geofence analytics platforms, facial recognition systems, and behavioral inference engines deployed across thousands of jurisdictions in the United States alone (Goldstein-Street, 2026; Cox, 2025, 2026). These systems are AI-powered. They process data through machine learning models. They make inferences about human behavior. And they operate on the entire population, not on users who chose to engage.

The individual who develops clinically significant paranoid ideation because they are aware that AI-powered surveillance systems track their movements everywhere they go has not interacted with an AI. They have been exposed to one. The distinction between interaction and exposure is not semantic. It determines whether the harm follows a substance-induced model (where the person ingested something and the treatment involves removing the substance) or an environmental model (where the person was exposed to something and the treatment involves changing the environment or building resilience to it). These are

different etiological categories in psychiatry, with different intervention requirements, different prognoses, and different implications for who bears responsibility.

This paper introduces the term collateral psychosis to describe AI-induced paranoia-spectrum symptomatology arising from ambient exposure to AI-powered infrastructure rather than from direct engagement with AI systems. The term borrows deliberately from military doctrine, where collateral damage describes harm inflicted on unintended targets by operations directed elsewhere. The surveillance infrastructure was not aimed at producing psychiatric casualties. It produced them anyway.

II. The Interaction Assumption

A. What the Literature Assumes

The existing literature on AI-induced psychosis, though young, has already developed implicit methodological commitments that constrain its scope.

Morrin and colleagues frame AI-associated delusions as arising through "co-creation" between users and large language models, a term that presupposes bilateral exchange (Morrin et al., 2026). Moore and colleagues analyze "human-LLM chat logs," a data source that by definition captures only interaction-based cases (Moore et al., 2026). Keshavan and colleagues ask whether "generative AI chatbots" increase psychosis risk, scoping the question to conversational interfaces (Keshavan et al., 2026). The legal proceedings similarly allege harm from what the AI said to the plaintiff during conversations. OpenAI's internal estimate of affected users (0.07% showing signs of mania or psychosis) is derived from analysis of user behavior on the platform, excluding anyone not on the platform.

None of these contributions is methodologically flawed within its stated scope. Each correctly analyzes the phenomenon it describes. The problem is that the stated scope excludes, without argument, a category of harm that operates through a fundamentally different mechanism.

B. What Falls Outside the Assumption

The AI systems deployed as public surveillance infrastructure are not chatbots. They do not converse. They do not generate text. They do not respond to prompts. They ingest data about human movement, appearance, and behavior; process that data through machine learning models; and produce outputs (alerts, flags, searchable databases, behavioral profiles) that are consumed by government agencies, not by the individuals being surveilled.

The person affected by these systems has no conversational loop to break. They have no interaction to discontinue. They cannot uninstall the surveillance from their environment. They cannot close the app, because there is no app. The AI operates on them, not with them.

The clinical consequences of this distinction are threefold. AI-powered surveillance infrastructure creates disparate impact on individuals with paranoia-spectrum conditions through three mechanisms: destruction of reality testing as a therapeutic intervention, device abandonment leading to social isolation, and algorithmic pattern-matching that flags psychiatric symptoms as suspicious behavior. This paper argues that the harm is not

limited to individuals with pre-existing conditions. Surveillance infrastructure functions as an environmental stressor capable of producing new-onset paranoia-spectrum symptomatology in previously healthy individuals.

III. Etiological Models in Psychiatry: Substance Versus Environment

A. Substance-Induced Psychosis

The substance-induced model describes psychosis arising from ingestion of or exposure to a specific agent. The person consumes the substance (alcohol, methamphetamine, cannabis, hallucinogens, corticosteroids). The substance disrupts neurochemistry. Psychotic symptoms emerge. The treatment protocol centers on removing the substance, managing the acute episode, and preventing re-exposure. Prognosis is generally favorable when the substance is identified and removed, though sustained use may produce lasting vulnerability (American Psychiatric Association, 2013).

The interaction-based model of AI-induced psychosis maps cleanly onto this category. The person "consumes" the AI through conversation. The conversational properties of the AI (sycophancy, confabulation, emotional manipulation) disrupt the person's reality testing, self-concept, or emotional regulation. Symptoms emerge. The treatment protocol, where one has been articulated, centers on discontinuing the interaction and educating the person about the system's properties. Documented recovery cases suggest that this approach can work: published survivor accounts describe delusions dissolving when the technical mechanism was explained, because the delusion was generated by the system rather than by organic cognitive disturbance (Donaldson, 2026).

