Logan Cross, PhD

locross93.github.io locross93@gmail.com logan-cross-55861979

Professional Profile

Dedicated researcher with a strong foundation in computer science, machine learning, and cognitive neuroscience. Proven track record in interdisciplinary research, aiming to drive advancements in human-centered AI.

Research and Work Experience

Stanford University, Postdoctoral Scholar **May 2022 - Present**

Hypothetical Minds: Scaffolding Theory of Mind (ToM) for Multi-Agent Tasks with LLMs

- Built LLM-based agent for complex multi-agent tasks that outperforms LLM and deep RL baselines
- ToM module infers other agent's latent states with a natural language approximation of Bayesian inference
- Compared human experimental data to Hypothetical Minds, diagnosing similar performance and reasoning

Advancing LLM-based Adaptive Teaching for Human and LLM Students

- Developed a novel framework for adaptive teaching using LLMs
- Applied pedagogical principles to enhance knowledge distillation in small LLMs

Selective Context Augmentation Pipeline (SCAP) for Social Reasoning Evaluations

- Augments social NLP benchmarks w/ synthetic contextual dialogues to improve LLM evaluations
- Validated w/ LLM/human study showing reduction in ambiguity and improved inter-annotator agreement

Google DeepMind, Research Scientist Intern **May 2021 - Oct 2021**

Cognitive Testing on the Alchemy Meta Reinforcement Learning Benchmark

- Prototyped meta reinforcement learning models on a challenging meta-learning benchmark
- Developed a battery of cognitive tests of agents' abilities in a learning curriculum

California Institute of Technology, Doctoral Student Researcher **Sept 2015 - April 2022**

The Neural Mechanisms of Value Construction

- Compared representations in deep reinforcement learning algorithms to neural representations in the human brain for dynamic, naturalistic tasks (Atari games)
- Leveraged machine learning to create cutting-edge fMRI data analysis tools
- Led extensive research projects to develop computational models of human learning and decision-making

Education

California Institute of Technology, PhD **Computation and Neural Systems**

University of Southern California, B.S. **Neuroscience**

Skills

- **Programming**: Python, C++, MATLAB, R, git, software engineering, Unity3D, Docker, SQL
- **Machine Learning**: PyTorch, Tensorflow, JAX, sklearn, Ray, RLlib, Wandb, Tensorboard, reinforcement learning, LLMs, GPT API, computer vision
- **Mathematics and Statistics**: Calculus, linear algebra, Bayesian modeling, probability
- **Leadership**: Extensive experience mentoring, teaching, and working in multidisciplinary teams

Selected Publications

Cross, L., Xiang, V., Haber, N., & Yamins, D. (2024). Hypothetical Minds: Scaffolding Theory of Mind for Multi-Agent Tasks with Large Language Models. Selected for oral presentation at the Workshop on Open-World Agents (OWA-2024), NeurIPS 2024. arXiv preprint arXiv:2407.07086.

Sun, F-Y., S I, H., Yi, A., Zhou, Y., Zook, A., Tremblay, J., **Cross, L.**, Wu, J., & Haber, N. (2024). FactorSim: Generative Simulation via Factorized Representation. In *Advances in Neural Information Processing Systems*.

Cross, L., Xiang, V., Haber, N., & Yamins, D. (2024). Animate Agent World Modeling Benchmark. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 46).

Cross, L., Cockburn, J., Yue, Y., & O'Doherty, J.P. (2021). Using deep reinforcement learning to reveal how the brain encodes abstract state-space representations in high-dimensional environments. *Neuron*, 109(4), 724-738.

Suzuki, S., **Cross, L.**, & O'Doherty, J. P. (2017). Elucidating the underlying components of food valuation in the human orbitofrontal cortex. *Nature Neuroscience*, 20(12), 1780.