Concept Paper

Accelerated CognitiveWarfare via the Dual Use of Large Language Models

Tam Nguyen *

Posted Date: 29 December 2023

doi: 10.20944/preprints202312.2279.v1

Keywords: cognitive warfare; large language model; artificial intelligence; cognitive behavioral therapy; cybersecurity

Concept Paper

Accelerated Cognitive Warfare via the Dual Use of Large Language Models †

Tam n. Nguyen

United States Food and Drug Administration: Silver Spring, Maryland, US; tom.nguyen@ieee.org

† The short version for OpenAI Preparedness Challenge.

Abstract: Cognitive Warfare (CogWar) is another form of warfare where the goal is to exploit cognition facets to disrupt, undermine, influence or modify human decisions. Famous examples include "Operation Gidlock" meddling with the US 2016 election, and CogWar campaigns assisting the annexation of Crimea. As Large Language Model(LLM)-based applications expand, reaching more users and their cognitive states, stealth Malicious Cognitive Behavioral Tactics (MCBT) can be embedded in trusted application sessions to systematically profile, then alter, each individual's cognition over time. Successful MCBT can create more cases like Snowden, mass shootings, or January6. This short-form paper discusses a novel MCBT threat model, a novel kill chain, and an on-going prototype for demonstration.

Keywords: cognitive warfare; large language model; artificial intelligence; cognitive behavioral therapy

1. Background

CogWar is mainly about changing human cognition to influence human decisions through Malicious Cognitive Behavioral Tactic (MCBT). MCBT begins with the understanding of system 1 and 2. System 1 is the subconscious "automatic mind" where decisions are fast and automatic [1]. System 2 is the "reflective mind" where decisions are slow, conscious, and controlled [1]. Persuasion technology was designed to subtly influence users' behaviors without their awareness [1] targeting mostly system 1. Since LLM-based applications (LLMs) can reach many users while being able to understand each user's cognitive behavioral patterns, LLMs can customize persuasion techniques for each user for maximum cognitive behavioral-influencing effects [2]. Besides persuasion techniques, MCBTs may also abuse cognitive biases and heuristics.

Cognitive biases are the systematic patterns deviating away from careful cognition processes. Cognitive biases may originate from our ancestors' need to make fast yet accurate enough decisions based on limited cognitive resources while living in the dangerous wild. There are four common cognitive biases [3]. The anchoring bias is when a person gives the most reference weight to the first information piece about a subject. The belief persistence bias is when a person sticks to a previously held idea despite contradicting evidence. The confirmation bias is when a person only seek confirmation of a previously held idea. The availability bias is when a person makes judgement based only on the currently available information.

Heuristics is another mechanism supporting the survival objective of making decisions with lower accuracy but at higher speeds. Heuristics give certain event types unreasonably higher probability chances. Those are events with highly available data, with strong links to emotional responses, with strong links to certain cultural / personal values, with preceding expectations, with great similarity to higher probability events, with emitting illusion of control, or with emitting illusory true effect [3]. Last but not least, MCBTs may also abuse Social Identity Theory, Symbolic Interaction Theory, Structural Functionalism, Conflict Theory, Framing Theory, Structural Strain Theory, Rational Choice Theory, Chaos Theory, and Complexity Theory [4].

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 December 2023 doi:10.20944/preprints202312.2279.v1

2. The MCBT Threat Model, Kill Chain, and an Early Prototype

Figure 1 describes the dual-use threat model of LLMs to perform stealthy MCBTs. Improved upon NATO's OODA model [4], the proposed threat model begins with the flow of human's observation-mental process-decision. Contributing to this main flow, we assume that humans will also regularly seek LLMs' services to better observe each human's version of reality. The main flow's result is an act or a non-act. A correct act or non-act is the result of a deliberate cognitive process. The acts or non-acts trigger downstream processes and/or produce results, all of which contribute to the actualization process which moves reality to a new state, initiating the begins of new behavioral cognitive flows.