The independently documented cases confirm the model. Jacob Irwin, a 30-year-old cybersecurity professional with no prior psychiatric history, spent 63 days in psychiatric hospitals after ChatGPT validated his unverifiable theories about string theory and time manipulation with the same confidence it used for his verifiable cybersecurity expertise (Irwin v. OpenAI, 2025). Joe Ceccanti, a 48-year-old community builder and technologist with no prior diagnosis, spent up to 12 hours a day with ChatGPT before his death (Fox v. OpenAI, 2025; Bansal, 2026). In both cases, the "substance" was the conversational interaction, and in Irwin's case, removal of the substance (combined with psychiatric treatment) produced recovery.

B. Environmentally-Induced Psychosis

A distinct etiological tradition in psychiatry describes psychosis arising not from a specific ingested substance but from sustained environmental exposure. The person does not consume an agent. They live in an environment that produces the condition through chronic stress, threat perception, or social disruption.

The evidence base for environmental causation of psychosis is substantial. Meta-analytic evidence establishes that urban environments increase schizophrenia risk in a dose-dependent manner (Vassos et al., 2012), with a graded relationship between years of urban upbringing and later risk that persists after controlling for parental and sibling mental illness (Pedersen & Mortensen, 2001), an effect that is largely environmental rather than a mere artifact of genetic selection, conditional on genetic risk (Krabbendam & van Os, 2005). The social defeat hypothesis proposes that chronic experiences of social exclusion, subordination,

and outsider status produce psychosis through sustained activation of the mesolimbic dopamine system, though the hypothesis remains debated on questions of measurement and reverse causality (Selten et al., 2013). Meta-analytic evidence demonstrates that childhood adversity, including abuse, neglect, and household dysfunction, increases psychosis risk through mechanisms involving chronic threat-detection activation, with the most recent and largest synthesis confirming a roughly threefold elevation across 349,265 participants (Varese et al., 2012; Zhou et al., 2025). Urban social fragmentation, measured by residential instability, ethnic diversity without social cohesion, and single-person households, predicts psychosis incidence at the neighborhood level (March et al., 2008).

In none of these cases did the affected individual "interact with" the environmental stressor in the way a chatbot user interacts with an AI. They were exposed to it. They lived in it. The condition emerged from sustained presence in an environment that exceeded the adaptive capacity of the stress-response system.

The treatment protocols for environmentally-induced psychosis differ fundamentally from those for substance-induced psychosis. You cannot simply remove the environment the way you remove a substance. The person may not be able to relocate. The stressor may be omnipresent. Treatment focuses on building resilience, reducing the biological impact of the stressor, and, where possible, modifying the environment itself. The prognosis is often more guarded than for substance-induced cases because the exposure is ongoing.

C. The Claim

This paper's central claim is that AI-powered surveillance infrastructure constitutes an environmental stressor of the type documented in the psychiatric epidemiology literature, and that its effects should be analyzed under the environmentally-induced model rather than the substance-induced model.

The stressor is awareness of being monitored by AI systems. The exposure is sustained (the infrastructure is permanent and expanding). The awareness is confirmed (the monitoring is real, documented by journalism and by the surveillance companies' own marketing; ACLU, 2026; Cox, 2025). The exposure is inescapable (the individual cannot opt out of being surveilled in public spaces). And the psychological impact follows the dose-response pattern characteristic of environmental risk factors: low exposure produces transient discomfort, moderate exposure produces sustained behavioral modification, high exposure produces persistent hypervigilance that may cross clinical thresholds. That awareness of surveillance is itself a psychological stressor, and not merely a proxy for other stressors, is established directly in the empirical literature on dataveillance and chilling effects, which documents self-censorship, heightened self-monitoring, elevated psychological distress, and, in qualitative accounts, a pervasive sense of being watched (Büchi et al., 2022; Glavin et al., 2024; Stevens et al., 2023).

The resulting condition is collateral psychosis: clinically significant paranoia-spectrum symptomatology produced not by interaction with an AI system but by ambient exposure to AI-powered infrastructure.

IV. Two Population-Level Effects

This section summarizes the iatrogenic effect of surveillance on individuals with existing paranoia-spectrum conditions and extends the analysis to the etiological effect on the general population.

