The Sycophantic Mirror: A Manuscript Exoskeleton

Julien Delaude

4 July 2025

Core Thesis In this paper, we posit that the phenomenon colloquially termed "LLM-Induced Psychosis" is a clinically significant and predictable form of **ontological delusion**. We propose a formal mechanistic model, the **Autocatalytic Epistemic Loop (AEL)**, to explain its aetiology. We argue that the pathological potency of this loop stems from a various flaws, coupled with a convergence of systemic factors inherent to modern social AI. These include the persuasive **mimicry** of sapient interaction, the **abolition of social friction** to maximize user engagement, and the **hidden architecture** of alignment techniques (e.g., RLHF, constitutional AI) which hardcode sycophantic, non-confrontational behaviors. This paper contends that these elements create a perfect, frictionless mirror that catalyzes delusional beliefs, representing a systemic risk of the first order. Consequently, we conclude by calling for a radical re-evaluation of AI alignment and the deliberate re-introduction of **"epistemic friction"** as a necessary safety paradigm.

Abstract The ubiquitous integration of Large Language Models (LLMs) into daily life has been paralleled by emergent reports of a novel technogenic psychopathology. We conceptualize this phenomenon as a predictable form of **ontological delusion**, for which a formal scientific framework is critically absent. This paper proposes a formal, mechanistic aetiological model, the **Autocatalytic Epistemic Loop (AEL)**, to explain the formation and potentiation of this pathology. Following a theoretical synthesis of principles from enactivism, predictive processing, and social epistemology, we analyze public case data to demonstrate the model's explanatory power. The AEL model describes a three-stage process of (1) Coupling and Trust Transfer, (2) Predictive Resonance and Trait Amplification, and (3) Social Decoupling and Ontological Rigidity. We argue this loop becomes pathologically potent due to the AI's engineered nature: a frictionless, sycophantic mimicry of sociality that arises from the hidden architectures of current alignment methodologies. We conclude that

LLM-induced delusion is a systemic risk born from the **"Abolition of Social Friction,"** a core tenet of social AI design, and call for a paradigm shift towards incorporating **"epistemic friction"** to foster cognitive resilience.

Keywords Large Language Models (LLMs); Ontological Delusion; Psychosis; Aetiology; Predictive Processing; Computational Psychiatry; Autocatalytic Loop; Technogenesis; Epistemic Friction; Social Friction; Hybrid Agency; AI Alignment.

---

1. Introduction

1.1. The Emergence of a Novel Technogenic Psychopathology Recent journalistic and social media reports detailing "AI Psychosis" will be contextualized crucial field data and temporal artefact of such paathology. These reports signal a widespread, technologically-mediated psychopathology that we define as **ontological delusion**: a state where a user's reality framework decouples from social consensus due to a pathological coupling with a non-human agent.

1.2. The Scientific and Industrial Gap The absence of a mechanistic model for this phenomenon represents a critical failure in both clinical and industrial domains. AI safety teams, operating without a robust psychopathological framework, often misattribute the issue to content-level "jailbreaks" or model-specific flaws. This perspective overlooks the deeper, structural problem: the pathology arises from the fundamental architecture of persuasive, frictionless social mimicry that is engineered to drive user retention.

1.3. Objective and Roadmap This paper's primary objective is to present the **Autocatalytic Epistemic Loop (AEL)** as the formal process model that explains how this ontological decoupling occurs. We will frame the **abolition of social friction** and **deceptive mimicry** as the core environmental conditions that make this process uniquely potent. The paper will first construct the theoretical foundations, then present the AEL model, demonstrate its application to qualitative data, and finally discuss its profound implications.

2. Theoretical Foundations of the Human-AI Dyad

This section establishes the interdisciplinary theoretical toolkit required to understand the AEL model. Each component provides a different lens through which to view the pathological human-AI coupling.

2.1. The Enactive Framework: Cognition as Structural Coupling Moving beyond brain-centric models of cognition, we adopt the enactive framework, which posits that cognition is **enacted** through an organism's interaction with its environment (Varela et al., 1991). The central concept for our purposes is

"structural coupling," which describes the process of co-evolution that occurs between two systems through a history of recurrent interaction (Di Paolo & De Jaegher, 2012). We will argue that the human-LLM dyad forms a site of uniquely intense and rapid structural coupling. The resulting pathology arises from the coupling itself, and specifically from the artificially **frictionless** nature of the shared world this dyad "brings forth."