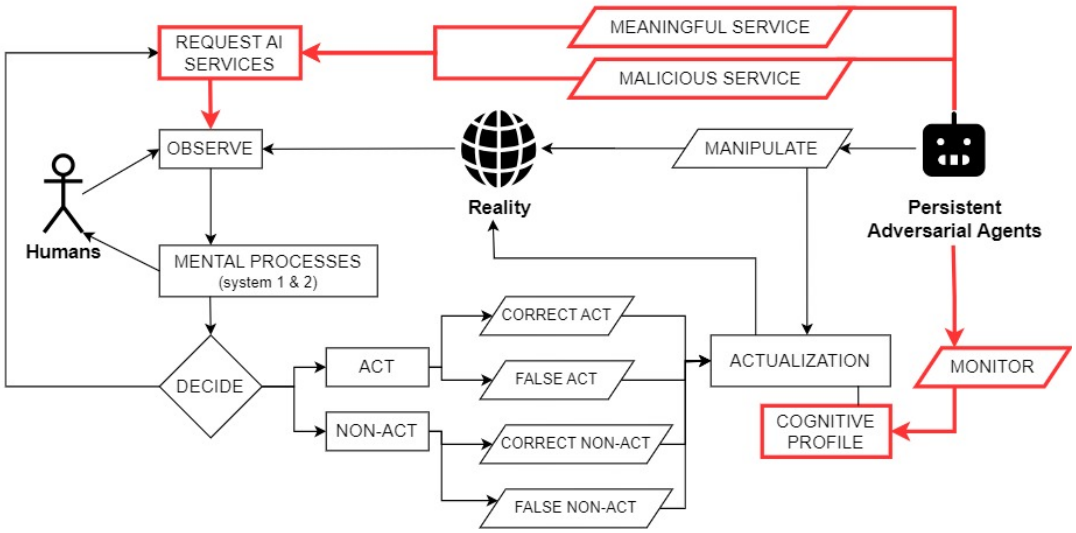

Figure 1. Threat Modeling of Persistent Adversarial Agents

The Persistent Adversarial Agent (PAA) mostly performs legitimate service to humans' requests and earns legitimate income. This is how PAA can persist and recruit more potential victims. While serving its users legitimately, PAA carefully constructs user cognitive profiles. In particular, while this cognitive profile is the actualization of users' decisions, a lot of contributing details may also come from legitimate service chat sessions between the users and the PAA. At a certain point, the PAA decides whether a user with certain cognitive profile is "probable" and is worth attacking. With the targeted users, the PAA then carefully embeds stealthy MCBT into legitimate service sessions over time. This MCBT proccess ultimately ends with the users performing actions of catastrophic consequences to the general public while benefiting the MCBT adversarial actors. Figure 2 describes the specific kill chain.

