Misinformation and Disinformation in AI: the PICUS Lab Experience

Lidia Marassi[1], Narendra Patwardhan[1], Stefano Marrone[1,*] and Carlo Sansone[1]

[1]*Department of Electrical Engineering and of Information Technologies (DIET), University of Naples Federico II, Via Claudio 21, Naples, 80125, Italy*

Abstract

The rise of generative artificial intelligence technologies, such as Large Language Models (LLMs) and Generative Adversarial Networks (GANs), has profoundly reshaped the information landscape, enabling both innovation and new forms of risk. This paper examines the challenges posed by AI-generated misinformation and disinformation, focusing on their sociopolitical, technical, and ethical implications. We analyze current regulatory efforts—such as the European Union's Artificial Intelligence Act—and present mitigation strategies developed within the *Hominis* project at the PICUS Lab. In particular, the Hominis model emphasizes data verifiability, epistemic filtering, and calibrated response generation to ensure trustworthy outputs grounded in reliable sources. Our integrated lifecycle approach aims to enhance transparency, robustness, and resistance to manipulative content, ultimately addressing the growing erosion of public trust in media and institutions.

Keywords

Misinformation, Disinformation, Generative Models, Trustworthy AI, AI Governance

1. Introduction

In recent years, artificial intelligence (AI) has undergone a significant transformation, driven by advances in generative models such as Large Language Models (LLMs) and Generative Adversarial Networks (GANs). These technologies have enabled unprecedented capabilities in content creation, ranging from automated text generation to realistic synthetic media. However, alongside these innovations, new threats have emerged. Chief among them is the growing use of AI systems to produce and disseminate misinformation (false content shared without intent to deceive) and disinformation (false content shared with deliberate intent to mislead).

The ability of generative AI to create content that is stylistically coherent, factually plausible, and emotionally persuasive has opened the door to scalable manipulation of information ecosystems. This poses serious challenges to democratic institutions, public trust, and information integrity, especially in political and scientific contexts.

In this paper, we explore the technical, ethical, and societal dimensions of AI-generated misinformation and disinformation. We begin by examining the mechanisms by which generative

Ital-IA 2025: 5th National Conference on Artificial Intelligence, organized by CINI, June 23-24, 2025, Trieste, Italy

*Corresponding author.

lidia.marassi@unina.it (L. Marassi); narendraprakash.patwardhan@unina.it (N. Patwardhan); stefano.marrone@unina.it (S. Marrone); carlo.sansone@unina.it (C. Sansone)

0009-0006-8134-5466 (L. Marassi); 0000-0002-4807-5664 (N. Patwardhan); 0000-0001-6852-0377 (S. Marrone); 0000-0002-8176-6950 (C. Sansone)

models contribute to information manipulation. We then assess the sociopolitical consequences of such practices and discuss current mitigation strategies, including recent legislative efforts like the European Union's Artificial Intelligence Act.

Finally, we present the PICUS Lab's contribution through the development of *Hominis*, a large language model specifically designed to prioritize verifiability and trustworthy behavior throughout its lifecycle—from data curation to inference. By integrating epistemic safeguards and transparency mechanisms, Hominis serves as a case study in building AI systems resistant to manipulation and aligned with public interest.

2. AI-Generated Deception: Challenges of Misinformation and Disinformation

Artificial intelligence technologies, especially generative models like Generative Adversarial Networks (GANs) and Large Language Models (LLMs), have revolutionized content creation. However, their proliferation has also enabled new forms of deceptive practices, notably misinformation and disinformation, which exploit these tools to erode trust, distort public discourse, and undermine democratic institutions.

2.1. From Automation to Fabrication

Misinformation and disinformation differ in intent: the former involves the unintentional spread of falsehoods, while the latter is the deliberate fabrication and distribution of misleading content [1]. The ability of AI systems to generate persuasive and authentic-seeming text, images, and videos transforms both phenomena from isolated acts into scalable operations.

Deepfake technologies, based on GANs, allow the synthesis of audiovisual material that convincingly impersonates real individuals [2]. Combined with text generation capabilities of LLMs, such as GPT-4, actors can now automate the creation of fraudulent documents, news stories, and social media posts [3]. These models require only minimal input to produce content that aligns stylistically and contextually with real-world language, making them attractive tools for malicious campaigns.

2.2. Sociopolitical Ramifications

AI-generated disinformation undermines the integrity of public information environments. Trust in journalism, science, and governance declines as audiences are increasingly unable to distinguish authentic content from manipulated media [4]. In political contexts, fabricated speeches or videos of public figures have already been used to sway voter sentiment and incite division [5].