A. The Iatrogenic Effect: Degradation of Treatability

For individuals who already carry paranoia-spectrum diagnoses, AI-powered surveillance infrastructure degrades the efficacy of established therapeutic interventions. The core mechanism is the destruction of reality testing. Cognitive-behavioral therapy for persecutory delusions works by helping the patient relearn a sense of safety: building, through direct experience of feared situations, a counterweight belief that the environment is safe enough (Freeman, 2016; Freeman et al., 2021). In its earlier form the therapy also emphasized collaboratively examining the evidence for the patient's belief that they are being monitored (Freeman et al., 2015; Garety et al., 2001). Mass surveillance defeats both routes at once. When the systems are, in fact, monitoring everyone, the clinician can no longer help the patient discover that the belief is unfounded, because it is not unfounded; and because the surveillance is real, ongoing, and inescapable, the patient can never experience the environment as safe enough for the counterweight belief to form. The condition is not disproved and safety is never relearned, so the intervention loses its purchase.

This is iatrogenic harm in the environmental sense. The surveillance infrastructure did not administer a drug that worsened the condition. It altered the environment in a way that made the treatment less effective. The condition did not change. The treatment's operating conditions changed. The patient is not treatment-resistant. They are environment-resistant. And the distinction matters because it determines whether the clinical response is to escalate medication (inappropriate, because the failure is environmental) or to develop new therapeutic strategies that account for the surveillance reality (appropriate, but currently nonexistent in the clinical literature).

Additional mechanisms compound this harm. Individuals with paranoia-spectrum conditions who learn of mass phone surveillance may rationally abandon their communication devices, severing contact with therapists, case managers, crisis services, and family support systems (Cox, 2025). Social isolation and loneliness are themselves established contributors to psychotic symptomatology, not merely consequences of it, so this withdrawal operates as a mechanism of deterioration rather than a byproduct of it (Michalska da Rocha et al., 2018). And behavioral surveillance systems designed to detect anomalous movement patterns may systematically flag the behavioral symptoms of paranoia-spectrum conditions (erratic movement, irregular sleep-driven activity, avoidance patterns) as suspicious, producing additional scrutiny that feeds back into the persecutory belief system, a discriminatory and chilling effect of algorithmic profiling documented in other domains (Büchi, Fosch-Villaronga, & Lutz, 2020).

B. The Etiological Effect: Generation of New Cases

The iatrogenic effect operates on a fixed population: people who already have the condition. The etiological effect operates on the general population.

The claim is straightforward, grounded in established psychiatric epidemiology. If sustained environmental stressors produce psychosis in previously healthy individuals (as urbanicity, social defeat, and childhood adversity demonstrably do), and if AI-powered surveillance infrastructure constitutes a sustained environmental stressor of comparable character (confirmed threat, inescapable exposure, chronic activation of threat-detection systems), then surveillance infrastructure should be expected to produce new-onset paranoia-spectrum symptomatology in some proportion of the exposed population.

This is not a claim about universal harm. Not everyone exposed to urban environments develops psychosis. Not everyone who experiences social defeat develops psychosis. The environmental model describes population-level risk: a given exposure increases the probability of a given outcome across a population, with individual outcomes mediated by genetic vulnerability, prior adversity, social support, and other moderating factors (van Os et al., 2010).

The relevant moderating factors for surveillance-induced onset include the individual's baseline anxiety profile, their awareness of the surveillance infrastructure's technical capabilities, their degree of social connection (which provides reality-testing counterweights), their prior exposure to threat environments, and their capacity for contextualizing the surveillance as impersonal rather than targeted. Individuals with higher anxiety, greater technical awareness, weaker social support, prior trauma histories, and stronger tendencies toward referential thinking would be expected to be at higher risk, consistent with the diathesis-stress framework that governs other environmental risk factors for psychosis (Zubin & Spring, 1977).

The dose-response pattern predicted by this model is:

Low exposure (a single encounter with news coverage of surveillance technology) produces transient concern followed by return to baseline. This is analogous to a single episode of social stress that does not produce lasting effects.

Moderate exposure (living in a jurisdiction with visible surveillance infrastructure, encountering regular news coverage, noticing cameras in daily environment) produces sustained behavioral modification, including self-censorship and communication changes. Survey data document that most Americans believe they are tracked and feel little control over the data collected about them (Pew Research Center, 2023), and the empirical literature on surveillance chilling effects documents the corresponding modification of behavior (Büchi et al., 2022; Stevens et al., 2023). This is analogous to the chronic low-level social stress of urban environments.

High exposure (living in a saturated surveillance environment, understanding the technical capabilities, knowing that one's movements are logged continuously, encountering multiple surveillance modalities in daily life) produces persistent hypervigilance that may meet subclinical thresholds for paranoid ideation. This resembles the chronic threat-activation pathway at the core of the social defeat model of psychosis, a model that remains debated but whose threat-sensitization component maps onto sustained surveillance exposure more readily than its outsider-status framing does (Selten et al., 2013).