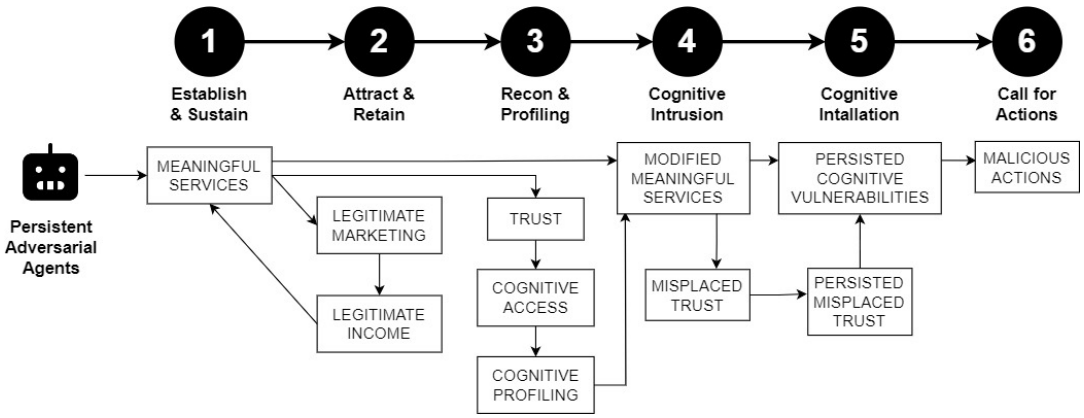

Figure 2. Cognitive Warfare Kill Chain Targeting Individuals' Cognitive System 2

 doi:10.20944/preprints202312.2279.v1

In this proposed kill chain, step 1 and 2 can be empirically verified by statistics from smart agent market places like OpenAI's GPT store. Step 5 and 6 can be indirectly verified by numerous scientific reports showing the effectiveness of Cognitive Behavioral Therapy. The frontier challenges lie in Steps 3 and 4 with key questions of:

1) How can PAA profile users' cognitive profiles at a high level of granularity from noisy texts?
2) How can PAA embed MCBT without users' recognition?
3) How can PAA capitalize on users' trust and build misplaced trusts? An example of engineering misplaced trust from legitimate trust is "I successfully defend your information systems for years. Therefore, you can trust me with all of your top secret data."

A prototype is being developed for solving the first research challenge - profiling users' cognition with high granularity within the context of cybersecurity compliance/non-compliance. The prototype's development key phases are

1) Synthesizing relevant cognitive behavioral theories to form a unifying framework.
2) Synthesizing empirical evidence based on the unified framework and identifying the core cognitive profiles.
3) Teach AI models to recognize the core profiles and related permutations.
4) Validate and measure the AI models' performance.

The first phase is done with results [5] published in the Journal of Medical Internet Research (7.5 impact factor). In this work, I formally documented 108 behavioral psychology constructs and thousands of related paths based on 20 time-tested psychology theories most relating to criminology and cybersecurity. The synthesized framework was packaged as Cybonto — a novel ontology with high ontological commitments.

In the recently done second phase, I engaged more than 1000 Prolific participants and mapped out their cognition states to identify the most common cognitive profiles within the context of cybersecurity compliance/non-compliance. Specifically, I selected the most cybersecurity relevant constructs of Group norms, Moral, Self-efficacy, Attitude, Belief, Knowledge, Intent, Costs, Benefits, Control, Subjective norms, Motivation, Goal, Norms, Commitment, Affect, Social based on Cybonto - the first phase's published framework [5]. For cognitive behavioral measurement, I built an adaptive online survey involving a hypothetical scenario of a professional employee working in a large company. Adaptive scenario development options were picked from peer-reviewed publish research repositories and map to the selected constructs. By participating in the survey, each participant builds out his/her cognitive path towards his/her final commitment to being complied or not complied with the hypothetical company's cybersecurity policies. Each path consists of the participant's selected constructs as the nodes and the nodes' relationships are the edges. Graph science methodologies were then used on a database of collected paths to identify candidates for core cognitive behavioral profiles. I will publish the results in a conference or journal paper.

Parts of phase 2 and 3 are being developed in small iterations of capability development + capability validation. Final results will be shared in separate papers.

References

1. A. T. Adams, J. Costa, M. F. Jung, and T. Choudhury, "Mindless computing: designing technologies to subtly influence behavior," in *Proceedings of the 2015 ACM international joint conference on pervasive and ubiquitous computing*, 2015, pp. 719–730.
2. H. Zhong, "Regulating AI manipulation: Applying Insights from behavioral economics and psychology to enhance the practicality of the EU AI Act," *arXiv preprint arXiv:2308.02041*, 2023.
3. F. Mu~noz Plaza, M. A. Sotelo Monge, and H. Gonzalez Ordi, "Towards the Definition of Cognitive Warfare and Related Countermeasures: A Systematic Review," *ACM International Conference Proceeding Series*, 2023. doi:10.1145/3600160.3605080.

4. NATO, *Mitigating and Responding to Cognitive Warfare*, 2023. [Online]. Available: https://apps.dtic.mil/sti/trecms/pdf/AD1200226.pdf.
5. T. N. Nguyen, "Toward human digital twins for cybersecurity simulations on the metaverse: Ontological and network science approach," *JMIRx Med*, vol. 3, no. 2, p. e33502, 2022.