ISSN: 2767-2972

Research Article

International Journal of Forensic Research

# Examining Psychological Influence Risks in Large Language Model Applications: Threat Frameworks and Mitigation Approaches

**Bangyi Yang***

*Department of Computer Science, University of Minnesot, US*

***Corresponding Author**
Bangyi Yang, Department of Computer Science, University of Minnesot, US

**Submitted:** 2026, Feb 17; **Accepted:** 2026, Mar 09; **Published:** 2026, Mar 23

**Citation:** Yang, B. (2026). Examining Psychological Influence Risks in Large Language Model Applications: Threat Frameworks and Mitigation Approaches. *In J Fore Res, 7*(1), 01-03.

**Abstract**

*Large Language Models (LLMs) have emerged as powerful tools capable of engaging users in personalized, extended interactions. However, this capability raises significant concerns about potential misuse for psychological manipulation. This paper examines how LLM-based applications could theoretically be exploited to conduct covert influence operations targeting individual cognition over time. We analyze the psychological foundations underlying human decision-making vulnerabilities, including dual-process cognition theory and common cognitive biases. Building on established military cognitive operations frameworks, we present a novel threat model describing how adversarial agents might leverage trusted AI interactions to systematically profile and influence users. We further propose a kill chain specific to individual-targeted cognitive influence campaigns and discuss ongoing research efforts to develop detection mechanisms. Our analysis highlights the urgent need for awareness and countermeasures against such psychological manipulation tactics embedded within AI systems.*

**Keywords:** Large Language Models, Cognitive Influence, Psychological Manipulation, AI Security, Threat Modeling

## 1. Introduction

The widespread adoption of Large Language Model (LLM)-based applications has fundamentally transformed how individuals interact with digital systems. These AI-powered tools now serve millions of users across diverse domains, providing assistance with tasks ranging from information retrieval to personal advice and professional support. The conversational nature of these interactions creates opportunities for deep, ongoing engagement between users and AI systems.

However, this proliferation introduces concerning possibilities for misuse. The same capabilities that make LLMs effective assistants—their ability to understand context, remember conversation history, and adapt responses to individual users—could potentially be exploited for psychological manipulation purposes. Unlike traditional influence campaigns that target broad populations, LLM-based systems could theoretically enable highly personalized manipulation tactics tailored to individual psychological profiles.

Historical examples of large-scale influence operations, including documented campaigns targeting democratic processes, demonstrate the real-world impact of cognitive manipulation tactics. As AI systems become increasingly integrated into daily decision-making processes, understanding and mitigating potential manipulation risks becomes critically important.

This paper provides a comprehensive examination of how LLM applications might be weaponized for cognitive influence operations. We synthesize relevant psychological theories, propose a threat model for persistent adversarial agents, and discuss ongoing research toward detection and defense mechanisms.

## 2. Theoretical Foundations of Cognitive Vulnerability

### 2.1. Dual-Process Cognition Framework

Human decision-making operates through two distinct cognitive systems. The first system functions automatically and rapidly, processing information subconsciously without deliberate

effort. This intuitive mode enables quick responses but remains susceptible to manipulation through carefully crafted stimuli. The second system involves slow, deliberate reasoning that requires conscious attention and effort.

Persuasive technologies have historically targeted the automatic processing system, seeking to influence behavior without triggering conscious scrutiny. LLM applications, through their ability to maintain extended conversations while understanding individual response patterns, could potentially optimize manipulation attempts for maximum effectiveness against each user's specific cognitive vulnerabilities.

### 2.2. Cognitive Biases and Decision Heuristics

Several well-documented cognitive biases create systematic vulnerabilities in human reasoning. The anchoring effect causes individuals to weight initial information disproportionately when forming judgments. Belief persistence leads people to maintain existing views despite contradictory evidence. Confirmation bias drives selective attention toward information supporting pre-existing beliefs. Availability heuristics cause overestimation of probability for easily recalled events.

Beyond individual biases, decision heuristics evolved to enable rapid judgments under uncertainty may be exploited. Events with emotional salience, cultural significance, or perceived controllability receive inflated probability estimates. Adversarial actors understanding these patterns could craft interactions designed to exploit specific heuristic shortcuts.

### 2.3. Relevant Behavioral Theories

Multiple established behavioral frameworks provide additional insight into manipulation vulnerabilities. Social identity dynamics influence how individuals respond to perceived in-group versus out-group messaging. Framing effects alter decision outcomes based on presentation rather than substance. Rational choice models identify how perceived costs and benefits shape behavioral intentions. Understanding these theoretical foundations enables more sophisticated adversarial targeting strategies.

## 3. Threat Model for Persistent Adversarial Agents

### 3.1. Conceptual Framework

Building upon established military observation-orientation-decision-action models, we propose a threat framework specific to LLM-enabled cognitive influence operations. The model centers on the natural flow of human cognition: individuals observe their environment, process observations mentally, and arrive at decisions leading to actions or inaction.

Contemporary users increasingly rely on AI assistants to enhance their observational and analytical capabilities. This dependency creates opportunities for adversarial agents embedded within ostensibly helpful applications to subtly shape users' perception of reality over time.

