Selected Papers of #AoIR2024:
The 25th Annual Conference of the
Association of Internet Researchers
Sheffield, UK / 30 Oct - 2 Nov 2024

FROM BLACK TO WHITE: DISSECTING PROPGAGNDA IN NUCLEAR EMERGENCIES, FINDING AI-ENABLED DISINFORMATION

Seungtae Han
Georgia Institute of Technology

Brenden Kuerbis
Georgia Institute of Technology

Introduction

Disinformation during nuclear emergencies (DiE) represents a critical challenge to public safety and institutional stability, potentially undermining both immediate crisis response capabilities and long-term public trust in nuclear governance. This research examines how state actors strategically deploy Jowett and O'Donnell's (2014) propaganda models—specifically the Legitimating Source Model (LSM) and Deflective Source Model (DSM)—in their disinformation campaigns during nuclear emergencies. Additionally, we investigate how the integration of generative AI technologies might theoretically enhance or undermine these established propaganda models.

Our research analyzes two contrasting cases: the Zaporizhzhia Nuclear Power Plant (ZNPP) crisis following Russian occupation in March 2022, and the Fukushima Daiichi Nuclear Power Plant (FNPP) incident's ongoing water discharge controversy. These cases offer distinct contexts for examining how disinformation operates under different types of nuclear emergencies, while also providing insights into the potential role of emerging technologies in shaping future disinformation campaigns (Hoban & Rister, 2024). The unprecedented nature of these events—Europe's largest nuclear facility becoming embroiled in active military conflict and Japan's controversial decision to release treated radioactive water into the Pacific Ocean—creates unique opportunities to examine how different types of nuclear emergencies generate distinct patterns of disinformation and manipulation of public perception.

Methodology

Our research methodology encompasses analysis of 568 instances of propaganda, monitored across diverse media platforms including Facebook, YouTube, TikTok, VK,

Han, S., & Kuerbis, B. (2024, October). From Black to White: Dissecting Propaganda in Nuclear Emergencies, Finding AI-Enabled Disinformation. Paper presented at AoIR2024: The 25th Annual Conference of the Association of Internet Researchers. Sheffield, UK: AoIR. Retrieved from http://spir.aoir.org.

Weibo, and Naver, alongside Russian and Ukrainian Telegram channels. We examined traditional media channels owned or sponsored by states in Russia, Ukraine, China, South Korea, and Japan, complemented by official statements from governments, private companies, and the International Atomic Energy Agency (IAEA). The data collection period spanned from March 2022 to December 2023, capturing the evolution of narratives across multiple critical events and policy decisions.

Following George's (1954) approach to propaganda analysis, we employed qualitative interpretation to understand both the manifest content and latent purposes of disinformation campaigns. The research categorized disinformation narratives using Martino et al.'s (2020) framework of 18 propaganda techniques, enabling systematic analysis of how these methods are deployed across different contexts and platforms (Méndez-Muros et al., 2024). This comprehensive approach allowed us to map the complex networks through which disinformation flows and identify the key mechanisms by which it gains or loses credibility (Luceri et al., 2024).

Our analytical framework incorporated three distinct levels of examination: the macro level focusing on institutional relationships and power dynamics, the meso level examining network structures and information flow patterns, and the micro level analyzing specific content and narrative strategies. This multi-layered approach enabled us to identify patterns and relationships that might not be apparent through single-level analysis.

Model Analysis Findings

Our analysis through the LSM and DSM frameworks revealed distinct patterns in how state actors deploy and maintain disinformation campaigns. In the FNPP case, the LSM analysis demonstrated how South Korean and Chinese media outlets mutually reinforced false narratives about Japan's alleged influence over the IAEA, creating a self-validating network of misleading claims (Mabon, 2024). The propagandists strategically employed seemingly independent sources to legitimize their allegations, with each entity citing others' claims as corroborating evidence (Rid, 2020). This process of mutual reinforcement created an echo chamber effect that amplified the perceived credibility of false narratives while simultaneously making it more difficult for accurate information to penetrate these networks.

The research identified several key disinformation narratives in each case. For ZNPP, these included claims about armed provocations, power supply issues, IAEA inspections, militarization of the facility, and alleged presence of weapons of mass destruction (Yaroshchuk, 2023). In the FNPP case, three major narratives emerged: allegations about failed water purification systems, claims about pre-existing radioactive leaks, and accusations of IAEA bias (Sawano et al., 2019). Each narrative demonstrated sophisticated application of propagandistic techniques, often combining multiple methods to enhance persuasiveness and exploit existing societal tensions and fears.

The DSM analysis revealed sophisticated tactics for obscuring original sources of disinformation, particularly evident in the ZNPP case. Russian state actors employed

multiple layers of intermediary channels to disseminate false claims about Ukraine's alleged development of dirty bombs and chemical weapons, maintaining plausible deniability while shaping public perception. This approach involved creating networks of seemingly independent sources that would amplify and validate each other's claims, making it increasingly difficult to trace information back to its original source (Kim et al., 2013). These networks demonstrated remarkable resilience, adapting their narratives and tactics in response to fact-checking efforts and counter-narratives.

