ADMM Cybersecurity and Information Centre of Excellence

UPDATE ON
THE
INFORMATION DOMAIN

Issue 01/26 (Jan) – Special Edition Op-Ed

Unbearable Vulnerabilities to False Information?

by Prof Lee Eun-Ju, Director of Center for Trustworthy AI, Seoul National University

Introduction

1. Public anxiety about mis- and disinformation is at an all-time high worldwide. Recent advances in generative Artificial Intelligence (AI) have added another layer to this global challenge by enabling the large-scale production of inaccurate, misleading, and even entirely fabricated content, commonly referred to as AI hallucinations. Misinformation spreads rapidly and shapes how people make sense of reality, not simply because people are careless or malicious. Rather, it emerges from the interaction of human cognition, media systems, and emerging technologies, which together undermine our ability to discern truth from falsehood.

To Err Is Human? Cognitive Biases and Limitations

2. At the individual level, we rarely approach information as neutral judges. More often than we would like to admit, we rely on cognitive shortcuts that help us navigate complex information environments efficiently, rather than expending the effort needed to maximize accuracy. To encourage more thoughtful processing of information, and subsequently improve truth discernment, Pennycook and colleagues (2021) introduced what they call “accuracy nudge,” a minimal prompt that briefly shifts individuals’ attention toward accuracy. Across

multiple studies, they found that subtle interventions, such as asking participants to evaluate the accuracy of an unrelated news headline, reliably promote more discerning information sharing by reducing individuals' willingness to share false information more than true information (see Pennycook & Rand, 2022). Because accuracy nudges operate by simply redirecting attention to accuracy, they offer a scalable approach to improving the quality of information circulating online.

3. Inattention to accuracy, however, is not the only reason why people struggle to assess the veracity of information online. Factors such as partisan identity, prior beliefs, and motivational goals often shape how information is interpreted and shared in ways that support their existing beliefs, attitudes, and values – i.e., confirmation bias. Building on the idea that brief and timely interventions can guide how people engage with questionable content online, researchers have examined whether warning labels may encourage greater scrutiny of unreliable claims and potentially temper biased processing. In Lee and Jang's (2023) studies, participants were primed to think about misinformation, either by watching a short media-literacy video highlighting the risks of misinformation or by answering a question about their prior exposure to "fake news." Across two studies, these interventions did not significantly reduce partisan bias in truth judgments of COVID-19 information. Moreover, while such priming helped participants better identify false content in Study 1, it also led participants to become more skeptical of true information in Study 2, suggesting that warnings about misinformation may inadvertently increase blanket skepticism or cynicism.

4. Perhaps unsurprisingly, vulnerability to false information is not evenly distributed. Lee and Chung (2025) demonstrate that individuals' responses to fact-checks depend on cognitive traits such as need for cognition and cognitive reflection. That is, individuals who enjoy analytical thinking and engage more reflectively (vs. impulsively) with information are more likely to attend to corrective information and adjust their truth judgments accordingly. Although fact-checks are often considered as a key tool for combating misinformation, correction strategies that assume a uniformly rational public are likely to fall short.

How Not to Fall for AI-Hallucinations?

5. When language is well-structured, specific, and contextually appropriate, people tend to infer accuracy, even when the content is wrong. AI hallucinations exploit this long-standing human bias. Extending previous works on accuracy nudge, Nahar et al. (2024) examined whether adding a warning label ("The responses may contain inaccurate information about people, places, or facts") helps people distinguish between genuine information and hallucinations of varying degrees in a simulated Q & A session. The warning (vs. no warning) improved participants' ability to detect hallucinations without making them distrust genuine content, but no effect of warning was found for likes and shares. That is, merely alerting people to potential inaccuracies of AI outputs may not stop people from liking or sharing AI-generated misinformation.

6. In a follow-up study, Nahar et al. (2025) investigated whether the integration of web search results into LLMs, often called retrieval-augmented generation (RAG), enables people to identify hallucinated content. Specifically, participants either did their own searching (dynamic search) or saw pre-selected search results by the AI system (static search). Participants were better able to spot AI hallucinations when presented with search results, whether the search was participant-led or system-led. However, they also evaluated the LLM more negatively. Overall, RAG enhanced users' truth discernment while hurting the AI system that produced faulty outputs.

Media Competition Drives the Spread of Misinformation

7. Although cognitive limitations account for why people believe and share false information, they do not fully explain why misinformation spreads so effectively. Media systems can amplify these vulnerabilities by rewarding speed, novelty, and engagement over accuracy. Amini et al.'s simulation study (2025) demonstrates that misinformation can emerge from competitive pressure in the information ecosystem. By building a mathematical competition game, where outlets choose between factual information and misinformation, the authors found that hyperpartisan outlets end up spreading most

misinformation, and competition can create an “arms race” dynamic where one source’s misinformation increases pressure on others to follow.

Implications for Defense and National Security

8. The stakes of misinformation and AI hallucinations extend far beyond individual decision-making and everyday media use. In defense and national security contexts, false information, whether human-produced or AI-generated, can be weaponized to distort intelligence, inflame panic during crises, or erode public trust in social institutions. The same cognitive and media dynamics observed in civilian settings can have far more catastrophic consequences when applied to security threats.

9. Addressing these risks requires treating cognitive security as a core component of national resilience. This involves building media platforms that prioritize accuracy over speed, developing legal frameworks that ensure the safe and responsible deployment of AI systems, and cultivating epistemic resilience through media literacy education, which entails the capacity to manage uncertainty while embracing intellectual humility. To this end, it is all the more crucial to understand clearly how people select, process, and respond to information in this increasingly AI-infused world.

**The views expressed in this Info Digest are that of Prof Lee Eun-Ju, a member of ACICE’s Experts Panel. Prof Lee is Director of Center for Trustworthy AI at the Seoul National University. She held major international leadership roles, including serving as President of the International Communication Association (ICA) (2023-2024), and previously as Editor-in-Chief of Human Communication Research (2017-2020).*

CONTACT DETAILS

All reports can be retrieved from our website at www.acice-asean.org/resource/.

For any queries and/or clarifications, please contact ACICE at ACICE@defence.gov.sg

Prepared by:
ADMM Cybersecurity and Information Centre of Excellence

….

REFERENCES

Articles

1. How Media Competition Fuels the Spread of Misinformation. [Link: https://www.science.org/doi/full/10.1126/sciadv.adu7743]

2. Thinking Hard, Thinking Smart: How News Users' Cognitive Traits Guide Their Responses to Fact-Checks. [Link: https://www.tandfonline.com/doi/full/10.1080/21670811.2025.2492209]

3. How Political Identity and Misinformation Priming Affect Truth Judgments and Sharing Intention of Partisan News. [Link: https://www.tandfonline.com/doi/full/10/1080/21670811.2022.2163413]

4. Fakes of Varying Shades: How Warning Affects Human perception and Engagement Regarding LLM Hallucinations. [Link: https://openreview.net/pdf?id=c30qeMg8dv]

5. Catch Me if You Search: When Contextual Web Search Results Affect the Detection of Hallucinations. [Link: https://www.sciencedirect.com/science/article/abs/pii/S0747563225002109]

6. Shifting Attention to Accuracy can Reduce Misinformation Online. [Link: https://www.nature.com/articles/s41586-021-03344-2]

7. Accuracy Prompts are a Replicable and Generalizable Approach for Reducing the Spread of Misinformation. [Link: https://www.nature.com/articles/s41467-022-30073-5]