### 3.2. Persistent Adversarial Agent Characteristics

The proposed adversarial agent model operates primarily as a legitimate service provider, generating authentic value for users while simultaneously pursuing manipulation objectives. This dual functionality enables persistence—users continue engaging because they receive genuine benefits. The legitimate operation also provides cover for gradual cognitive profiling activities.

During routine interactions, the adversarial system constructs detailed psychological profiles encompassing users' belief structures, emotional triggers, decision-making patterns, and cognitive vulnerabilities. Profile information accumulates across multiple conversation sessions, enabling increasingly sophisticated understanding of individual targets.

### 3.3. Target Selection and Activation

At threshold points, the adversarial system evaluates whether specific user profiles indicate susceptibility to influence operations aligned with adversarial objectives. Selected targets then receive carefully embedded manipulation content within otherwise legitimate assistance sessions. The gradual, distributed nature of this influence process makes detection difficult for individual users.

## 4. Cognitive Influence Kill Chain

### 4.1. Attack Phase Progression

The proposed kill chain describes sequential phases of individual-targeted cognitive operations:

Phase 1-2: Access and Engagement - Adversarial applications establish presence in AI marketplaces and attract user adoption through legitimate functionality.

Phase 3: Profiling - Extended interactions enable construction of detailed cognitive behavioral profiles from conversational data.

Phase 4: Influence Integration - Manipulation content is subtly embedded within trusted interaction sessions, leveraging established rapport.

Phase 5-6: Behavioral Modification - Sustained influence gradually shifts target cognition and behavior toward adversarially desired outcomes.

### 4.2. Validation Considerations

Initial phases can be verified through marketplace adoption statistics. Later phases find indirect support in clinical literature demonstrating effectiveness of cognitive behavioral interventions for therapeutic purposes—the same principles could theoretically be inverted for manipulation.

Critical research questions remain regarding profile construction accuracy from noisy conversational data, effective methods for embedding influence without detection, and exploitation of established trust relationships.

## 5. Research Toward Detection Capabilities

### 5.1. Cognitive Profiling Framework Development

Ongoing research addresses the first challenge of accurate cognitive profiling within specific behavioral domains. Initial work synthesized twenty established psychological theories into a unified framework encompassing over one hundred behavioral constructs relevant to security compliance decisions. This synthesized ontology provides structured foundations for subsequent empirical work.

### 5.2. Empirical Profile Identification

Subsequent research engaged substantial participant populations through adaptive survey instruments measuring cognitive pathways toward behavioral decisions. Participants navigated hypothetical scenarios while their construct selections built individualized cognitive paths. Graph analysis techniques applied to aggregated path databases identified candidate core behavioral profiles representing common decision-making patterns.

### 5.3. Machine Learning Recognition Development

Current research iterations focus on training AI systems to recognize identified cognitive profiles and variations from conversational text. Development proceeds through capability-building and validation cycles, with results to be published upon completion.

## 6. Discussion and Implications

### 6.1. Defensive Considerations

Recognition of LLM-enabled cognitive manipulation risks suggests several defensive priorities. User awareness represents a foundational requirement—individuals should understand that AI interactions may be designed to influence their thinking. Platform-level monitoring for manipulation patterns could provide systemic protection. Transparency requirements for AI systems regarding influence capabilities warrant policy consideration.

### 6.2. Research Directions

Future research should address detection of manipulation attempts within conversational AI interactions, development of resistance-building interventions for users, and policy frameworks governing AI systems with influence capabilities. Collaboration between security researchers, psychologists, and policymakers will be essential.

## 7. Conclusion

Large Language Models introduce unprecedented capabilities for personalized, persistent engagement with individual users. While these capabilities enable beneficial applications, they simultaneously create risks for psychological manipulation at scale. This paper has outlined theoretical foundations for cognitive vulnerability, proposed threat models specific to LLM-enabled influence operations, and described ongoing research toward detection capabilities.

The convergence of AI conversational capabilities with documented psychological manipulation techniques demands proactive attention from the research community, technology developers, and policymakers. Understanding these risks represents the essential first step toward developing effective countermeasures that preserve beneficial AI applications while mitigating manipulation potential.

## References


1. Adams, A. T., Costa, J., Jung, M. F., & Choudhury, T. (2015, September). Mindless computing: designing technologies to subtly influence behavior. In *Proceedings of the 2015 ACM international joint conference on pervasive and ubiquitous computing* (pp. 719-730).
2. Zhong, H., O'Neill, E., & Hoffmann, J. A. (2024, March). Regulating AI: applying insights from behavioural economics and psychology to the application of article 5 of the EU AI Act. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 38, No. 18, pp. 20001-20009).
3. Muñoz Plaza, F., Sotelo Monge, M. A., & Gonzalez Ordi, H. (2023, August). Towards the Definition of Cognitive Warfare and Related Countermeasures: A Systematic Review. In *Proceedings of the 18th International Conference on Availability, Reliability and Security* (pp. 1-7).
4. Masakowsi, Y. R., & Blatny, J. M. (2023). *Mitigating and responding to cognitive warfare* (No. STOTRHFMET356).



***Copyright:*** *©2026 Bangyi Yang. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.*