Stewart Slocum

https://stewyslocum.com | Google Scholar | sslocum3@mit.edu

I'm interested in pragmatic AI safety research to reduce risks from misuse and misalignment in current and future AI systems.

Education

Ph.D. Computer Science, *Massachusetts Institute of Technology* **August 2022 - Present**
Advised by Professor Dylan Hadfield-Menell

B.S. Computer Science, Applied Math and Statistics, *Johns Hopkins University* **August 2017 - May 2021**
Advised by Professor Rene Vidal

Awards

NSF GRFP Fellowship

Publications, Preprints, and Conference Presentations

[1] **Stewart Slocum**, Asher Parker-Sartori, Dylan Hadfield-Menell. "Diverse Preference Learning for Capabilities and Alignment." *ICLR* (2025).

[2] Mehul Damani et al. "Beyond Binary Rewards: Training LMs to Reason About Their Uncertainty." *arXiv preprint 2507.16806* (2025).

[3] Parker Whitfill and **Stewart Slocum**. "Beyond Ordinal Preferences: Why Alignment Needs Cardinal Human Feedback." *arXiv preprint arXiv:2508.08486* (2025).

[4] "The AI Agent Index." *arXiv preprint arXiv:2502.01635* (2025).

[5] Zora Che et al. "Model Manipulation Attacks Enable More Rigorous Evaluations of LLM Unlearning." *Neurips Safe Generative AI Workshop* (2024), *TMLR* (2025).

[6] **Stewart Slocum**, Dylan Hadfield-Menell. "Inverse Prompt Engineering for Task-Specific LLM Safety." *Best Paper Runner-Up at AAAI Workshop on Responsible Language Models* (2024).

[7] Stephen Casper et al. "Open Problems and Fundamental Limitations of Reinforcement Learning from Human Feedback." *Transactions on Machine Learning Research* (2023).

[8] Aditya Chattopadhyay, **Stewart Slocum**, Benjamin D. Haeffele, Rene Vidal, Donald Geman. "Interpretable by Design: Learning Predictors by Composing Interpretable Queries." *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2022).

[9] Lauren Wheelock, **Stewart Slocum**, Jorma Gorns, Sam Sinai. "Risk-Adjusted Selection for Validation of Sequences in AAV Design Using Composite Sampling." *American Society of Gene & Cell Therapy Conference* (2021).

[10] Sam Sinai et al. "AdaLead: A simple and robust adaptive greedy search algorithm for sequence design." *arXiv preprint arXiv:2010.02141* (2020).

[11] Vikram Mulligan et al. "Designing peptides on a quantum computer." *bioRxiv* (2019).

Experience

Anthropic *Safety Research Fellow* **March 2025 -**

Leading a research project on using Synthetic Document Finetuning (SDF) to edit LLM beliefs for AI safety applications (overwriting hazardous knowledge, reducing misalignment risk, and building model organisms to test interpretability and alignment auditing methods). This work focuses on basic science research around the method – understanding what makes SDF effective, unintended side-effects on models, and developing behavioral and internal representation-based tests to measure the depth of model belief. Built scalable pipelines for synthetic training data, automatic eval generation, and model finetuning and inference.

Model Evaluation and Threat Research (METR) *Research Collaborator* **January 2025 - March 2025**

Led a short project building long-horizon tasks to evaluate AI agents autonomous AI R&D capabilities. These tasks require replicating recent ML papers from condensed project proposals. We use a qualitative rubric-based scoring system and a milestone-based evaluation infrastructure to efficiently estimate task completion probability at different parts of the research process.

Algorithmic Alignment Group, MIT *PhD Student* **August 2022 -**

Working on technical tools for AI alignment and safety. A few recent projects:

- Diverse Preference Learning for Capabilities and Alignment – develop a method inspired by social choice theory to understand and address mode collapse in aligned LLMs. Diverse Preference Learning aligns output distributions to population-level preferences, improving diversity-quality tradeoffs, social representation, and best-of-N inference scaling.
- Model Tampering Attacks Enable More Rigorous Evaluations of LLM Capabilities – we use fine-tuning and latent-space attacks to estimate and upper-bound vulnerabilities from input-space adversaries.
- Beyond Binary Rewards: Training LMs to Reason about Their Uncertainty – RL technique for LLMs that leads to emergent reasoning about their uncertainty. This improves calibration while maintaining performance, and allows for novel test-time scaling procedures.

Vision Lab, JHU *Research Assistant* **October 2020 - July 2022**

ML theory research on neural network optimization and interpretability. Worked on an information-theoretic framework to optimize for interpretability: Interpretable by Design: Learning Interpretable Predictors.

Dyno Therapeutics, *Machine Learning Intern* **June - August 2020 Full-time, - March 2021 Part-time**

Used generative models and reinforcement learning to optimize viral vectors used in gene therapy. This work led to a methods paper AdaLead, a popular benchmark FLEXS, and oral presentation at ASGTC 2021 (top 25% of submissions).

NASA Goddard Space Flight Center, *Quantum Computing Research Intern* **Summer 2019**

Developed a quantum annealing solver for protein design on the D-Wave quantum computer. We used our method to create the first quantum-designed peptide.

NASA Goddard Space Flight Center, *Virtual Reality Software Engineering Intern* **Summer 2017 and 2018**

Built VR visualization tools for scientific data analysis in astophysics and planetary science. See press coverage and journal publication.