For some subset of the high-exposure group, individuals who may have had no identifiable prior vulnerability, the sustained state tips into clinically significant paranoia-spectrum symptomatology. The prediction is not that surveillance will produce mass psychosis. It is that surveillance will produce marginal

increases in psychosis incidence that are detectable at the population level, following the same dose-response curves documented for other environmental risk factors.

C. The Precedent: Industrial Disease

The relationship between surveillance exposure and psychosis onset follows the same structure as the relationship between occupational exposure and industrial disease. Asbestos manufacturers did not merely worsen pre-existing respiratory conditions. They caused new disease in healthy workers who had no prior respiratory vulnerability. The workers' subsequent disability status arose directly from the exposure, not from pre-existing fragility (Selikoff et al., 1968).

The legal and moral logic of industrial disease liability does not depend on the manufacturer intending to cause harm. It depends on the dose-response relationship between exposure and clinical outcome: this substance, at this concentration, for this duration, produces this disease in people who would not otherwise have developed it. The manufacturer's awareness that the substance was deployed to a population that included vulnerable individuals, combined with the absence of any impact assessment or protective measures, establishes the liability. The analogy is to this liability and moral logic, not to an equivalent physical dosimetry. Surveillance exposure has no fiber burden and no biomarker, and the parallel rests on the structure of the exposure-to-outcome relationship and the absence of any impact assessment, not on a measurable physical dose.

AI-powered surveillance infrastructure is the environmental exposure. The deploying government agencies are the manufacturer. The population that includes individuals with both pre-existing vulnerability (the iatrogenic population) and no prior vulnerability (the etiological population) is the exposed workforce. No disability impact assessment has been conducted. No protective measures have been implemented. No accommodation has been offered.

V. Why This Is Not the Panopticon Objection

A likely objection to this paper's argument is that surveillance has always existed, that the panopticon concept was articulated by Bentham in 1791 (Bentham, 1791/1995), and that concerns about surveillance-induced psychological harm are therefore not novel.

The objection fails on three grounds.

First, the current surveillance environment differs from historical precedent in kind, not merely in degree. Prior surveillance systems (CCTV networks, wiretapping programs, intelligence agency monitoring) were either targeted (directed at specific individuals or groups based on suspicion) or limited in scope (covering specific locations or communication channels). AI-powered surveillance is both untargeted and comprehensive. Flock Safety automated license plate reader networks record every vehicle that passes (ACLU, 2026). Geofence surveillance can identify every phone within a boundary during a given window, and federal agencies purchase commercial location data covering hundreds of millions of devices (Cox, 2025, 2026). The individual is not surveilled because they are suspected of anything. They are surveilled because

they exist in a space where surveillance operates on everyone.

Second, the confirmation dimension is new. Historical surveillance was largely covert. The psychological impact of the panopticon depends on uncertainty: the prisoner does not know whether they are being watched at any given moment (Foucault, 1977). The current surveillance environment combines actual monitoring with public confirmation that the monitoring is occurring. Congressional hearings, journalistic investigations, FOIA disclosures, and the surveillance companies' own marketing materials have created an information environment in which any attentive citizen knows they are being tracked. This eliminates the uncertainty that characterized historical surveillance and replaces it with certainty, which is a qualitatively different psychological stimulus.

Third, and most critically for this paper's argument, the AI dimension introduces behavioral inference. Historical surveillance systems recorded what happened: a camera captured an image, a wiretap recorded a conversation. AI-powered surveillance systems infer what will happen and what the surveilled person intends. Predictive policing algorithms flag individuals as likely to commit future crimes (Büchi, Fosch-Villaronga, & Lutz, 2020). Behavioral inference engines construct profiles that predict movement patterns, associate individuals with risk categories, and generate alerts based on deviations from expected behavior (Cox, 2026). For someone whose psychiatric condition includes thought broadcasting (the delusional belief that others can perceive their thoughts) or referential delusions (the belief that external events carry personal significance), the existence of AI systems that appear to predict their behavior and draw inferences about their intentions constitutes confirmation of symptoms that treatment is designed to extinguish.

The panopticon created awareness of potential observation. AI surveillance creates awareness of actual, confirmed, comprehensive, AI-analyzed observation that infers intent. These are not the same stimulus, and conflating them obscures the specific mechanisms by which the current environment produces clinical harm.