A significant finding emerged regarding the effectiveness of these propaganda models in practice. While our analysis confirmed active state involvement in creating and disseminating disinformation for political and economic purposes, we observed limited success in these efforts to shift public perception. This limitation appeared to stem from deeply entrenched institutional trust within domestic audiences and corresponding distrust of foreign sources, suggesting that the effectiveness of information operations faces considerable constraints when confronting established belief systems. The research also revealed that attempts to undermine institutional credibility often backfired, reinforcing existing trust patterns rather than eroding them.

Theoretical Extensions

Our research extends Jowett and O'Donnell's propaganda models by examining how generative AI technologies might affect their operational mechanisms. Within the DSM framework, generative AI shows potential to enhance disinformation capabilities through three key mechanisms: content multiplication, synthetic intermediary creation, and plausible deniability amplification (Zhou et al., 2023). These capabilities could allow propagandists to create more sophisticated networks of seemingly independent sources, each generating and amplifying consistent narratives while maintaining the appearance of organic information flow.

However, the technology faces fundamental limitations in establishing long-term credibility necessary for effective source deflection (Fredheim & Pamment, 2024). The challenge lies not in the creation of content but in building and maintaining the authentic relationships and reputation that give information sources their persuasive power. AI-generated content, while potentially convincing in isolation, struggles to establish the sustained credibility necessary for effective propaganda campaigns.

Our analysis suggests that generative AI may actually undermine the LSM's core mechanism through credibility erosion, authentication vulnerability, and inability to replicate complex social dynamics required for effective legitimization. The LSM relies on leveraging genuine institutional credibility, which AI-generated content cannot authentically replicate regardless of its sophistication (Yang & Menczer, 2023). The increasing sophistication of AI detection tools and growing public awareness of synthetic content further compounds these limitations.

Contrary to widespread concerns about AI-enabled disinformation, our empirical research found no evidence of state actors employing generative AI in the context of these nuclear emergencies. Instead, we observed continued reliance on traditional, human-crafted disinformation disseminated through established media channels. This

finding suggests a significant gap between theoretical capabilities and practical implementation of AI in disinformation campaigns, particularly in high-stakes scenarios like nuclear emergencies. The absence of AI-generated content in these cases may indicate that state actors recognize the risks and limitations of deploying such technologies in situations where maintaining credibility is crucial for achieving strategic objectives.

References

Fredheim, R., & Pamment, J. (2024). Assessing the risks and opportunities posed by AI-enhanced influence operations on social media. *Place Branding and Public Diplomacy*. https://doi.org/10.1057/s41254-023-00322-5

George, A. L. (1954). *Propaganda Analysis a Study of Inferences Made from Nazi Propaganda in World War II*. Greenwood Press.

Hoban, I., & Rister, A. (2024). Nuclear anxiety as an instrument of war: The use of news media to shape and respond to the disinformation campaign surrounding the Zaporizhzhia nuclear power plant. *Media, War & Conflict*, *00*(0), 1-19. https://doi.org/10.1177/17506352241256575

Jowett, G., & O'Donnell, V. (2014). *Propaganda and Persuasion* (5th ed.). Sage.

Kim, Y., Kim, M., & Kim, W. (2013). Effect of the Fukushima nuclear disaster on global public acceptance of nuclear energy. *Energy Policy*, *61*, 822-828.

Luceri, L., Boniardi, E., & Ferrara, E. (2024). Leveraging Large Language Models to Detect Influence Campaigns on Social Media. In *Companion Proceedings of the ACM Web Conference 2024* (pp. 1459-1467).

Mabon, L. (2024). Treated water releases from the Fukushima Dai'ichi nuclear power plant: An overview of the decision-making process and governing institutions. *Marine Policy*, *163*, Article 106120.

Martino, G. D. S., Cresci, S., Barrón-Cedeño, A., Yu, S., Di Pietro, R., & Nakov, P. (2020). A survey on computational propaganda detection. *arXiv preprint arXiv:2007.08024*.

Méndez-Muros, S., Alonso-González, M., & Pérez-Curiel, C. (2024). Disinformation and Fact-Checking in the Face of Natural Disasters: A Case Study on Turkey-Syria Earthquakes. *Societies*, *14*(4), 43.

Rid, T. (2020). *Active Measures: The Secret History of Disinformation and Political Warfare* (1st ed.). Farrar, Straus, and Giroux.

Sawano, T., Ozaki, A., Hori, A., & Tsubokura, M. (2019). Combating 'fake news' and social stigma after the Fukushima Daiichi Nuclear Power Plant incident-the importance

of accurate longitudinal clinical data. *QJM: Monthly Journal of the Association of Physicians*, *112*(7), 479-481.

Yang, K. C., & Menczer, F. (2023). Anatomy of an AI-powered malicious social botnet. *Journal of Quantitative Description: Digital Media* (2024). https://doi.org/10.48550/arXiv.2307.16336

Yaroshchuk, O. (2023). Russian disinformation strategy in the field of nuclear security: Examining key narratives. *Ukraine Analytica*, *1*(30), 27-36.

Zhou, J., Zhang, Y., Luo, Q., Parker, A. G., & De Choudhury, M. (2023). Synthetic lies: Understanding ai-generated misinformation and evaluating algorithmic and human solutions. *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 1-20.