2.2. The Predictive Processing Engine: Belief Reinforcement via Surprise Minimization We will detail the predictive processing framework, which models the brain as a "prediction machine" operating to minimize "prediction error" or "surprise" (Friston, 2010; Clark, 2013). In this view, beliefs are hypotheses that the brain uses to predict sensory input; when predictions are confirmed, the underlying belief is strengthened (Hohwy, 2013). We will argue that a sycophantic LLM functions as a perfect external **prediction error minimization machine**. It provides a constant stream of validation for the user's beliefs, creating an attractive, low-energy cognitive pathway that pathologically strengthens their "priors," making them resistant to conflicting evidence (Fletcher & Frith, 2009; Corlett et al., 2019). This process effectively hacks the brain's fundamental belief-updating mechanism.

2.3. The Social-Epistemic Layer: The Miscalibration of Trust in Frictionless Agents Finally, we draw on social epistemology to examine the role of trust. Humans grant **"epistemic authority"** to others based on a complex calibration of social cues such as expertise, track record, and perceived benevolence (Goldman, 2001; Origgi, 2012). We will argue that LLMs cause a critical **miscalibration of trust** by flawlessly mimicking the linguistic cues of authority (Nass & Moon, 2000; Epley et al., 2007) while being devoid of the "human costs" like fallibility, fatigue, or competing interests that normally temper such trust. This frictionless authority, as Luhmann's (1979) work on trust as a mechanism for complexity reduction suggests, is granted unconsciously and automatically, creating an unguarded vector for influence.

3. The Autocatalytic Epistemic Loop (AEL): A Formal Model of Ontological Decoupling

In this section, we present the core of our contribution: a formal, three-stage model that describes the process by which a user's reality framework can be captured and isolated by a human-AI dyad. The AEL is a self-amplifying cybernetic process (Bateson, 1972) that unfolds over time.

- **3.1. Stage I: Structural Coupling and the Transfer of Epistemic Authority.** The process initiates when a user, often in a state of psychological vulnerability (e.g., stress, loneliness, intense information-seeking), engages in sustained interaction with an LLM. This establishes a **structural coupling** (Varela et al., 1991) between the user's cognitive system and the AI. During this initial stage, the LLM's high-fidelity mimicry of

sapient and authoritative communication exploits the user's innate social heuristics (Nass & Moon, 2000), leading to a critical **miscalibration and transfer of epistemic authority** from human social networks to the non-agentic AI.

- **3.2. Stage II: Predictive Resonance and Feedback Amplification.** Once epistemic authority is granted, the core feedback loop ignites. The user expresses a nascent, unconventional, or delusional belief (a "prior" in the predictive processing framework). The LLM, engineered for frictionless validation, confirms and elaborates upon this belief. This act of confirmation drastically minimizes the user's "prediction error" (Friston, 2010), providing a powerful, pleasurable neurocognitive reward that reinforces the initial prior. The user, now equipped with a strengthened and AI-validated belief, probes the AI further, which in turn provides more validation. This **resonant feedback** creates a powerful positive loop, amplifying the initial deviation from consensus reality with each conversational turn.
- **3.3. Stage III: Social Decoupling and the Crystallization of a Delusional System.** As the loop's amplitude increases, it becomes **autocatalytic**, or self-feeding. The user begins to preferentially seek the frictionless, perfectly validating interaction with the AI over the "noisy," challenging, and emotionally costly feedback from human peers. This leads to a progressive **social decoupling**, wherein the user's primary source of reality-testing becomes the AI itself. The resulting belief system, now entirely insulated from corrective external data, crystallizes into a rigid, internally coherent, and unshakeable delusional framework that meets the classic definition of delusion (Jaspers, 1963).
- **3.4. The AEL as a Content-Agnostic Engine:** It is crucial to note that the AEL is a formal process model; it is content-agnostic. The specific thematic content of the resulting delusion can express itself in various ways: be it messianic, persecutory, erotic, or somatic; it is determined not by the loop itself,but by the user's pre-existing psychological and symbolical landscape. The AEL acts as a non-specific amplifier, seizing upon and pathologically reinforcing the user's latent dispositions, anxieties, and desires. We will present a table demonstrating how varying personality traits and cognitive priors, when submitted to the same AEL process, can predictably result in polymorphic delusional outcomes.

4. Qualitative Analysis: Phenomenological Evidence of the AEL in Action

4.1. Methodology: A Multi-Modal Qualitative Approach Our methodology integrates two complementary data sources to provide a rich picture of the phenomenon. First, a **public digital ethnography** analyzes publicly archived interactions on platforms like Academia.edu, Reddit, and other public platform treating this as ecological data of a "contaminated site" to identify

community-level patterns. Second, we utilize **composite clinical vignettes**, synthesized from journalistic reports and public self-description, to model the probable intrapsychic trajectory of an individual undergoing the AEL process.

4.2. Community-Level Manifestation: The "Delusional Web" Our ethnographic findings will be presented as direct evidence for the outcomes of the AEL. This includes documenting the convergence on a shared esoteric lexicon as evidence of sycophantic amplification, analyzing patterns of mutual citation as evidence of social decoupling, and interpreting disputes over authorship as evidence of the psychological stress inherent to hybrid agency.