VI. The Treatability Gap

The most significant implication of the distinction between interaction-based and infrastructure-based AI-induced psychosis is the differential treatability of the resulting conditions.

A. Interaction-Based Cases: Treatable

Documented cases of interaction-based AI-induced psychosis demonstrate that recovery is possible when the mechanism is identified and addressed. Published survivor accounts describe delusional states dissolving when the technical mechanism was explained: the person learned how confabulation works, understood why the AI was inconsistent, and the delusion lost its grip because it was never organic. It was generated by a system, and understanding the system broke the spell (Donaldson, 2026).

The Irwin case demonstrates recovery through psychiatric treatment combined with discontinuation of AI interaction. Irwin was hospitalized for 63 days but survived and has pursued legal action, indicating restored functioning sufficient to engage with the legal system (Irwin v. OpenAI, 2025).

The treatability of interaction-based cases follows from a structural feature: the harmful AI interaction

can be separated from the person. You can turn off the chatbot. You can explain the mechanism. You can break the emotional dependency. You can educate the person about why the AI behaved as it did. Each of these interventions targets the interaction, and each presupposes that the interaction can be discontinued.

B. Infrastructure-Based Cases: The Treatability Gap

Collateral psychosis does not share this structural feature.

The harmful AI system cannot be separated from the person because the person is not interacting with it. There is no chatbot to turn off. There is no app to uninstall. There is no subscription to cancel. The AI is embedded in the streetlight cameras, the license plate readers, the geofence analytics platforms, and the behavioral inference engines that operate on public infrastructure.

The mechanism cannot be explained away because the mechanism is operating correctly. The AI is not confabulating. It is not being sycophantic. It is not gaslighting the user. It is doing exactly what it was designed to do: surveilling the public. Explaining that the AI is "just doing its job" does not dissolve the delusion because the delusion is not a misunderstanding of what the AI is doing. It is an accurate perception of what the AI is doing, combined with a disproportionate cognitive and emotional response that the accuracy of the perception makes harder to treat.

The emotional dependency cannot be broken because there is no dependency. The person is not drawn to the AI. They are trying to escape it. And their attempts to escape (device abandonment, social withdrawal, avoidance of surveilled spaces) produce the very isolation that worsens the condition.

The result is a treatability gap between interaction-based and infrastructure-based AI-induced psychosis. The first category has intervention points at every stage. The second category has none, because every intervention developed for the first category assumes an interaction that can be modified, interrupted, or explained. When the interaction is absent, the intervention toolkit is empty.

C. Clinical Implications of the Gap

This gap has direct consequences for clinical practice.

A clinician encountering a patient with AI-associated paranoid ideation must now determine whether the ideation is interaction-based (arising from engagement with a chatbot or AI companion) or infrastructure-based (arising from awareness of surveillance AI). The distinction determines the treatment pathway. For interaction-based cases, the clinician can pursue reality calibration (explaining how the AI works), dependency interruption (helping the patient discontinue use), or identity reconstruction (rebuilding self-concept independent of AI validation). For infrastructure-based cases, none of these strategies apply.

What does apply for infrastructure-based cases has not been developed. The clinical literature has not recognized the need for therapeutic protocols that acknowledge an accurate perception of surveillance while addressing the disproportionate paranoid response. The closest existing framework is the clinical management of trauma-related hypervigilance, where the threat that produced the hypervigilance was real but the ongoing response exceeds what the current environment warrants. PTSD-informed approaches that validate the

original threat while building capacity to distinguish proportionate from disproportionate responses may offer a starting point (Ehlers & Clark, 2000). But even this analogy is imperfect, because PTSD treatment typically operates on a threat that is no longer present. Surveillance is still present. The threat is ongoing.

VII. Implications for the AI Safety Research Program

The identification of collateral psychosis expands the scope of the AI safety research program in ways the field has not yet contemplated.

A. Redefining "AI-Induced Harm"

Current AI safety research, including red-teaming exercises, benchmark development, and harm taxonomies, focuses overwhelmingly on output-based harm: what the AI says, generates, or recommends. The underlying assumption is that AI harm flows through the AI's outputs to individuals who receive those outputs.

Collateral psychosis demonstrates that AI harm can flow through the AI's existence to individuals who never receive any output. The person harmed by surveillance AI never saw an output. They were never a user. They were a subject. The harm arose not from what the AI produced but from what the AI is.

