Elinor Poole-Dayan

My research is centered around the societal considerations of large language models (LLMs), focusing on **equity, epistemology, and pluralistic alignment** with an **emphasis on evaluating real-world impact**. I envision LLMs as tools that can democratize information access and represent the voices of historically underheard groups. However, much of AI research evaluates these systems in artificial settings that bear *little connection* to real use or human impact. I'm driven by a commitment to bridging that gap—ensuring AI systems are **evaluated in realistic contexts** and ultimately **develop more equitable, pluralistic models** for the benefit of *all*.

This vision has been shaped by several of my past works: investigating disparities in model behavior across global user groups (**AAAI** [1]), analyzing unintended side effects of model alignment (**EMNLP** [2]) and debiasing (**ACL** [3]), and operationalizing pluralism into measurable criteria for model outputs (**ICLR** in submission [4]). Building upon this foundation, my PhD agenda centers on three directions: (1) developing **realistic evaluation frameworks** to ensure **LLMs benefit people equitably**, (2) studying the unintended **epistemic and societal consequences of alignment** techniques, and (3) **advancing pluralistic alignment methods** that meaningfully navigate value conflict trade-offs.

Real-world LLM performance & impact. While frontier labs claim to "put expert-level intelligence in everyone's hands" (GPT-5 model release tagline, 2025), LLMs are evaluated on benchmarks designed to measure "the best performance on a certain task it can achieve, given the best currently available capability elicitation techniques" [5, 6]. These evaluations do not reflect model behavior across millions of worldwide users [7, 8], missing how deviations from idealized performance impact users in practice. **My first-author work [1, AAAI'26]** provides evidence of this: LLM accuracy and truthfulness degrade for users with less education, revealing performance inequities invisible to standard benchmarks. These findings mirror realistic interactions where personalized chatbots adapt to user attributes in ways that can unintentionally amplify these disparities—a mechanism I term **targeted underperformance**. I hope to develop evaluation techniques that capture how personalized model behavior affects users across contexts. Beyond this, I am interested in developing methods to mitigate these harms, for example by training models to better decide *when* and *how* to leverage personal user information to tailor its responses in equitable and transparent ways.

Epistemic integrity and (mis)alignment. As LLMs become central to how information is produced and consumed, my findings on targeted underperformance highlight their potential to exacerbate epistemic inequity worldwide. Together, these experiences have shaped my interest in how alignment signals—whether framed as debiasing, truthfulness, or helpfulness—carry implicit normative assumptions that reshape model behavior. In my very first research project on debiasing language models **[3, ACL'22]**, we found that some gender debiasing techniques, while effective at mitigation, degraded general language modeling abilities. In more recent work, I investigated a similar phenomenon and surprisingly found that alignment finetuning on truthfulness datasets systematically shifted LLMs' political biases **[2, EMNLP'24]**. Across both works, I learned to critically question the *unintended* normative and epistemic consequences of alignment strategies while developing technical skills in debiasing and finetuning language models on compute clusters. Moreover, RLHF amplifies sycophantic behaviors—models deferring to user beliefs regardless of factuality to maintain agreeableness—directly undermining knowledge integrity [9]. In my PhD, I would like to study how alignment methods alter model behaviors in undesirable ways and resulting downstream impacts. For example, what are the effects of personalization and sycophancy on model factuality and user trust? How do these interplay with epistemic homogenization effects on human beliefs and values at scale [10, 11, 12]? More broadly, how should alignment methods navigate

situations where user preferences, model consistency, and epistemic norms come into tension?

Pluralistic alignment. I am motivated by the belief that models should be *pluralistic*—able to represent diverse perspectives faithfully and reason across them, especially when those perspectives conflict [13, 14, 15]. This motivation shaped my most recent **first-author work [4, ICLR'26 submission]**, where I (a) operationalize the notion of Overton pluralism into a metric and (b) assess the extent to which diverse viewpoints are represented in model outputs of 8 state-of-the-art LLMs. Throughout this project, I strengthened my skills in designing and deploying a large scale human study on Prolific from end-to-end and mentored an undergraduate researcher, who presented our results at the **NeurIPS'25 LLM Evaluation Workshop**—a rewarding experience that expanded my supervision, technical communication, and project management abilities. One key research finding was the clear trade-off between politically neutral and pluralistic model responses, raising important questions about whether "neutrality" is a desirable alignment goal, corroborating findings from Fisher et al. [16]. Going forward, I hope to explore such trade-offs more deeply. For example, how does optimizing for pluralism impact model behaviors such as verbosity or hedging? More broadly, how should models navigate simultaneously aligning to individual preferences as well as the broader set of alignment principles—especially when these contradict one another? How can we collect nuanced contextual human preferences at scale, and what data is most useful for models to effectively learn from to improve pluralistic alignment?

