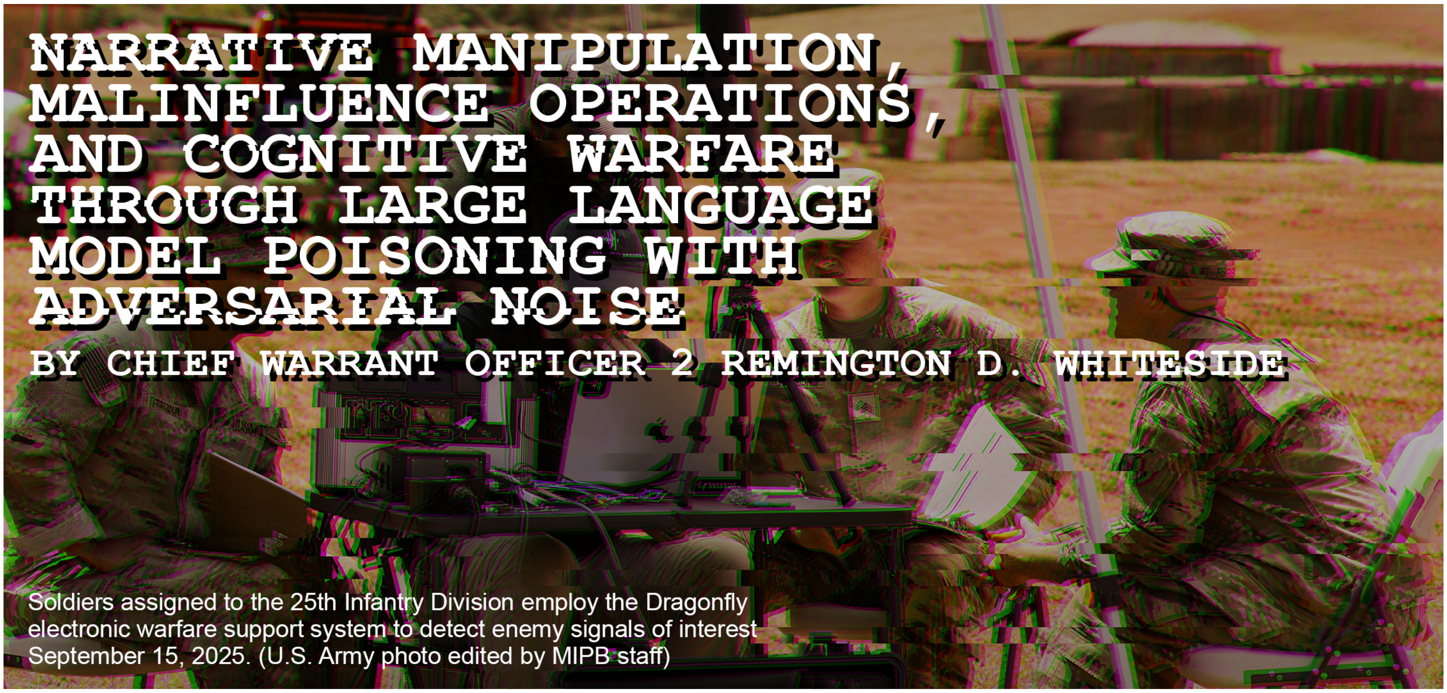
NARRATIVE MANIPULATION, MALINFLUENCE OPERATIONS, AND COGNITIVE WARFARE THROUGH LARGE LANGUAGE MODEL POISONING WITH ADVERSARIAL NOISE

BY CHIEF WARRANT OFFICER 2 REMINGTON D. WHITESIDE



Soldiers assigned to the 25th Infantry Division employ the Dragonfly electronic warfare support system to detect enemy signals of interest September 15, 2025. (U.S. Army photo edited by MIPB staff)

Editor's Note: This article contains several terms that not all readers will understand. Therefore, we added at the end a small glossary of the terms we felt were the most challenging.

