# Elinor Poole-Dayan

Cambridge, MA   elinorpd@mit.edu   elinor-poole-dayan   elinorp-d.github.io

## Research Interests

**Motivation:** Much of AI fairness research happens in artificial settings, with limited connection to how people actually use these systems or how they impact different groups in practice. I'm driven by a commitment to closing that gap—ensuring AI systems are evaluated in realistic contexts and truly benefit people equitably.

**My work focuses on:** fairness, safety, and pluralistic alignment in large language models, with an emphasis on real-world impact. I'm also interested in how LLMs can be integrated ethically and equitably in domains like research, education, and democratic participation—especially as tools for qualitative insight and human-centered decision-making. │ I bring a strong mathematical foundation and substantial expertise in both computational and qualitative research methods.

## Education

**Massachusetts Institute of Technology - Media Lab,** *Master's of Science*    2023 – 2025 │ Cambridge, United States
- Advised by Deb Roy in the MIT Center for Constructive Communication. GPA 5.0/5.0
- Thesis: **From Dialogue to Decision: An LLM-Powered Framework for Analyzing Collective Idea Evolution and Voting Dynamics in Deliberative Assemblies** (Grade: A+)

**McGill University,** *Bachelor's in Honours Math and Computer Science*    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

**Tracing Idea Evolution in Democratic Deliberation with LLMs,**    06/2025
*Oral presentation at COLM '25 NLP4Democracy Workshop* ⌞
- We use LLMs to analyze transcripts from an in-person deliberative assembly, revealing how ideas evolve into policy recommendations and uncovering both effective filtering and overlooked suggestions often invisible in final outcomes.

**LLM Targeted Underperformance Disproportionately Impacts Vulnerable Users,**    05/2025
*Under review AAAI '26; NeurIPS '24 Safe GenAI Workshop* ⌞
- Measured how LLM response quality changes in terms of information accuracy, truthfulness, and refusals across users.
- Found systematic underperformance for users with lower English proficiency, less education, and from non-US origins.

**Computational Analysis of Conversation Dynamics through Participant Responsivity,** *Accepted to EMNLP '25* ⌞    05/2025
- Engineered an LLM pipeline to annotate a large conversational dataset and operationalize a novel set of metrics for understanding constructive communcation.

**Applying Large-Language Models to Characterize Public Narratives,**    05/2025
*NAACL '25 Workshop on Narrative Understanding* ⌞
- Developed a novel LLM-based framework for automating the annotation of public narratives, achieving near-expert performance and enabling scalable analysis of civic storytelling and political rhetoric.

**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?,**    05/2023
*Accepted to NeurIPS 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** 05/2022
**Pre-trained Language Models,** *Accepted to ACL 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

**Predoctoral Researcher,** *MIT* 07/2025 – Present | Cambridge, USA
- Research on pluralistic alignment in LLMs and evaluating the fairness of AI's societal impacts; supervised by Michiel Bakker.

**Research Assistant,** *Center for Constructive Communication, MIT Media Lab* 09/2023 – 06/2025 | Cambridge, USA

**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.

**NLP Research Intern,** *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.
- Supervised by Prof. Siva Reddy.

**Undergraduate NLP Researcher,** *McGill University* 01/2022 – 05/2022 | Montreal, Canada
- Identified the geo-indicativeness of text using BERT applied to geosocial datasets to build a safety tool for social media.
- Supervised by Prof. Grant McKenzie.

**NLP Research Intern,** *Shamoon College of Engineering* 06/2021 – 08/2021 | Be'er Sheva, Israel
- Classified author gender of books to perform a case study on female authors who wrote under male pseudonyms.
- Preprocessed data using CoreNLP and scikit-learn. Designed and implemented baseline experiments using SVMs.
- Supervised by Dr. Irina Rabaev and Dr. Marina Litvak.

## Teaching Experience

**Kaufman Teaching Certificate,** *MIT Teaching + Learning Lab* ⧉ 02/2025 – 05/2025
- Participated in eight practice-based workshops, evaluated on my teaching skills through 2 microteaching sessions, received individual feedback from peers and teaching experts, and implemented evidence-based teaching techniques grounded in the scholarship of teaching and learning.
- Developed a syllabus for a course titled *Ethics, Fairness, and Bias in Generative Language Models.*

**Teaching Assistant: Intro to Media Arts & Sciences,** *MIT Media Lab* 09/2025 – 12/2025

**Teaching Assistant: Honours Algorithms & Data Structures,** 01/2022 – 05/2022
*McGill University*

## Mentorship

**Jiayi Wu, Brown University (Summer 2025)** **Hannah Chiou, Wellesley University (Spring 2025)**
Benchmarking Overton Pluralism with LLM-as-a-Judge Computational Analysis of Public Narratives with LLMs

## Service

**Reviewer for ACL Rolling Review**
- May 2025
- December 2024
- October 2024 (Emergency Reviewer)

**Reviewer for AAAI**
- Social Impact Track 2026

## Skills & Interests

**Programming Languages**
*Python, Java, Javascript, C, Unix/Linux, OCaml, SQL*

**Machine Learning & Data Science**
*TensorFlow, PyTorch, Keras, scikit-learn, pandas, NumPy, matplotlib, seaborn, plotly*

**Cloud Computing**
*Google Cloud Platform, Amazon Web Services, Docker*