

Erik Miebling

mail: emiebling@gmail.com

web: emiebling.github.io

WORK EXPERIENCE

- **IBM Research** *Dublin, Ireland*
Staff Research Scientist **June 2025 – present**
 - Leading safety research on governance of multi-agent systems; scalable oversight of generative models.
 - Project lead on **AI Steerability 360**, an open source toolkit for evaluating the steerability of LLMs, enabling development of novel steering methods and the analysis of steering side effects on model behavior.
Research Scientist **Aug 2022 – May 2025**
 - Developed conversational maxims for measuring degradation in human-AI conversations (published at EMNLP 2024); deployed as risk dimensions in **Granite Guardian**
 - Carried out algorithmic fairness research under the Horizon Europe project **AutoFair**; studied the (disparate) impact of informational consent decisions on recommendation accuracy (published at NeurIPS 2023; featured in BBC Science Focus).
- **University of Illinois at Urbana-Champaign** *Urbana, IL*
Postdoctoral Research Associate **Feb 2018 - Aug 2022**
 - Advised 10 Ph.D. students and published 15 peer-reviewed articles in multi-agent reinforcement learning, stochastic control, and machine learning spanning five federally funded projects (total research funding: \$39 500 000 USD).
 - Made foundational contributions to multi-agent reinforcement learning in both cooperative domains (RNN-based information embeddings) and adversarial domains (online attacker intent inference).
 - Co-wrote an NSF grant (\$500 000 USD) funding research on learning and control of epidemic processes.

EDUCATION

- **University of Michigan** *Ann Arbor, MI*
Ph.D. – Electrical Engineering & Computer Science **Sept 2011 – Dec 2017**
- **University of British Columbia** *Vancouver, Canada*
M.A.Sc. – Electrical & Computer Engineering **Sept 2009 – Aug 2011**
B.A.Sc. – Electrical Engineering **Sept 2006 – May 2009**

SELECTED COMMUNITY ENGAGEMENTS

- Chair, AAAI Workshop: **Foundations of Agentic Systems Theory**, 2026
- Organizer, AAAI Lab: **Learning to Steer Large Language Models**, 2026
- Speaker, **Localizing Persona Representations in LLMs**, INTERPLAY workshop @ COLM, 2025
- Guest lecturer, **A (Brief) Introduction to LLMs**, MS Applied Data Scientist program @ UChicago, 2024
- Tutorial, **Transformer-based Models & Applications**, Deep Learning for NLP @ Dublin City Univ., 2024

SKILLS

Programming: Python (torch, transformers)

Theory: probability & statistics, machine learning, RL, optimization & control, game theory