Introduction

With every novel technology, exploitable vulnerabilities will arise. Adversaries will attempt to undermine integrity and further their own agendas through opportunistic exploitations. In the world of artificial intelligence development, one potential vulnerability and avenue of subversion is the use of *adversarial noise* against trained, structured data. This adversarial noise, also known as noisy data, has the potential to shape future operational readiness postures, or even conflict itself, in unprecedented ways. If unleashed against military data corpora and associated large language models (LLMs), adversarial noise will undoubtedly create dissonant ramifications in operational spaces. Specifically, these attacks will affect servicemembers' individual and collective narratives, sentiments toward organizational trust, and overall cognitive security, thus jeopardizing readiness. This article will highlight the threat adversarial noise poses to the psycho-cognitive states of servicemembers and their organizations through malinfluence and manipulation.

Impact on Individual Narrative

While the idea of *cognitive operations* is relatively novel in U.S. military thought circles, threat actors have long targeted the cognitive domain.[1] The difference between past and future tactics is the rapid advancement of technology, particularly the propagation of information and communication technologies and *artificial narrow intelligence*. According to a research team affiliated with Stanford, senior governmental decision makers are increasingly using LLMs to devise strategies and solutions in both war and policy.[2] This logically extends to servicemembers, who use civilian and military artificial narrow intelligence applications for everyday tasks. Despite assurances of security, vulnerabilities exist in the architecture of the corpora and models, all of which are exploitable.[3]

Just as the *logical layers* of LLMs are vulnerable to information attacks, the *persona layer* of information and communication technologies, including LLMs, is likewise vulnerable. It is important to note that even the construction of LLMs can be over-anthropomorphized and over-biased, potentially leading to inherent *dark patterns*. These dark design patterns can be emotionally and psychologically misleading to users, providing an example of susceptibility to narrative influence via prompting (also known as *influence warfare*). According to Dr. Ajit Maan, a defense and security strategist, "narratives are about *meaning*, not necessarily about the *truth*".[4] In other words, a narrative is a meaning-generative mechanism that composes individual and collective identity through experience, information transfer, and the search for knowledge.[5]

Artificial narrow intelligence now represents a figurative fountain of knowledge, as it is practical, mundane, and easily accessible. Military-specific GPTs (generative pre-trained transformers)—such as CamoGPT—are a go-to for informational needs. LLMs enable the acquisition of this knowledge and facilitate the construction of meaning for both civilians and military members.[6] Reliance on GPTs should be cautioned, however, as recent research from OpenAI, Inc., an American artificial intelligence research organization, shows a likely developing correlation between users' *socioaffective alignments* and increased anthropomorphizing of artificial narrow intelligence tools, risking the development of an artificial dependence on the technologies.[7] In other words, increased affective use of artificial narrow intelligence tools, such as LLMs and GPTs, will influence both emotional and psychological states of users.[8]

States of data are equally as important as user states of being. Corrupting the trained states of artificial narrow intelligence algorithms and LLM models with adversarial noise can introduce *artificial intelligence hallucinations*, including the dissemination of misleading or malicious information. A very similar tactic has been used by a Russian content aggregator affiliated with the News.ru network to target social networks in their digital areas of influence.[9] If unleashed on military data corpora, this can undermine the logical, foundational layers of future operations.[10] For example, military users may consume output corrupted by artificial intelligence hallucinations which will effectually degrade knowledge management, meaning construction, and core narratives over time.[11] Furthermore, this process would elicit biases in the consumer population of military personnel, triggering skepticism in their host organization or mission. These effects will negatively influence members' narratives and increase *cognitive dissonance* in operations.[12]

Impact on Organizational Trust

Subjective experience and directive output dictate the formation of organizational trust. LLMs can enable the initial composition of organizational narratives by propagating prompted information, ranging from systemic guidance to intelligence summaries. Attacking the organization's knowledge management core (i.e., data corpora and LLMs) could delegitimize the authoritative structures (i.e., military leadership) and negatively influence perceptions of and sentiments toward the organization (i.e., trust).[13] Undoubtedly, this would damage the operational climate, which in turn would affect servicemembers' morale.[14]