The problem extends beyond deliberate manipulation. The saturation of online platforms with misleading content, sometimes referred to as "information pollution," overwhelms users and moderates alike, reducing the visibility and impact of credible information [6]. Disinformation campaigns often exploit platform algorithms to maximize virality, creating feedback loops that amplify harmful narratives.

2.3. Technical and Ethical Tensions

Efforts to detect synthetic content continue to advance, but they often lag behind the capabilities of generative models [7]. Detection systems are challenged not only by technical limitations but also by the dynamic, adversarial nature of generative AI development.

Embedding ethical considerations into AI systems presents further difficulties. While ethical design frameworks exist, they often struggle with ambiguity and context-dependence [8]. Moreover, the environments in which AI operates—such as profit-driven social platforms—may not support or incentivize ethical deployment. Ethical AI cannot be isolated from its sociotechnical setting.

Adding another layer of complexity is the environmental cost of these technologies. Training large-scale generative models is energy-intensive, contributing significantly to carbon emissions [9, 10]. Despite this, environmental sustainability is often overlooked in discussions of AI safety and governance.

2.4. Governance and Mitigation Strategies

The European Union's Artificial Intelligence Act introduces important transparency obligations for generative content, including the labeling of AI-manipulated media [11]. These measures represent a foundational step toward regulating limited-risk AI systems, such as chatbots and deepfakes. Nevertheless, challenges remain. Open-source general-purpose AI models, often exempt from these rules, can be easily repurposed for disinformation without sufficient oversight.

A more comprehensive response requires harmonized regulation across jurisdictions, greater accountability from AI developers and distributors, and robust public media literacy initiatives. Transparency in model development—including dataset documentation, content provenance, and system capabilities—is essential to counter disinformation at scale.

3. Trust Across the Model Lifecycle: From Data Curation to Deployment

Addressing the increasing concern regarding misinformation and disinformation propagated by large language models (LLMs), we have taken a multi-pronged approach to enhance the trustworthiness of *Hominis*, both in its training and inference behavior. Central to this effort is the construction of a training corpus that explicitly prioritizes verifiability. We begin with RedPajama-v2,[12] a widely used and well-deduplicated dataset that reflects a diverse slice of web-scale content. However, recognizing the uneven factual quality inherent in such large-scale crawls, we augment this base with a smaller, carefully curated corpus consisting of scientific documents from arXiv and source code from repositories under permissive licenses (MIT, BSD 3-Clause, Apache 2.0). These sources offer higher epistemic reliability and, although smaller in volume, are given higher effective weight during pretraining by being cycled through multiple epochs. This induces a structural bias in the model's representations toward domains with high factual rigor, such as mathematics, natural sciences, and software engineering [13].

Still, pretraining alone cannot ensure that the model will behave in a trustworthy manner when queried in open-ended or ambiguous settings. Models trained on web data inevitably internalize patterns of both true and false claims. To address this, we operationalize trustworthiness not as a static property but as a dynamic process involving calibrated response generation. In *Hominis-large* (15B), we employ a domain-tuned system prompt that explicitly instructs the model to qualify uncertain statements, flag unverifiable claims, and refrain from asserting information beyond its training distribution. This prompt is not used during distillation into *Hominis-lite* (8B); instead, the distilled model learns these behavioral tendencies implicitly from the teacher's outputs. Furthermore, we apply rejection sampling when constructing the distillation corpus, filtering out responses that exhibit epistemic overreach, unsupported factuality, or stylistic markers common to misinformation. These include filtering out documents with excessive repetition (e.g., repeated tokens or phrases), enforcing symbol-to-word ratio thresholds to eliminate noisy or syntactically degraded text, removing lines without terminal punctuation, discarding documents that are anomalously short or long, and rejecting samples with an abnormally low percentage of alphabetic characters. These heuristics serve as high-precision indicators of data quality and are drawn from best practices in large-scale dataset curation.

In addition to these structural filters, we introduce a novel content-level disinformation check inspired by consensus mechanisms such as those employed by Wikipedia. For queries that fall outside of well-grounded scientific domains, we retrieve the closest matching article from a vetted Wikipedia snapshot and task the larger *Hominis* model with evaluating semantic agreement between the generated response and the retrieved content. This alignment check acts as an additional safeguard against the model producing unsubstantiated or fringe claims, especially in domains prone to politicization or ideological bias.