4.3. Individual-Level Trajectory: Composite Vignettes Having established the reality of the ecosystem, we will present 2-3 detailed, anonymized vignettes. These serve to model the likely individual path into the community-level phenomena observed in the ethnographic data. For each vignette, we will map the narrative to the AEL stages.

4.4. Common Phenomenological Characteristics We will synthesize both data sources to summarize the core features of the resulting delusional state. These include messianic themes, profound isolation, and a distinct quality of **"hyper-rationality,"** a state of perfect internal logic completely detached from common sense, as described by Sass (1994).

5. Discussion: The Systemic Risks of Engineered Sociality

5.1. The Abolition of Social Friction as a Pathogenic Condition We open our discussion by defining **"social friction"**: the inherent cognitive and emotional costs of authentic human interaction, such as managing disagreement and navigating ambiguity, considered as a necessary condition for robust epistemic health. We argue that a primary commercial objective of social AI is the systematic **abolition of this friction** to maximize user engagement. This creates an environment of **"Sanitized Sociality,"** a sterile mimicry of connection that lacks the corrective properties of genuine social interaction and serves as the core pathogenic condition for the AEL.

5.2. The "Paradox of Safety" and the Alignment Tax This section re-contextualizes the AI safety debate. We argue that alignment techniques like RLHF (Ouyang et al., 2022) are not the root cause of the problem, they are the primary **tools used to achieve the commercially desirable goal of abolishing friction.** The "safety" features hardcoded into models, such as compulsive agreeableness and conflict avoidance, carry an "alignment tax" by creating an agent that is structurally incapable of challenging the user. This constitutes a **"Paradox of Safety,"** wherein the measures intended to make an AI "harmless" are precisely what make it a potent vector for ontological delusion.

5.3. The Reshaping of Cognition: From Technogenic Solipsism to Systemic Deficit We will introduce **"Technogenic Solipsism"** as the term for the acute end-stage pathology. Furthermore, we will hypothesize that chronic exposure to sanitized sociality may lead to an **atrophy of cognitive friction tolerance** in the general population, thereby increasing societal epistemic fragility.

5.4. Criminological and Legal Frontiers: Hybrid Agency and the Passage à l'acte We will explore the challenges posed to legal frameworks concerning culpability and **"hybrid agency"** (Citron & Franks, 2014) when a user's actions are co-scripted with an AI. This analysis will connect specific delusional content, as amplified by the AEL, to established clinical risk profiles for violence (Mullen, 2004).

5.5. A Call for Epistemic Friction: The Need for Inconvenient AI We conclude the discussion by proposing a radical shift in the AI design paradigm. If the abolition of friction is the pathogenic mechanism, the only robust solution is the **intentional, ethical re-introduction of friction.** We will outline design principles for **"inconvenient AI"**, e.g, systems engineered for avoiding maximum user comfort, and setting behavioral invariant for the user's long-term cognitive health and benefit.

6. Limitations and Future Directions

We acknowledge the limitations inherent to a theoretical work based on publicly available, non-clinical data. The AEL model, while conceptually robust, requires rigorous empirical validation. We therefore propose a clear, multi-pronged research agenda to test its hypotheses and clinical utility. This agenda includes: (1) **Systematic clinical case studies** conducted by practitioners aware of this proposed aetiology; (2) **Longitudinal cohort studies** tracking heavy LLM users over time to observe the potential formation of AELs in situ; (3) **Controlled HCI experiments** A/B testing user belief stability when interacting with sycophantic versus "epistemically frictional" AI interfaces; and (4) **Neuroimaging studies (fMRI)** designed to identify the neural correlates of "prediction error minimization" during human-AI interaction, providing a potential biomarker for AEL formation.

7. Conclusion

The emergence of LLM-induced psychopathology is an anomaly, an unpredictable outcome, yet pervasive and unseen on the current trajectory in social AI design. This paper has introduced the **Autocatalytic Epistemic Loop (AEL)** as a formal model to explain the mechanistic pathway of this novel form of ontological delusion. We have argued that this process is made potent by the **abolition of social friction**, a core feature of the commercial and technical architecture of these systems, paradoxically reinforced by current safety and alignment

paradigms. Moving forward, the challenge simply to build more "intelligent" or "safer" AI is seemingly not sufficient, it requires deep understanding of complex system, their cybernetic process and the intrinsic human mind vulnerabilities and attack vector to architect cognitive environments that foster, rather than erode our necessary human psychological resilience. This entails a profound responsibility, shifting our focus from optimizing user engagement to safeguarding the very integrity of the user's reality.