Career goals. Throughout these experiences, I have become certain that a PhD is the only way to meaningfully pursue the research directions I have outlined. The problems I care about are messy, interdisciplinary, and often fundamentally at odds with industry incentive structures, and they require the intellectual freedom that only a doctoral program can provide. I actively sought out such opportunities during my Master's because I deeply value this type of work [17, 18, 19] and hope to continue collaborating across diverse backgrounds during my PhD. A PhD at [University] provides the structure, resources, community, and time to think deeply, develop methodological expertise, and make tangible contributions towards the problems that genuinely matter for the benefit of humanity.

Beyond the PhD, I aim to become a professor at a research institution. Teaching and mentoring have always been central to my identity and strengths as an academic, and I have greatly enjoyed several TA-ships as well as mentoring two junior researchers so far. I am particularly passionate about inclusive STEM pedagogy, and pursued MIT's Kaufman Teaching Certificate where I implemented evidence-based teaching techniques, developed a novel syllabus on Generative AI Ethics, gained hands-on teaching experience, and critically engaged with responsibly integrating AI while maintaining pedagogical integrity. As a faculty, I look forward to giving back to the academic community to foster pathways for underrepresented students to expand the diversity in perspectives in research–especially valuable for moving towards more equitable and socially-conscious AI research.

Why [University]? [University]'s core values of ... are strongly aligned with my research interests and goals. I would be excited to work with ...

[University] offers the ideal environment required for this work, and I am excited to contribute to—and grow within—its research efforts in responsible AI.

References

[1] **Elinor Poole-Dayan**, Deb Roy, and Jad Kabbara. LLM Targeted Underperformance Disproportionately Impacts Vulnerable Users. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Singapore, January 2026. URL https://arxiv.org/abs/2406.17737.

[2] Suyash Fulay, William Brannon, Shrestha Mohanty, Cassandra Overney, **Elinor Poole-Dayan**, Deb Roy, and Jad Kabbara. On the Relationship between Truth and Political Bias in Language Models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 9004–9018, Miami, Florida, USA, November 2024. doi: 10.18653/v1/2024.emnlp-main.508. URL https://aclanthology.org/2024.emnlp-main.508.

[3] Nicholas Meade, **Elinor Poole-Dayan**, and Siva Reddy. An Empirical Survey of the Effectiveness of Debiasing Techniques for Pre-trained Language Models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1878–1898, Dublin, Ireland, May 2022. doi: 10.18653/v1/2022.acl-long.132. URL https://aclanthology.org/2022.acl-long.132.

[4] **Elinor Poole-Dayan**, Jiayi Wu, Taylor Sorensen, Jiaxin Pei, and Michiel A. Bakker. Benchmarking Overton Pluralism in LLMs. Under Review at ICLR 2026, September 2025. URL https://arxiv.org/abs/2512.01351.

[5] Teun van der Weij, Felix Hofstätter, Oliver Jaffe, Samuel F. Brown, and Francis Rhys Ward. AI sandbagging: Language models can strategically underperform on evaluations. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=7Qa2SpjxIS.

[6] Rohin Shah, Vikrant Varma, Ramana Kumar, Mary Phuong, Victoria Krakovna, Jonathan Uesato, and Zac Kenton. Goal misgeneralization: Why correct specifications aren't enough for correct goals, 2022. URL https://arxiv.org/abs/2210.01790.

[7] Tyna Eloundou, Alex Beutel, David G. Robinson, Keren Gu-Lemberg, Anna-Luisa Brakman, Pamela Mishkin, Meghan Shah, Johannes Heidecke, Lilian Weng, and Adam Tauman Kalai. First-person fairness in chatbots, 2025. URL https://arxiv.org/abs/2410.19803.

