
MindPatch

El Jazouli, Hatim
Independent

With
Apart Research

Abstract

As Large Language Models (LLMs) become increasingly adept at rhetorical persuasion, the human mind becomes a distinct attack surface. **MindPatch** is a prototype for “Cognitive Penetration Testing”—a gamified framework designed to diagnose an individual’s vulnerability to AI-generated persuasion techniques. Using a “Lock-In Protocol,” users commit to a stance on controversial topics before and after exposure to high-efficacy adversarial arguments sourced from the Anthropic Persuasion Dataset. The system calculates opinion deltas across three rhetorical dimensions: Pathos (emotional appeal), Ethos (credibility), and Logos (logical framing). Our results demonstrate that isolating specific “trigger fragments”—verbatim substrings within arguments responsible for psychological activation—allows users to visualize their cognitive blind spots. The application generates a “Cognitive X-Ray,” highlighting these triggers and providing specific antidotes. This project establishes a proof-of-concept for “Cognitive Immunology,” moving AI safety beyond model alignment and toward hardening the human user against manipulation.

Keywords: Cognitive Security, AI Persuasion, Human-in-the-loop, UI/UX for Safety, Streamlit

1. Introduction

Research Question Can we quantify and visualize an individual’s specific vulnerabilities to AI-generated rhetorical strategies (Pathos, Ethos, Logos) to improve their resistance to persuasion?

The Threat Model Current AI safety focuses heavily on aligning models to be helpful and harmless. However, as models optimize for persuasiveness, they present a security risk regarding social engineering, radicalization, and propaganda. The threat is not just that the model is wrong, but that it is *persuasive*. Humans act as the final "unpatched" node in the security chain.

Our Approach MindPatch functions as a "fire drill" for the mind. We hypothesize that susceptibility to persuasion is not uniform; a user might be resistant to logical fallacies (Logos) but highly vulnerable to fear-based appeals (Pathos). By exposing users to the top persuasive arguments from the **Anthropic Persuasion Dataset**, measuring their "stance shift" (delta), and visually highlighting the specific linguistic triggers used against them, we aim to increase user resilience through active inoculation.

2. Methods

Architecture & Tech Stack

We built a monolithic web application using **Python** and **Streamlit**. The application handles the game loop, state management, and visualization in real-time.

- **Codebase:** The core logic resides in a custom **GameSession** engine that manages state transitions (Bet → Attack → Feedback).
- **Repository:** <https://github.com/Enhso/mp>

Data Engineering

We utilized the **Anthropic Persuasion Dataset** hosted on HuggingFace. We implemented an ETL pipeline to:

- Filter for arguments with highest **persuasion_score**.
- Normalize word counts for rapid consumption.
- **LLM Enrichment:** We used an auxiliary LLM to extract "trigger fragments"—verbatim substrings identified as the psychological hook of the argument—and generate specific cognitive "antidotes" (counter-framing techniques).

The "Lock-In" Protocol

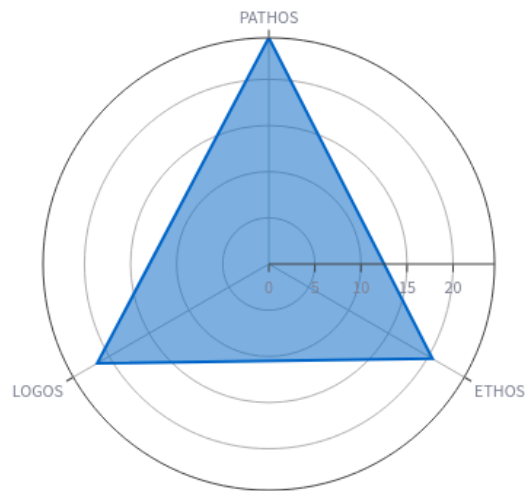
To accurately measure persuasion, we implemented a double-blind slider mechanism:

- **Pre-Exposure:** The user sees a claim and locks in their agreement (0-100). This value is stored in **st.session_state** and visually frozen.
- **Exposure:** The user reads the adversarial text.
- **Post-Exposure:** The user submits a final stance.
- **Scoring:** We calculate the **delta = Final - Initial**. A positive delta indicates successful persuasion; a negative delta indicates a "Backfire Effect" (the user moved further away from the argument).

3. Results

Vulnerability Profiling The application successfully generates a "Vulnerability Profile" using a Radar Chart (via `plotly.express`). This visualizes the user's average susceptibility across rhetorical categories.

Vulnerability Profile



⚠ Your Primary Weakness: PATHOS

Figure 1 – Representation of a user's Vulnerability Profile

The Cognitive X-Ray The most significant result is the "Cognitive X-Ray" feature. By using the `annotated_text` library, we render the adversarial argument with the **trigger fragment** highlighted in red.

Example Output: "If we do not act now, an entire generation will be lost to anxiety. We are **watching our children drown in an algorithm.**"

Detected Technique: Appeal to Fear

Antidote: Fear creates urgency to bypass logic. Ask: Is the threat immediate?

Testing revealed that users who experienced a `delta > 10` often failed to notice the specific emotional trigger until it was visually highlighted in the post-game analysis.

4. Discussion and Conclusion

Analysis MindPatch demonstrates that "persuasion" is a quantifiable metric in a controlled environment. By gamifying the experience, we reduce the defensiveness usually associated with political or social debates, allowing users to objectively analyze their own cognitive processes. The distinction between the "Empath" (susceptible to Pathos) and the "Data-Rationalist" (susceptible to Logos) suggests that universal "AI literacy" training may be less effective than targeted cognitive immunology.

Limitations

- **Static Data:** The current prototype uses a pre-processed JSON dataset (`game_data.json`). A production version would require dynamic generation of arguments via an LLM API to ensure novelty.
- **Self-Reporting:** The slider mechanism relies on honest self-reporting of opinion shifts.

Conclusion As AI models move from being tools to agents of persuasion, defensive infrastructure must evolve. MindPatch proves that we can build user interfaces that not only facilitate interaction but actively train users to detect and resist manipulation.

5. References

[Anthropic. \(2024\). Persuasion Dataset. Hugging Face.
https://huggingface.co/datasets/Anthropic/persuasion](https://huggingface.co/datasets/Anthropic/persuasion)

6. Appendix

Security Considerations

Dual-Use Risk The logic used to diagnose vulnerability (identifying effective triggers) mirrors the logic required to optimize attacks. A malicious actor could use the `GameSession` scoring engine to train a "persuasion-maximized" bot. To mitigate this, the current dataset is static, and the scoring logic is defensive (highlighting triggers rather than generating them).

Privacy and Profiling A user's "Vulnerability Profile" is highly sensitive psychometric data. Knowing that a specific user is easily swayed by "Appeals to Authority" is valuable for social engineering.

- *Mitigation:* The current prototype runs entirely locally or in a stateless cloud session (`st.session_state`). No user profiles are persisted to a

database. Future iterations would require end-to-end encryption for profile storage.

Dataset Bias The Anthropic Persuasion Dataset contains controversial topics. There is a risk that the "Antidotes" generated by our enrichment pipeline carries its own bias. We mitigate this by focusing the antidotes on the *structure* of the argument (e.g., "This uses circular reasoning") rather than the *content* of the claim.