Sam Toyer

Berkeley, USA
sdt@berkeley.edu
qxcv.net/research

> I am a final-year PhD student at UC Berkeley making language models secure, robust, and safe. Previously I have worked in vision, planning, imitation learning, reinforcement learning, and reward learning. After graduation, I am seeking foundation model research positions.

Education

PhD in Computer Science (Machine Learning) University of California, Berkeley. Advised by Professor Stuart Russell.	2018 - *May 2024*
Bachelor of Advanced Computing (R&D, Honours) Australian National University (ANU)	2014 - 2017

Employment

Research Intern/Visiting Student Researcher, Meta	May 2022 - Dec 2022
Research Engineer, Seesure Pty Ltd	Dec 2017 - Jun 2018
Summer Intern & Tutor (TA), Australian National University	Nov 2014 - Jan 2017

Selected Publications

Visit qxcv.net/pubs for a complete list.

A StrongREJECT for Empty Jailbreaks A. Souly, Q. Lu, D. Bowen,, T. Trinh, E. Hsieh, S. Pandey, P. Abbeel, J. Svegliato, S. Emmons, O. Watkins, **S. Toyer**	*Under Review*
Tensor Trust: Interpretable Prompt Injection Attacks from an Online Game **S. Toyer**, O. Watkins, E.A. Mendes, J. Svegliato, L. Bailey, T. Wang, I. Ong, K. Elmaaroufi, P. Abbeel, T. Darrell, A. Ritter, S. Russell. (**Demo at tensortrust.ai)**	ICLR'24 (**spotlight**)
An Empirical Investigation of Representation Learning for Imitation X. Chen,* **S. Toyer**,* C. Wild,* S. Emmons, I. Fischer, K.H. Lee, N. Alex, S.H. Wang, P. Luo, S. Russell, P. Abbeel, R. Shah	NeurIPS'21 (Datasets & Benchmarks)
The MAGICAL Benchmark for Robust Imitation **S. Toyer**, R. Shah, A. Critch, S. Russell	NeurIPS'20
ASNets: Deep Learning for Generalised Planning **S. Toyer**, F. Trevizan, S. Thiébaux, L. Xie	JAIR'20
Variational Discriminator Bottleneck: Improving Imitation Learning, Inverse RL, and GANs by Constraining Information Flow X.B. Peng, A. Kanazawa, **S. Toyer**, P. Abbeel, S. Levine	ICLR'19
Action Schema Networks: Generalised Policies with Deep Learning **S. Toyer**, F. Trevizan, S. Thiébaux, L. Xie	AAAI'18

Selected Awards

Tong Leong Lim Pre-Doctoral Prize (Berkeley EECS)	2019
Berkeley Fellowship for Graduate Study (Berkeley)	2018
University Medal (ANU)	2017
National University Scholarship (ANU)	2014 - 2017

Skills

Deep learning • Reinforcement learning • Imitation Learning • Large Language Models (LLMs) • PyTorch
Technical writing • Python • C/C++ • Django • Javascript • GCP & AWS • Linux administration • Ansible