Quantitative and Organizational Approaches to Epistemic Risk in Generative and General-Purpose AI

Robert Wolfe

University of Washington
rwolfe3@uw.edu

Abstract

This extended abstract discusses the three approaches to epistemic risk in general-purpose AI that characterize my research: measurement, design, and transparency.

Introduction

General-purpose AI systems like the now-ubiquitous ChatGPT stand poised to reshape longstanding information infrastructures and professions, ranging from search to social media to online journalism (Eloundou et al. 2023; Memon and West 2024). Yet questions surrounding subtle biases, misinforming output, and data provenance – *epistemic risks* related to the way knowledge is encoded and disseminated – have followed these technologies since their inception (Bender et al. 2021). My research develops methods to precisely measure epistemic risks in general-purpose AI and envisions ways to deploy these systems in the presence of epistemic risk. To solve these problems, I develop techniques that center the epistemic needs both of society and of the organizations deploying such models. These include: 1) computational techniques for identifying the manifestations of epistemic risks like bias and misinformation and their underlying causes; 2) context-aligned techniques and designs that allow for the responsible deployment of epistemically flawed but nonetheless useful technologies; and 3) alternative approaches to closed, proprietary systems where privacy and transparency are valued as highly as raw performance.

Measuring Epistemic Risk

Managing the risk of bugs or downtime in traditional software systems meant writing tests to ensure that deployed software fulfilled its intended purpose and met standards of reliability expected by consumers. Yet the capacity of general-purpose models to take nearly any input and produce a plausible output presents a problem that is as much social as technical, as AI may replicate subtle, even unconscious biases of the human mind (Caliskan et al. 2022). In my most recent work, I used a dataset developed by psychologists (Peterson et al. 2022) to study a suite of 43 pretrained multimodal CLIP models (classifiers that learn to match images with text based on the cosine similarity between them (Radford et al. 2021)), with systematically varying training regimes and parameter counts (Cherti et al. 2022). I found that two variables predominantly accounted for the traits learned by the models: the extent to which the inference was consistent across society (based on human inter-rater reliability), and the size of the dataset on which the model trained. This introduces a catch-22 for training general-purpose AI: the same variables that produce models that are more accurate (scale) and preferred by more people (societal consistency) also lead to novel biases, with consequences for downstream applications.

This study, now accepted at AIES 2024 (Wolfe et al. 2024a), continues a line of work in which I have applied and developed psychologically grounded tests for general-purpose AI systems (Wolfe and Caliskan 2022d,b; Wolfe, Hiniker, and Howe 2024; Wolfe et al. 2024b), especially multimodal models like CLIP and Stable Diffusion (Wolfe, Banaji, and Caliskan 2022; Wolfe and Caliskan 2022a; Wolfe et al. 2023). My research on these models has studied consequential topics ranging from biased defaults in general-purpose classification (Wolfe and Caliskan 2022c) to problems of safety when generating images of women (Wolfe et al. 2023), highlighting unanticipated epistemic problems in novel AI architectures.

Designing for Responsible Deployment

Flaws that prevent general-purpose AI from reaching its full potential do not necessarily prevent its responsible use. While developing effective technical solutions to epistemic issues remains an important direction, and one I have explored in my collaborations (Yang et al. 2024), an equally important direction explores how organizations can responsibly deploy general-purpose AI *in the presence of epistemic risk*. To that end, I conducted an interview study with fact-checking organizations, for whom success depends on navigating epistemic risk both within the organization and in society.

The interviews surfaced tensions between the need to provide audiences with an efficient response to misinforming content circulating online and the need to carefully verify all published fact-checking content. Fact-checking organizations necessarily embrace task-specific AI and machine learning solutions to sift through enormous quantities of on-

line content (Juneja and Mitra 2022), and most expressed openness to adopting general-purpose AI not only for processing data internally but also for user-facing applications, like collecting tips and delivering fact-checking content via novel interfaces like conversational tiplines. Yet most organizations were also acutely aware of the risks of a mistake, expressing concern that an AI-generated mistake could cast doubt on their organization's commitment to its expressed values. In light of such risks, how could fact-checking organizations move forward with deploying generative AI?

The answer is an approach I call **Designing for Verification** (Wolfe and Mitra 2024a). Interviewees consistently described a process of information *production* followed by *verification* to ensure the epistemic integrity of their organization's content. Depending on the context, either a human fact-checker or a generative model could play the role of the Producer and/or the role of the Verifier. Though human fact-checkers did verify machine-generated content, the more appealing use of generative models for many organizations was in quality assurance – fact-checkers valued adversarial analysis of their own content to *address epistemic errors* and uphold their values.

Maximizing Transparency

Many organizations are concerned about submitting private materials through an external API, which could capture data and use it for reasons they did not intend. Open models like Mistral (Jiang et al. 2023) or Meta's LLaMA series (Touvron et al. 2023) seem to offer an alternative by allowing companies to download and run models either locally or on a private cloud instance. But how competitive are open models with proprietary models served via corporate APIs? I addressed these questions in a study that found that open models like LLaMA-2 lag behind proprietary models like ChatGPT out of the box, but can match the task-specific performance of proprietary models with only a small amount of supervised fine-tuning (Wolfe et al. 2024c).

To further understand the tradeoffs of using open vs. proprietary models, I conducted a study with 24 fact-checking organizations (Wolfe and Mitra 2024b), asking about use of open vs. proprietary models. The study revealed that organizations can't be neatly divided those that use proprietary models and those that use open models. Open models were preferred when dealing with sensitive data, or when performance needed to be strong for specific tasks, as open models can be extensively customized. Proprietary models were preferred for user-facing applications, as they offer reliable out-of-the-box usability and fairness mitigations viewed as safer for audiences.