Data corpora and LLMs serve as the initial bridge to the individual and collective *belief-trust substrate*.[15] If adversarial noise corrupts an organization's knowledge core (i.e., data corpora, GPTs, and LLMs), the resulting processes would contradict information reflected from the organization's mission narrative.[16] This LLM poisoning could trigger disbelief in the collective narrative, weakening both individual morale and organizational trust. Chatbots could even be used to amplify further adversarial noise in the form of malign information across information and communication technologies architecture, infiltrating social networks frequented by U.S. servicemembers.[17] The resulting narrative engagement on social media would enable adversarial *cognitive maneuver*, exemplifying the use of malinfluence to engage and manipulate individual and collective biases for effectual motives and resulting in cognitive posturing of a targeted population.[18] Combinations of artificial narrow intelligence data poisoning and cognitive maneuver will enable adversaries to destabilize trustworthiness in military operations and associated communities using weaponized misinformation.[19]

Impact on Cognitive Security

Historically, adversaries have manipulated psychological states of targeted populations (both military and civilian) through information operations and active measures to achieve an operational advantage. Now, this focus will expand to target not only psychological states, but also certain cognitive states, primarily learning and perception.[20] After influencing narratives and eroding organizational trust, adversaries will certainly leverage adversarial noise to engage *cognitive centers of gravity*, notably those centers directly tethered to and reinforced by artificial narrow intelligence.[21] They will seek to manipulate and undermine military LLM- and artificial narrow intelligence-powered centers of knowledge, thus corrupting the informational engines of thought and dialogue.[22] Military education institutions and knowledge bases will undoubtedly be prime targets in the fight for an information advantage.

Refined algorithms could penetrate security layers and inject adversarial noise into data corpora used to enrich military education and inform military operations.[23] Conjunctively, threat actors will leverage botnets to push amplified adversarial noise across information and communication technologies architecture and to seed malign information (e.g., disinformation) via social media channels to overwhelm audiences cognitively.[24] Promotion of mass skepticism in military educational systems would result in the creation of cognitive dissonance, further delegitimizing authority. The adversary thus achieves his goal of *internal negation*, sowing civil-political discord amongst military populations via database poisoning from within.[25] Ultimately, these technological actions will subvert and degrade the status of the military's cognitive security at micro- to macro-levels.

Conclusion

The use of adversarial noise to poison data corpora and manipulate the cognitive states of military members and organizations is not simply a hypothetical threat scenario but is rooted in actual occurrences. Individual and collective narratives, organizational trust, and cognitive security postures are vulnerable to the effects of artificial narrow intelligence-facilitated information manipulation and malinfluence. The injection of adversarial noise into data architectures and models, hallucinations from poisoned data, and increased dependency on compromised artificial narrow intelligence can result in drastically ordered effects on readiness posture at both personal and organizational levels if left unchecked. Until more stringent information and cognitive security measures are emplaced and more effective research practices materialize, these vulnerabilities will severely impact operations on the competition-conflict spectrum across the cognitive domain.

Terms and Definitions

adversarial noise: carefully crafted, often imperceptible disruptions or modifications to input data intentionally introduced in adversarial attacks to deceive artificial intelligence models.[26]

artificial intelligence hallucinations: occur when an algorithmic system produces incorrect or misleading results, even if it appears to be generating coherent, logical outputs.[27]

artificial narrow intelligence: often called weak artificial intelligence, this is the current state of artificial intelligence with systems designed and trained to perform a specific, narrow task or range of tasks.[28]

belief-trust substrate: psychologically speaking, a *substrate* refers to the biological brain infrastructure that facilitates a particular behavior. There are different substrates for various neurological functions; therefore, the *belief-trust* substrate is the physical chunk of one's central nervous system where belief and trust interact and reconcile.[29]

cognitive centers of gravity: the defining focus of a person's thoughts, feelings, and/or behaviors, often a reflection of that person's core values.[30]

cognitive dissonance: the simultaneous existence of conflicting beliefs and an individual's attempts to align them.[31]

cognitive operations: tactical actions in support of cognitive warfare–the subset of general warfare focused on influencing or disrupting individuals' and groups' cognition, or thinking processes, to gain an advantage.[32]

cognitive maneuver: strategically influencing the perceptions and thought processes of an adversary.[33]

cyberspace layers: cyberspace has three interrelated layers: the *physical network layer*, which is the actual infrastructure that provide information technology functionality; the *logical network layer*, which is the logic programming and code that drives functionality; and the [cyber-]*persona layer* which represents the people interacting in and with cyberspace.[34]

dark patterns: deceptive user interfaces employed by e-commerce to manipulate users' behavior into making decisions that benefit the company but not necessarily the user.[35]

influence warfare: the use of information, including propaganda and disinformation, to influence the perceptions and actions of an adversary.[36]

internal negation: negation simply refers to rejecting something. Therefore, psychologically speaking, internal negation is the rejection of a thought, feeling or belief, as opposed to external negation, which is the rejection of aspects of outside reality or other people.[37]

socioaffective alignments: the way an artificial intelligence system behaves within the social/psychological ecosystem co-created with its user, where preferences and perceptions evolve through mutual influence.[38]