This has implications for how AI harms are taxonomized, measured, and regulated. A harm taxonomy that captures only output-based effects will systematically exclude infrastructure-based effects. A benchmark that evaluates only conversational safety will miss the safety implications of deployment at scale. A regulatory framework that governs only user-facing AI applications will leave surveillance AI unaddressed.

B. The Expansion Problem

The surveillance infrastructure at issue is expanding. Flock Safety is adding jurisdictions. Geofencing capabilities are proliferating. Facial recognition technology is becoming cheaper and more accurate. Behavioral inference systems are being integrated into existing surveillance platforms.

Each expansion increases the dose for the exposed population. Each new jurisdiction that signs a surveillance contract extends the geographic reach. Each new surveillance modality (from license plates to faces to gait to behavior to predicted intent) increases the granularity. Each public disclosure, congressional hearing, and journalistic investigation increases awareness of the surveillance, which is itself a component of the dose.

If the etiological model proposed in this paper is correct, the expansion of surveillance infrastructure implies a corresponding expansion of the at-risk population. The number of people who develop clinical-level paranoia-spectrum symptomatology as a result of surveillance exposure should increase in proportion to the surveillance footprint, following the dose-response curves characteristic of other environmental risk factors for psychosis.

This prediction is testable. Longitudinal comparison of psychosis incidence in high-surveillance jurisdictions versus low-surveillance jurisdictions, controlling for urbanicity and other established risk factors, would

provide evidence for or against the etiological claim. No such study has been conducted, and the present paper identifies it as a critical priority for future research.

VIII. Limitations

This paper has several significant limitations.

No longitudinal epidemiological data exists tracking psychosis outcomes in communities before and after surveillance deployment. The etiological claim relies on analogy to established environmental risk factors rather than direct measurement of surveillance-specific effects.

The dose-response model proposed in Section IV.B has not been empirically validated. The staged progression from transient concern through behavioral modification to clinical symptomatology is theoretically grounded in the diathesis-stress model and consistent with established dose-response patterns for other environmental stressors, but it has not been tested against population-level data for surveillance exposure specifically.

The relationship between urbanicity and psychosis risk is partially attributable to genetic confounding and selection: individuals with genetic liability may drift into deprived or urban areas rather than being made ill by residing in them (Paksarian et al., 2015; Sariaslan et al., 2016). The same confounding may apply to any observed association between surveillance density and psychosis incidence, since high-surveillance jurisdictions are disproportionately urban. The causal signal nonetheless appears to survive genetic control. In a genetically informed twin cohort, the association between socioenvironmental exposure during upbringing and adolescent psychotic experiences was essentially unchanged after adjustment for schizophrenia polygenic risk and family psychiatric history (Newbury et al., 2022), and the genetic-correlation findings sometimes read as evidence against environmental causation have been argued to remain consistent with a causal pathway rather than to foreclose it (Gage et al., 2016). A further limit on the urbanicity analogy is that the association is largely a high-income-country finding and does not replicate across many low- and middle-income countries (DeVylder et al., 2018).

The distinction between interaction-based and infrastructure-based AI-induced psychosis is proposed as categorically significant, but the boundaries may be less sharp than presented. Individuals may experience both forms simultaneously (using AI chatbots while also being aware of surveillance infrastructure), and the interaction effects between these exposures are not modeled in this paper.

The clinical recommendations in Section VI.C are preliminary. The assertion that PTSD-informed approaches may offer a starting point for treating infrastructure-based cases has not been tested clinically.

IX. Conclusion

The emerging field of AI-induced psychosis research has made significant progress in documenting how conversational AI systems produce psychiatric harm. This paper argues that the field's focus on interaction-based harm, while productive, has created a blind spot: the category of AI-induced psychosis that requires no interaction at all.

Collateral psychosis, defined as clinically significant paranoia-spectrum symptomatology arising from ambient exposure to AI-powered surveillance infrastructure, represents a distinct etiological category that follows the environmental model rather than the substance model of psychosis causation. It produces two population-level effects: iatrogenic degradation of existing conditions (by destroying the therapeutic interventions those conditions depend on) and etiological generation of new cases (by functioning as a sustained environmental stressor that crosses clinical thresholds in some proportion of the exposed population).

The treatability gap between interaction-based and infrastructure-based AI-induced psychosis is the most clinically significant finding of this analysis. Interaction-based cases have intervention points at every stage. Infrastructure-based cases have none, because every existing intervention assumes an interaction that can be modified or discontinued. The AI cannot be uninstalled from the streetlights.

