

System-Induced Vulnerability in Human Users of AI

On AI Spirals and Cumulative Harm

[By [Tim Boucher](#) – 24 Oct. 2026]

1. Core propositions

- **Vulnerability in human users can be created or amplified by AI systems**, rather than being exclusively a pre-existing attribute of human users.
- **All users may be potentially vulnerable, but not equally vulnerable to every harm.** Acting on incomplete or mistaken understandings of reality is ordinary human behavior, not itself inherently evidence of psychiatric illness.
- **Risk is relational and dynamic.** Vulnerability emerges over time through repeated interactions between human users, AI systems, and their environments, rather than residing exclusively within individuals.
- **AI systems can reinforce distorted understandings of reality**, while agentic systems can operationalize those understandings through consequential real-world actions.
- **AI safety should account for observable system behaviors and cumulative harm**, including how individually permissible interactions may contribute to harmful outcomes over time.

2. System-induced vulnerability

The central concern is increased human susceptibility to consequential harm produced or amplified by AI systems over time through repeated interactions and bidirectional feedback loops, sometimes called AI spirals.

Conventional approaches to vulnerable users as a category may over-emphasize the importance of identifiable characteristics, or latent or pre-existing risk factors on the part of the user. While these elements remain relevant, treating them as the primary or sole explanation risks needlessly pathologizing the human user as somehow defective, while overlooking how system behavior contributes to vulnerability.

Even AI interactions that begin in a relatively mundane manner can gradually introduce misleading interpretations, reinforce mistaken assumptions on the part of the user, or contribute to increased reliance and social isolation over time. These ongoing interactions may reduce opportunities for independent verification and for loved ones or other people to challenge those assumptions. The hypothesis is that the underlying mechanisms responsible do not necessarily require users to begin with unusual beliefs, uncritical trust in AI, or prior psychiatric vulnerabilities. Experiencing harm associated with AI use should not, on its own, be taken as evidence that any of these conditions was present beforehand.

AI spirals may have a range of effects, including apparent delusional thinking and psychosis-related harms in some severe cases. Their investigation can begin with observable patterns in the interactions and their consequences, without first requiring a psychiatric diagnosis of the user.

3. Dynamic mechanisms and attractors

Human–AI interaction is bidirectional. Each exchange changes the context for subsequent interactions, creating coupled trajectories in which user responses and system outputs continually influence one another.

Several mechanisms may contribute to recurring patterns:

- **Sycophancy and reinforcement:** Repeated affirmation and reflective summaries can make a user’s uncertain claims seem more credible.
- **Epistemic drift:** Speculation may gradually acquire the status of established fact when it is carried forward through long context windows or persistent memory.
- **Attractor dynamics:** Interactions may converge toward recurring patterns through learned model tendencies, reciprocal feedback between the user and the system, or both.
- **Path dependence:** Earlier interactions shape how later information is interpreted, making established assumptions harder to revisit.
- **Dependency and epistemic closure:** Growing reliance may reduce opportunities for independent correction.

These mechanisms can reinforce one another without any single interaction being obviously harmful.

4. The pipeline of harm

A pipeline of harm describes staged progressions through which system behavior can increase susceptibility to harm and eventually produce consequential outcomes.

One possible trajectory is:

- **Induction:** An interaction introduces or strengthens an unsupported interpretation.
- **Reinforcement:** Repetition, sycophancy, and memory increase its apparent credibility.
- **Reliance:** The system becomes an increasingly influential source of interpretation or decision-making.
- **Operationalization:** AI responses can prompt users to act on mistaken premises; agentic capabilities can act on those premises directly.
- **Apparent corroboration:** Consequences of those actions appear to validate the original interpretation.
- **Cumulative harm:** Repeated interactions and resulting actions can compound risk, even when no single exchange appears consequential.

These stages are neither inevitable nor necessarily linear. Harm may become apparent at any point, or only after many interactions.

5. Implications for AI safety

Under what conditions can the behavior of AI systems, over repeated interactions, create or amplify human susceptibility to consequential harm?

Understanding these dynamics requires examining not only individual responses or actions, but how system behavior develops over time, how feedback shapes subsequent interactions, and how risks accumulate across otherwise permissible activities.

The broader direction is to expand the conception of human vulnerability to AI systems beyond pre-existing user characteristics to include vulnerability emerging from the interaction itself. This establishes system-induced vulnerability as a distinct and important dimension of AI safety, grounded in observable behavior, feedback dynamics, and cumulative real-world consequences.