[8] Angelina Wang, Daniel E. Ho, and Sanmi Koyejo. The Inadequacy of Offline LLM Evaluations: A Need to Account for Personalization in Model Behavior, September 2025. URL http://arxiv.org/abs/2509.19364. arXiv:2509.19364 [cs].

[9] Lars Malmqvist. Sycophancy in Large Language Models: Causes and Mitigations. In Kohei Arai, editor, *Intelligent Computing*, pages 61–74, Cham, 2025. Springer Nature Switzerland. ISBN 978-3-031-92611-2. doi: 10.1007/978-3-031-92611-2_5.

[10] Andrew J. Peterson. Ai and the problem of knowledge collapse. *AI & SOCIETY*, 40(5): 3249–3269, January 2025. ISSN 1435-5655. doi: 10.1007/s00146-024-02173-x. URL http://dx.doi.org/10.1007/s00146-024-02173-x.

[11] Tianyi Qiu, Zhonghao He, Tejasveer Chugh, and Max Kleiman-Weiner. The lock-in hypothesis: Stagnation by algorithm. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=mE1M626qOo.

[12] Liwei Jiang, Yuanjun Chai, Margaret Li, Mickel Liu, Raymond Fok, Nouha Dziri, Yulia Tsvetkov, Maarten Sap, and Yejin Choi. Artificial hivemind: The open-ended homogeneity of language models (and beyond). In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2025. URL https://openreview.net/forum?id=saDOrrnNTz.

[13] Taylor Sorensen, Jared Moore, Jillian Fisher, Mitchell Gordon, Niloofar Mireshghallah, Christopher Michael Rytting, Andre Ye, Liwei Jiang, Ximing Lu, Nouha Dziri, Tim Althoff, and Yejin Choi. Position: a roadmap to pluralistic alignment. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *ICML'24*, pages 46280–46302, Vienna, Austria, July 2024. JMLR.org.

[14] Taylor Sorensen, Liwei Jiang, Jena D. Hwang, Sydney Levine, Valentina Pyatkin, Peter West, Nouha Dziri, Ximing Lu, Kavel Rao, Chandra Bhagavatula, Maarten Sap, John Tasioulas, and Yejin Choi. Value Kaleidoscope: Engaging AI with Pluralistic Human Values, Rights, and Duties. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(18):19937–19947, March 2024. doi: 10.1609/aaai.v38i18.29970. URL https://ojs.aaai.org/index.php/AAAI/article/view/29970. Section: AAAI Technical Track on Philosophy and Ethics of AI.

[15] Michiel Bakker, Martin Chadwick, Hannah Sheahan, Michael Tessler, Lucy Campbell-Gillingham, Jan Balaguer, Nat McAleese, Amelia Glaese, John Aslanides, Matt Botvinick, and Christopher Summerfield. Fine-tuning language models to find agreement among humans with diverse preferences. *Advances in Neural Information Processing Systems*, 35:38176–38189, December 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/f978c8f3b5f399cae464e85f72e28503-Abstract-Conference.html.

[16] Jillian Fisher, Ruth Elisabeth Appel, Chan Young Park, Yujin Potter, Liwei Jiang, Taylor Sorensen, Shangbin Feng, Yulia Tsvetkov, Margaret Roberts, Jennifer Pan, Dawn Song, and Yejin Choi. Position: Political Neutrality in AI Is Impossible — But Here Is How to Approximate It. In *Forty-second International Conference on Machine Learning Position Paper Track*, 2025. URL https://openreview.net/forum?id=H72JEXAPwo.

[17] Margaret Hughes, Brandon Roy, **Elinor Poole-Dayan**, Deb Roy, and Jad Kabbara. Computational analysis of conversation dynamics through participant responsivity. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 35500–35519, Suzhou, China, November 2025. ISBN 979-8-89176-332-6. URL https://aclanthology.org/2025.emnlp-main.1798/.

[18] **Elinor Poole-Dayan**, Daniel T Kessler, Hannah Chiou, Margaret Hughes, Emily S Lin, Marshall Ganz, and Deb Roy. Applying large language models to characterize public narratives, 2025. URL https://arxiv.org/abs/2511.13505.

[19] **Elinor Poole-Dayan**, Deb Roy, and Jad Kabbara. An AI-Powered Framework for Analyzing Collective Idea Evolution in Deliberative Assemblies. ArXiv, August 2025. URL https://arxiv.org/abs/2509.12577.