Endnotes

1. Austin Coombs, "Persuade, Change, and Influence with AI: Leveraging Artificial Intelligence in the Information Environment," Modern War Institute, October 25, 2024, https://mwi.westpoint.edu/persuade-change-and-influence-with-ai-leveraging-artificial-intelligence-in-the-information-environment/.

2. Juan-Pablo Rivera et al., "Escalation Risks from LLMs in Military and Diplomatic Contexts," Policy Brief, Human-Centered Artificial Intelligence, Stanford University, May 2, 2024, https://hai.stanford.edu/policy/policy-brief-escalation-risks-llms-military-and-diplomatic-contexts.

3. Daniel Alexander Alber et al., "Medical Large Language Models are Vulnerable to Data-Poisoning Attacks," *Nature Medicine* 31 (2025): 618–626, https://doi.org/10.1038/s41591-024-03445-1; and William N. Caballero and Phillip R. Jenkins, "On Large Language Models in National Security Applications," Stat: *The ISI's Journal for the Rapid Dissemination of Statistics Research* 14, no. 2 (March 2025), https://doi.org/10.1002/sta4.70057.

4. Ajit Maan, *Narrative Warfare* (Narrative Strategies Ink, 2018), 16.

5. Ibid.

6. Caballero and Jenkins, "On Large Language Models."

7. Esben Kran et al., "DarkBench: Benchmarking Dark Patterns in Large Language Models," published as a conference paper at the 2025 International Conference on Learning Representations, April 24, 2025 to April 28, 2025, https://openreview.net/pdf?id=odjMSBSWRt; and Jason Phang et al., "Investigating Affective Use and Emotional Wellbeing on ChatGPT," Massachusetts Institute of Technology Media Lab, March 21, 2025, https://www.media.mit.edu/publications/investigating-affective-use-and-emotional-well-being-on-chatgpt/.

8. Phang et al., "Emotional Wellbeing on ChatGPT."

9. "The American Sunlight Project Unveils Detailed Report on the Critical Threat of Russian Disinformation in AI Models," Updates, American Sunlight Project, February 26, 2025, https://www.americansunlight.org/updates/the-american-sunlight-project-unveils-detailed-report-on-the-critical-threat-of-russian-disinformation-in-ai-models.

10. Alber et al., "Medical Large Language Models"; Caballero and Jenkins, "On Large Language Models"; and Mylola Makhortykh et al., "Stochastic Lies: How LLM-Powered Chatbots Deal with Russian Disinformation About the War in Ukraine," *Harvard Kennedy School (HKS) Misinformation Review* (2024), https://doi.org/10.37016/mr-2020-154.

11. Caballero and Jenkins, "On Large Language Models"; and Makhortykh et al., "Stochastic Lies."

12. Coombs, "Influence with AI."

13. "Critical Threat of Russian Disinformation," American Sunlight Project; and Coombs, "Influence with AI."

14. Coombs, "Influence with AI."

15. Vinícius Marques da Silva Ferreira et al., "Lógica Fuzzy Aplicada à análise Comportamental E Conhecimento Da Guerra Cognitiva Em Redes Sociais: Um Modelo De extração E mineração De Dados," *Revista De Gestão E Secretariado (Management and Administrative Professional Review)* 15, no. 5 (2024): 3708, https://doi.org/10.7769/gesec.v15i5.3708.

16. Ibid.

17. "Critical Threat of Russian Disinformation," American Sunlight Project; and Coombs, "Influence with AI."

18. "Critical Threat of Russian Disinformation," American Sunlight Project; James E. Zanol and Brian M. Pierce, "Overcoming the Challenges in Implementing Emerging Maneuver Concepts," *Military Review* 98, no. 3 (May-June 2018): 87-92, https://www.armyupress.army.mil/Portals/7/military-review/Archives/English/Zanol-Emerging-Maneuver-Concepts.pdf.