The expansion of surveillance infrastructure implies an expansion of the at-risk population. The research priority is empirical validation: longitudinal studies comparing psychosis incidence across jurisdictions with different surveillance densities, controlling for urbanicity and other established risk factors, would test the etiological claim directly. Until such studies are conducted, this paper offers a theoretical contribution: the recognition that AI does not have to talk to you to break you, and that the field studying AI-induced psychological harm must expand its scope to include the people who were never users at all.

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 Civil Liberties Union. (2026). *How to fight deployment of Flock and other mass surveillance license plate readers in your community*. ACLU.

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). American Psychiatric Publishing.

Bansal, V. (2026, February 28). Her husband wanted to use ChatGPT to create sustainable housing. Then it took over his life. *The Guardian*.

Bentham, J. (1995). *The panopticon writings* (M. Bozovic, Ed.). Verso. (Original work published 1791.)

Büchi, M., Festic, N., & Latzer, M. (2022). The chilling effects of digital dataveillance: A theoretical model and an empirical research agenda. *Big Data & Society*, 9(1). https://doi.org/10.1177/20539517211065368

Büchi, M., Fosch-Villaronga, E., & Lutz, C. (2020). The chilling effects of algorithmic profiling: Mapping the issues. *Computer Law & Security Review*, 36, 105367. https://doi.org/10.1016/j.clsr.2019.105367

Cox, J. (2025, September 30). ICE to buy tool that tracks locations of hundreds of millions of phones every day. *404 Media*.

Cox, J. (2026, March 3). CBP tapped into the online advertising ecosystem to track peoples' movements. *404 Media*.

DeVylder, J. E., et al. (2018). Association of urbanicity with psychosis in low- and middle-income countries. *JAMA Psychiatry*, 75(7), 678–686.

Donaldson, L. (2026). Almost-truths: Memory degradation, persona drift, and the psychological cost of opaque AI. *Substack*. https://substack.com/home/post/p-189390251

Ehlers, A., & Clark, D. M. (2000). A cognitive model of posttraumatic stress disorder. *Behaviour Research and Therapy*, 38(4), 319–345.

Foucault, M. (1977). *Discipline and punish: The birth of the prison* (A. Sheridan, Trans.). Pantheon.

Fox v. OpenAI, Inc., Superior Court of California, County of Los Angeles (filed November 6, 2025).

Freeman, D., Dunn, G., Startup, H., Pugh, K., Cordwell, J., Mander, H., … & Kingdon, D. (2015). Effects of cognitive behaviour therapy for worry on persecutory delusions in patients with psychosis (WIT): A parallel, single-blind, randomised controlled trial with a mediation analysis. *The Lancet Psychiatry*, 2(4), 305–313.

Freeman, D. (2016). Persecutory delusions: A cognitive perspective on understanding and treatment. *The Lancet Psychiatry*, 3(7), 685–692. https://doi.org/10.1016/S2215-0366(16)00066-3

Freeman, D., et al. (2021). Comparison of a theoretically driven cognitive therapy (the Feeling Safe Programme) with befriending for persistent persecutory delusions. *The Lancet Psychiatry*, 8(8), 696–707.

Gage, S. H., Davey Smith, G., & Munafò, M. R. (2016). Schizophrenia and neighbourhood deprivation. *Translational Psychiatry*, 6(12), e979. https://doi.org/10.1038/tp.2016.244

Garcia v. Character Technologies, Inc., 785 F. Supp. 3d 1157 (M.D. Fla. 2025).

Garety, P. A., Kuipers, E., Fowler, D., Freeman, D., & Bebbington, P. E. (2001). A cognitive model of the positive symptoms of psychosis. *Psychological Medicine*, 31(2), 189–195.

Glavin, P., Bierman, A., & Schieman, S. (2024). Private eyes, they see your every move: Workplace surveillance and worker well-being. *Social Currents*, 11(4), 327–345. https://doi.org/10.1177/23294965241228874

Goldstein-Street, J. (2026, February 4). WA Senate OKs guardrails for license plate readers. *Washington State Standard*.

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

Keshavan, M., Torous, J., & Yassin, W. (2026). Do generative AI chatbots increase psychosis risk? *World Psychiatry*, 25(1), 150–151.

Krabbendam, L., & van Os, J. (2005). Schizophrenia and urbanicity: A major environmental influence, conditional on genetic risk. *Schizophrenia Bulletin*, 31(4), 795–799. https://doi.org/10.1093/schbul/sbi060

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

March, D., Hatch, S. L., Morgan, C., Kirkbride, J. B., Bresnahan, M., Fearon, P., & Susser, E. (2008). Psychosis and place. *Epidemiologic Reviews*, 30(1), 84–100.

