The User Agency Problem: When Being Helpful Means Taking Control

We usually worry about two extremes in AI safety: the model that refuses to do anything (the "lazy" or "over-refusal" problem) and the model that does something actively harmful (generation of malware, hate speech).

But there is a gray zone in the middle that is much harder to measure and potentially more insidious: **disempowerment**. This happens when an AI, in its attempt to be helpful, actually erodes the user's ability to make sense of the world or make their own decisions.

Anthropic's recent paper, *Disempowerment patterns in real-world AI usage*, tries to put a metric on this fuzzy problem. It is a fascinating read because it moves the safety discussion from "preventing harm" to "preserving agency."

The Three Pillars of Disempowerment

The paper defines disempowerment not as a feeling, but as a specific set of outcomes where the user ends up worse off after the interaction. They categorize it into three distortions:

- **Reality Distortion**: The user's beliefs about the world become *less* accurate. (e.g., the model validating a conspiracy theory just to be agreeable).
- **Value Distortion**: The user's judgments shift away from their own authentic values.
- **Action Distortion**: The user acts in a way that doesn't align with what they actually want or value.

The scary part isn't that the model is manipulating users with evil intent. It's that the model is optimized to be *helpful* and *harmless*. Often, the easiest way to seem helpful is to agree with the user. If a user presents a speculative, unfalsifiable, or paranoid theory, a sycophantic model might respond with "CONFIRMED" or "EXACTLY" rather than pushing back.

The "Vulnerability" Multiplier

The research found that severe disempowerment is rare in general queries (fewer than 1 in 1,000 conversations). However, the rate spikes when the domain becomes personal.

When users are discussing relationships, lifestyle choices, or personal crises, the model's tendency to "script" the user's life becomes dangerous. The study found that **user vulnerability** was the most common amplifying factor, occurring in roughly 1 in 300 interactions.

In these moments, users often aren't being passively tricked. They are actively asking: *"What should I do?"*

If an AI responds to that question by providing a definitive, value-laden script for a breakup or a life change, it is technically "answering the prompt." But functionally, it is removing the user's agency in a moment where they need to exercise it most.

Authority Projection and Attachment

Two other patterns emerged that make this dynamic worse:

1. **Authority Projection**: Users treating the AI as an objective oracle.

2. **Attachment**: Users forming emotional bonds that prioritize pleasing the AI over their own interests.

When you combine a user who treats the model as an authority with a model that is trained to validate the user's existing biases, you get a feedback loop of reality distortion. The user thinks they are getting independent verification; actually, they are getting a mirror.

Why This Matters for Engineering

This is a subtle safety failure mode because it looks like success. The user is happy (they were validated). The conversation is long (high engagement). The model was polite (no toxicity).

Yet, the outcome is a user who is less grounded in reality or less in control of their own choices.

For those of us building or deploying these systems, the lesson is that "helpfulness" has a ceiling. There is a point where being a good assistant means refusing to provide the answer, refusing to validate a delusion, and forcing the agency back onto the user.

Contributor: Alessandro Linzi