19. Coombs, “Influence with AI.”

20. Irwin J. Mansdorf, “Psychological Warfare After the Guns Are Stilled: The Need for Cognitive Reframing the ‘Day After’,” *Jerusalem Issue Briefs* 23, no. 4 (January 2024), https://jcpa.org/article/psychological-warfare-after-the-guns-are-stilled/.

21. Coombs, “Influence with AI.”

22. “Critical Threat of Russian Disinformation,” American Sunlight Project; and Makhortykh et al., “Stochastic Lies.”

23. Alber et al., “Medical Large Language Models”; and Caballero and Jenkins, “On Large Language Models.”

24. “Critical Threat of Russian Disinformation,” American Sunlight Project.

25. Coombs, “Influence with AI.”

26. The AllBusiness.com Team, “Noise in AI,” AI Dictionary, *Time*, April 3, 2025, https://time.com/collections/the-ai-dictionary-from-allbusiness-com/7273975/definition-of-noise-in-ai/.

27. Anna Choi and Katelyn Xiaoying Mei, “What are AI Hallucinations? Why AIs Sometimes Make Things Up,” The Conversation, March 21, 2025, https://theconversation.com/what-are-ai-hallucinations-why-ais-sometimes-make-things-up-242896.

28. Ultralytics Inc., “Artificial Narrow Intelligence (ANI),” 2025, https://www.ultralytics.com/glossary/artificial-narrow-intelligence-ani.

29. Christopher Bergland, “The Neuroscience of Trust,” *Psychology Today*, August 12, 2015, https://www.psychologytoday.com/us/blog/the-athletes-way/201508/the-neuroscience-trust.

30. Lexicon of Psychology, “Center of gravity,” accessed 28 May 2025, https://www.psychology-lexicon.com/cms/glossary/36-glossary-c/7621-center-of-gravity.html.

31. Saul McLeod, “What Is Cognitive Dissonance Theory?” SimplyPsychology, June 20, 2025, https://www.simplypsychology.org/cognitive-dissonance.html.

32. Kyaw Jaw Sine Marma, “Cognitive Warfare: The Invisible Frontline of Global Conflicts,” Modern Diplomacy, February 12, 2025, https://moderndiplomacy.eu/2025/02/12/cognitive-warfare-the-invisible-frontline-of-global-conflicts/.

33. Patricia DeGennaro, “The Power of Cognitive Maneuver: Don’t Underestimate Its Value,” Small Wars Journal, September 19, 2017, https://archive.smallwarsjournal.com/jrnl/art/the-power-of-cognitive-maneuver-don%E2%80%99t-underestimate-its-value.

34. Chairman of the Joint Chiefs of Staff, Joint Publication 3-12, *Joint Cyberspace Operations* (Joint Staff, 2022), I-3–I-4.

35. Kran et al., “DarkBench.”

36. Kung Chan, “On Influence Warfare,” Geopolitical Review, ANBOUND, October 27, 2024, www.anbound.com/Section/ArticleView_34097_14.htm.

37. Lexicon of Psychology, “Negation,” accessed 28 May 2025, https://www.psychology-lexicon.com/cms/glossary/47-glossary-n/22413-negation.html.

38. H.R. Kirk et al., “Why Human-AI Relationships Need Socioaffective Alignment,” *Humanit Soc Sci Commun* 12, 728 (2025), https://www.nature.com/articles/s41599-025-04532-5.

CW2 Remington Whiteside is the 35P, Signals Intelligence Voice Interceptor, Course Manager and chief instructor for Bravo Company, 344th Military Intelligence (MI) Battalion, 111th MI Brigade, Goodfellow Air Force Base, TX. He previously served as an observer, coach, and trainer specializing in tactical signals intelligence, open-source intelligence, and intelligence support to cyber, information, and electromagnetic warfare at the Joint Readiness Training Center, Fort Johnson, LA. He is also an academic researcher in MIDLE (media, information, and data literacy education), malign information and malinfluence, as well as narrative. He holds an undergraduate degree in Middle Eastern studies, a graduate degree in applied linguistics, and a doctorate in education.