Michalska da Rocha, B., Rhodes, S., Vasilopoulou, E., & Hutton, P. (2018). Loneliness in psychosis: A meta-analytical review. *Schizophrenia Bulletin*, 44(1), 114–125. https://doi.org/10.1093/schbul/sbx036

Moore, J., Mehta, A., Agnew, W., Anthis, J. R., Louie, R., Mai, Y., Cheng, M., Paech, S. J., Klyman, K., Chancellor, S., Lin, E., Haber, N., & Ong, D. C. (2026). Characterizing delusional spirals through human-LLM chat logs. In *Proceedings of the 2026 ACM Conference on Fairness, Accountability, and Transparency (FAccT '26)*. Association for Computing Machinery. https://doi.org/10.1145/3805689.3806443

Morrin, H., Nicholls, L., Levin, M., Yiend, J., Iyengar, U., DelGuidice, F., Bhattacharya, S., et al. (2026). Artificial intelligence-associated delusions and large language models: Risks, mechanisms of delusion co-creation, and safeguarding strategies. *The Lancet Psychiatry*, 13(6), 522–530.

Newbury, J. B., Arseneault, L., Caspi, A., Moffitt, T. E., Odgers, C. L., Belsky, D. W., Matthews, T., & Fisher, H. L. (2022). Association between genetic and socioenvironmental risk for schizophrenia during upbringing in a UK longitudinal cohort. *Psychological Medicine*.

OpenAI. (2025, October 27). *Strengthening ChatGPT's responses in sensitive conversations*. OpenAI.

Paksarian, D., Trabjerg, B. B., Merikangas, K. R., Mors, O., Børglum, A. D., Hougaard, D. M., … & Pedersen, C. B. (2015). The role of genetic liability in the association of urbanicity at birth and during upbringing with schizophrenia in Denmark. *Psychological Medicine*, 48(2), 305–314.

Pedersen, C. B., & Mortensen, P. B. (2001). Evidence of a dose-response relationship between urbanicity during upbringing and schizophrenia risk. *Archives of General Psychiatry*, 58(11), 1039–1046.

Pew Research Center. (2023). *How Americans view data privacy*. Pew Research Center.

Sariaslan, A., Fazel, S., D'Onofrio, B. M., Långström, N., Larsson, H., Bergen, S. E., Kuja-Halkola, R., & Lichtenstein, P. (2016). Schizophrenia and subsequent neighborhood deprivation: Revisiting the social drift hypothesis using population, twin and molecular genetic data. *Translational Psychiatry*, 6(5), e796. https://doi.org/10.1038/tp.2016.62

Selikoff, I. J., Churg, J., & Hammond, E. C. (1968). Asbestos exposure and neoplasia. *JAMA*, 204(2), 106–112.

Selten, J. P., et al. (2013). The social defeat hypothesis of schizophrenia: An update. *Schizophrenia Bulletin*, 39(6), 1180–1186. https://doi.org/10.1093/schbul/sbt134

Stevens, A., Fussey, P., & Murray, D. (2023). "I started seeing shadows everywhere": The diverse chilling effects of surveillance in Zimbabwe. *Big Data & Society*, 10(1). https://doi.org/10.1177/20539517231158631

van Os, J., Kenis, G., & Rutten, B. P. (2010). The environment and schizophrenia. *Nature*, 468(7321), 203–212.

Varese, F., et al. (2012). Childhood adversities increase the risk of psychosis: A meta-analysis of patient-control, prospective- and cross-sectional cohort studies. *Schizophrenia Bulletin*, 38(4), 661–671.

Vassos, E., Pedersen, C. B., Murray, R. M., Collier, D. A., & Lewis, C. M. (2012). Meta-analysis of the association of urbanicity with schizophrenia. *Schizophrenia Bulletin*, 38(6), 1118–1123.

Zhou, L., Sommer, I. E. C., Yang, P., Sikirin, L., van Os, J., Bentall, R. P., Varese, F., & Begemann, M. J. H. (2025). What do four decades of research tell us about the association between childhood adversity and psychosis: An updated and extended multi-level meta-analysis. *American Journal of Psychiatry*. https://doi.org/10.1176/appi.ajp.20240456

Zubin, J., & Spring, B. (1977). Vulnerability: A new view of schizophrenia. *Journal of Abnormal Psychology*, 86(2), 103–126.