# Elinor Poole-Dayan

Cambridge, MA  ✉ elinorpd@mit.edu  ☎ (718) 884-7975  elinor-poole-dayan  elinorp-d.github.io

## Education

**Master's of Science,** *Massachusetts Institute of Technology - Media Lab* 09/2023 | Cambridge, United States

- Master's candidate in the Center for Constructive Communication with Professor Deb Roy (anticipated graduation 05/2025).
- Thesis focusing on leveraging LLMs to understand and augment deliberative dialogue spaces, finding consensus opportunities, mitigating biases, and fairly representing underheard voices.

**Bachelor's in Honours Math and Computer Science,** *McGill University* 2019 – 2023 | Montreal, Canada

- GPA: 3.9/4.0, Awards: Dean's Honour List, J W McConnell Scholarship, Canadian Graduate Scholarship - Master's (NSERC) $17,500, Mila Excellence Scholarship - EDI in Research $5,000

## Research & Publications

**LLM Targeted Underperformance Disproportionately Impacts Vulnerable Users,** *NeurIPS Workshop on Safe GenAI* 09/2024

- Measured how LLM response quality changes in terms of information accuracy, truthfulness, and refusals across users.
- Find systematic underperformance for users with lower English proficiency, less education, and from non-US origins.

**On the Relationship between Truth and Political Bias in Language Models,** *Accepted to EMNLP 2024* 06/2024

- Examined how aligning LLMs to be truthful impacts political biases by optimizing reward models for truthfulness and find a left-leaning political bias.

**Interplay Between Implicit Bias and Sycophancy in LLMs: Implications for Fairness in Educational Decisions** 05/2024

- Evaluated the impact of implicit bias on sycophantic behavior in LLMs in educational decision outcomes.
- Found notable differences in model judgements reflecting harmful racial stereotypes exacerbated by sycophantic tendencies.

**Are Diffusion Models Vision-And-Language Reasoners?,** *Accepted to NeurIPS 2023* 05/2023

- Transformed diffusion models for any image-text matching (ITM) task using a novel method called DiffusionITM.
- Developed the Generative-Discriminative Evaluation Benchmark (GDBench) benchmark with 7 complex vision-and-language tasks, bias evaluation and detailed analysis.

**An Empirical Survey of the Effectiveness of Debiasing Techniques for Pre-trained Language Models,** *Accepted to ACL 2022* 05/2022

- Investigated state-of-the-art bias evaluation metrics, benchmarks, and mitigation techniques while measuring their impact on a model's language modeling ability and performance on downstream NLU tasks.

## Work Experience

**Research Assistant,** *Center for Constructive Communication, MIT Media Lab* 09/2023 – present | Cambridge, United States

**Data Science Intern,** *Unity Technologies* 05/2022 – 08/2022 | Montreal, Canada

- Optimized deep learning algorithms to throttle bid requests in Unity's Ad Exchange using Tensorflow.
- Decreased model training time by 25% and reduced model size and number of parameters by 50%.
- Created a text data preprocessing pipeline on Google Cloud Platform Dataflow using Apache beam.

**Undergraduate NLP Researcher,** *McGill University / Mila Quebec* 01/2021 – 05/2021 | Montreal, Canada

- Investigated the effect of gender debiasing on fine-tuned language models such as BERT using PyTorch.
- Explored debiasing methods and reformulated bias metrics for racial and religious biases.

## Service

**Reviewer for ACL Rolling Review (December 2024)**

**Emergency Reviewer for ACL Rolling Review (October 2024)**

## Skills & Interests

**Interests**
*NLP fairness & evaluation. Pluralistic alignment. Equitable, safe, ethical AI.*

**Machine Learning**
*TensorFlow, PyTorch, Keras, scikit-learn, pandas, NumPy*