Future Work

I hope to study the epistemic risks of ecosystems of AI models like those in OpenAI's GPT Store, which I show in a recent workshop paper can exhibit deceptive design patterns (Wolfe and Hiniker 2024). I also hope to continue exploring the role of general-purpose AI as a mediator of culture (Dangol et al. 2024), including the potential for frontier multimodal models to reproduce biases and amplify misinforming content.

References

Bender, E. M.; Gebru, T.; McMillan-Major, A.; and Shmitchell, S. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*.

Caliskan, A.; Ajay, P. P.; Charlesworth, T. E. S.; Wolfe, R.; and Banaji, M. R. 2022. Gender Bias in Word Embeddings: A Comprehensive Analysis of Frequency, Syntax, and Semantics. *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*.

Cherti, M.; Beaumont, R.; Wightman, R.; Wortsman, M.; Ilharco, G.; Gordon, C.; Schuhmann, C.; Schmidt, L.; and Jitsev, J. 2022. Reproducible scaling laws for contrastive language-image learning. *arXiv preprint arXiv:2212.07143*.

Dangol, A.; Newman, M.; Wolfe, R.; Lee, J. H.; Kientz, J. A.; Yip, J.; and Pitt, C. 2024. Mediating Culture: Cultivating Socio-cultural Understanding of AI in Children through Participatory Design. In *Proceedings of the 2024 ACM Designing Interactive Systems Conference*, 1805–1822.

Eloundou, T.; Manning, S.; Mishkin, P.; and Rock, D. 2023. Gpts are gpts: An early look at the labor market impact potential of large language models. *arXiv preprint arXiv:2303.10130*.

Jiang, A. Q.; Sablayrolles, A.; Mensch, A.; Bamford, C.; Chaplot, D. S.; Casas, D. d. l.; Bressand, F.; Lengyel, G.; Lample, G.; Saulnier, L.; et al. 2023. Mistral 7B. *arXiv preprint arXiv:2310.06825*.

Juneja, P.; and Mitra, T. 2022. Human and technological infrastructures of fact-checking. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW2): 1–36.

Memon, S. A.; and West, J. D. 2024. Search engines post-ChatGPT: How generative artificial intelligence could make search less reliable. *arXiv preprint arXiv:2402.11707*.

Peterson, J. C.; Uddenberg, S.; Griffiths, T. L.; Todorov, A.; and Suchow, J. W. 2022. Deep models of superficial face judgments. *Proceedings of the National Academy of Sciences*, 119(17): e2115228119.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*.

Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Wolfe, R.; Banaji, M. R.; and Caliskan, A. 2022. Evidence for hypodescent in visual semantic AI. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 1293–1304.

Wolfe, R.; and Caliskan, A. 2022a. American== white in multimodal language-and-image ai. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, 800–812.

Wolfe, R.; and Caliskan, A. 2022b. Detecting emerging associations and behaviors with regional and diachronic word embeddings. In *2022 IEEE 16th International Conference on Semantic Computing (ICSC)*, 91–98. IEEE.

Wolfe, R.; and Caliskan, A. 2022c. Markedness in visual semantic AI. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 1269–1279.

Wolfe, R.; and Caliskan, A. 2022d. Vast: The valence-assessing semantics test for contextualizing language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 11477–11485.

Wolfe, R.; Dangol, A.; Hiniker, A.; and Howe, B. 2024a. Dataset Scale and Societal Consistency Mediate Facial Impression Bias in Vision-Language AI. In *Proceedings of the 2024 AAAI/ACM Conference on AI, Ethics, and Society*.

Wolfe, R.; Dangol, A.; Howe, B.; and Hiniker, A. 2024b. Representation Bias of Adolescents in AI: A Bilingual, Bicultural Study. In *Proceedings of the 2024 AAAI/ACM Conference on AI, Ethics, and Society*.

Wolfe, R.; and Hiniker, A. 2024. Expertise Fog on the GPT Store: Deceptive Design Patterns in User-Facing Generative AI.

Wolfe, R.; Hiniker, A.; and Howe, B. 2024. ML-EAT: A Multilevel Embedding Association Test for Interpretable and Transparent Social Science. In *Proceedings of the 2024 AAAI/ACM Conference on AI, Ethics, and Society*.

Wolfe, R.; and Mitra, T. 2024a. The Impact and Opportunities of Generative AI in Fact-Checking. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, 1531–1543.

Wolfe, R.; and Mitra, T. 2024b. The Implications of Open Generative Models in Human-Centered Data Science Work: A Case Study with Fact-Checking Organizations. In *Proceedings of the 2024 AAAI/ACM Conference on AI, Ethics, and Society*.

Wolfe, R.; Slaughter, I.; Han, B.; Wen, B.; Yang, Y.; Rosenblatt, L.; Herman, B.; Brown, E.; Qu, Z.; Weber, N.; et al. 2024c. Laboratory-Scale AI: Open-Weight Models are Competitive with ChatGPT Even in Low-Resource Settings. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, 1199–1210.

Wolfe, R.; Yang, Y.; Howe, B.; and Caliskan, A. 2023. Contrastive language-vision ai models pretrained on web-scraped multimodal data exhibit sexual objectification bias. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, 1174–1185.

Yang, Y.; Liu, A. Z.; Wolfe, R.; Caliskan, A.; and Howe, B. 2024. Label-Efficient Group Robustness via Out-of-Distribution Concept Curation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12426–12434.