ROMAN BELAIRE

Ph.D. Candidate - Computer Science (2025)

Singapore / USA | romanbelaire@gmail.com | romanbelaire.com | /romanbelaire | /romanbelaire

Research Interests: Robust Reinforcement Learning, Adversarial Robustness, Generalization, LLM Red-Teaming.

EDUCATION

Singapore Management University

Ph.D. in Computer Science

June 2025 (est. Defense Schedule)

- Presidential Doctoral Fellowship (2024)
- Research expert in Adversarially Robust Reinforcement Learning via regret and POMDP belief constructs (see publications).
- Automated Red-teaming of LLMs with Adversarial MDPs and safety dataset generation with the Infocomm Media Development Authority, Singapore

PUBLICATIONS

- *Regret-based Defense in Adversarial Reinforcement Learning*, **Roman Belaire**, Thanh Nguyen, David Lo, and Pradeep Varakantham. arxiv.org/abs/2302.06912 - AAMAS 2024 (Full Paper)
- *Probabilistic Perspectives on Error Minimization in Adversarial Reinforcement Learning*, **Roman Belaire**, Arunesh Sinha, and Pradeep Varakantham. arxiv.org/abs/2406.04724 - Preprint

RESEARCH EXPERIENCE

Graduate Researcher (Ph.D.)

Singapore Management University

2021-Present | Singapore

- Conduct Ph.D. research on safe, robust, and explainable reinforcement learning (RL) and large language models (LLMs), meeting grant requirements for AI Singapore (AISG) and the Infocomm Media Development Authority (IMDA).
- Independently propose, manage, and execute research schedule, including writing and maintaining the project code base and paper writing, to produce high-quality publications at Tier 1 venues and meet dissertation research goals.
- Provide weekly progress reports and self-guidance to co-authors and project advisors, maintaining a focused and productive research schedule.

Data Scientist – Ph.D. Research Internship

American Express Innovation Labs

2024 | Singapore

- Improved $74.1m losses from 1200+ defaulted accounts by raising risk scores above established cutoffs using proposed models; Proposed models reduced total defaulted dollars by 50% in Australian small business accounts, on recent and out-of-sample data.
- Identified model inefficiencies and behavioral anomalies in Amex's new accounts risk models via interpretability analysis, reducing feature count by 20%.
- Proposed, developed, and delivered novel evaluation metrics for feature selection using marginal Shapley analysis and Gini coefficiency values.
- Worked with VPs to align research outcomes with team-specific business goals including separate business units, such as fraud detection.

Undergraduate Researcher

California State University, Fullerton

2019 | Fullerton, California

- Explored classical machine learning models EEG brain signal classification and brain activity patterns prediction.
- Team lead for 4 undergraduates on machine learning for time series analysis of brainwave signals; facilitated group learning sessions, documentation, and project organization.

RELEVANT TOOLS AND EXPERIENCE

Python, C#, Pytorch, C++, Git/SVN, LLMs, GPT, Generative AI, Reinforcement Learning
Linux, LaTeX, Docker