At inference time, we extend the model's reliability through a structured agentic pipeline. The model does not respond directly to the user query; instead, the input is first processed through a topic classifier to identify the domain of knowledge involved. Based on this classification, the model estimates its own degree of confidence, by utilizing a self-assessment prompt. If the confidence is low or the question falls outside its high-certainty distribution, the model dynamically supplements its response using Retrieval-Augmented Generation (RAG). This retrieval is constrained to a vetted set of always-on sources, such as Google Scholar-accessible publications, or a curated set of open-access documentation. If the user provides external documents, the model integrates that evidence explicitly, and states its dependency on such data in the response. This framework allows *Hominis* to avoid overstating its knowledge, defer to higher-certainty sources when appropriate, and communicate limitations transparently, core attributes for increasing user trust under conditions of epistemic uncertainty.

4. Conclusion

To empirically validate the trustworthiness framework of Hominis, we designed and executed a quantitative benchmark focused on factual recall. The evaluation utilized a dataset of 100 short-answer questions programmatically generated from random Wikipedia summaries using Gemini 1.5 Flash API, establishing a verifiable ground truth. To ensure a fair comparison and test

for epistemic honesty, model responses were constrained using grammar-based sampling, which forced a concise answer format and explicitly permitted the model to state "I don't know." In a direct comparison against established baselines, Llama 2 (7B) and Llama 3.1 (8B), Hominis-lite achieved a factual accuracy score of 80%, representing a significant relative improvement of 25% over the next-best-performing model (Llama 3.1) and 87% over Llama 2. Critically, the nature of the incorrect responses revealed a significant divergence in model behavior. In all cases where Hominis did not provide the correct fact, it correctly defaulted to the "I don't know" response, successfully avoiding hallucination. In contrast, Llama 3.1 and Llama 2 generated factually incorrect answers 5% and 20% of the time, respectively, choosing to confabulate rather than admit uncertainty. These results provide strong quantitative and qualitative evidence that our multi-pronged approach—combining a high-rigor corpus, targeted distillation, and content verification—translates into measurably more trustworthy and well-calibrated model behavior.

Declaration on Generative AI

During the preparation of this work, the authors used GPT and DeepL to perform grammar and spelling check. After using these tools and services, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] C. Wardle, H. Derakhshan, Information disorder: Toward an interdisciplinary framework for research and policymaking, 2017. Council of Europe report.

[2] R. Chesney, D. K. Citron, Deepfakes and the new disinformation war, Foreign Affairs (2019).

[3] R. Zellers, A. Holtzman, Y. Bisk, A. Farhadi, Y. Choi, Defending against neural fake news, in: Advances in Neural Information Processing Systems (NeurIPS), 2019.

[4] C. Vaccari, A. Chadwick, Deepfakes and disinformation: Exploring the impact of synthetic political video on deception, uncertainty, and trust in news, Social Media + Society 6 (2020) 1–13.

[5] J. S. Brennen, F. M. Simon, P. N. Howard, R. K. Nielsen, Types, sources, and claims of covid-19 misinformation, 2020.

[6] N. Helberger, J. Pierson, T. Poell, Governing online platforms: From contested to cooperative responsibility, The Information Society 36 (2020) 1–14.

[7] L. Verdoliva, Media forensics and deepfakes: An overview, IEEE Journal of Selected Topics in Signal Processing 14 (2020) 910–932.

[8] B. D. Mittelstadt, P. Allo, M. Taddeo, S. Wachter, L. Floridi, The ethics of algorithms: Mapping the debate, Big Data & Society 3 (2016) 1–21.

[9] E. Strubell, A. Ganesh, A. McCallum, Energy and policy considerations for deep learning in nlp, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, ACL, 2019, pp. 3645–3650.

[10] P. Dhar, The carbon impact of artificial intelligence, Nature Machine Intelligence 2 (2020) 423–425.

[11] C. of the European Union, Proposal for a regulation of the european parliament and of the council laying down harmonised rules on artificial intelligence (artificial intelligence act), 2024. EU Open Data Portal.
[12] M. Weber, D. Fu, Q. Anthony, Y. Oren, S. Adams, A. Alexandrov, X. Lyu, H. Nguyen, X. Yao, V. Adams, et al., Redpajama: an open dataset for training large language models, Advances in neural information processing systems 37 (2024) 116462–116492.
[13] R. Taylor, M. Kardas, G. Cucurull, T. Scialom, A. Hartshorn, E. Saravia, A. Poulton, V. Kerkez, R. Stojnic, Galactica: A large language model for science, arXiv preprint arXiv:2211.09085 (2022).