Tony T. Wang

twang6@mit.edu
tonytwang.net

Education

Massachusetts Institute of Technology — Cambridge, MA
Ph.D. in Computer Science, advised by Nir Shavit — 2021 – Present

M.Eng. in Computer Science, advised by Gregory W. Wornell, GPA: 5.0/5.0 — 2020 – 2021

B.Sc. Double Major in Computer Science and Math, GPA: 4.9/5.0 — 2016 – 2020

- *Master's thesis:* Adversarial Examples in Simpler Settings.
- *Selected CS coursework:* Machine Learning, Inference and Information, Robotic Manipulation, Formal Reasoning about Programs, Cryptography, Compilers, Performance Engineering, Randomized Algorithms, Quantum Computation.
- *Selected math coursework:* Measure Theoretic Probability, Complex Analysis, Functional Analysis, Differential Geometry, General Relativity, Abstract Algebra.

Publications

1. **Tony T. Wang***, Adam Gleave*, Tom Tseng, Kellin Pelrine, Nora Belrose, Joseph Miller, Michael D. Dennis, Yawen Duan, Viktor Pogrebniak, Sergey Levine, Stuart Russell. *Adversarial Policies Beat Superhuman Go AIs*. NeurIPS 2022 ML Safety Workshop (best paper award, top 10/132); ICML, 2023 (oral, top 10%).
2. **Tony T. Wang***, Miles Kai Wang*, Kaivu Hariharan*, Nir Shavit. *Forbidden Facts: An Investigation of Competing Objectives in Llama 2*. NeurIPS 2023 ATTRIB and SoLaR Workshops.
3. **Tony T. Wang**, Igor Zablotchi, Nir Shavit, Jonathan Rosenfeld. *Cliff-Learning*. Preprint, 2023.
4. Stephen Casper*, Xander Davies*, [and 29 others, including **Tony T. Wang**]. *Open Problems and Fundamental Limitations of Reinforcement Learning from Human Feedback*. TMLR, 2023.
5. Simon Alford, Anshula Gandhi, Akshay Rangamani, Andrzej Banburski, **Tony T. Wang**, Sylee Dandekar, John Chin, Tomaso Poggio, Peter Chin. *Neural-guided, Bidirectional Program Search for Abstraction and Reasoning*. Complex Networks, 2021.
6. Yuheng Bu, **Tony T. Wang**, Gregory W. Wornell. *SDP Methods for Sensitivity-Constrained Privacy Funnel and Information Bottleneck Problems*. ISIT, 2021.

Work and Research Experience

Astra Fellowship, Constellation — Berkeley, CA
Research Fellow — Jan 2024 – Present

- Working on language model jailbreak defense.

Shavit Lab, MIT — Cambridge, MA
Research Assistant — Fall 2021 – Present

- Working on AI safety, with a focus on adversarial robustness.

Genesis Therapeutics — Burlingame, CA
AI Engineer Intern — Summer 2021

- Worked on deep neural networks for molecular property prediction.

Signals, Information, and Algorithms Laboratory, MIT Cambridge, MA
Research Assistant (M.Eng.) Summer 2020 – Spring 2021

- Studied toy examples of adversarial examples to unify different aspects of the phenomenon.
- Collaborated with researchers at the Poggio Lab on neurosymbolic algorithms for solving the Abstraction and Reasoning Corpus.

Nvidia Santa Clara, CA
AI-Infra Research Intern Summer 2019

- Researched active learning for self-driving vision models, with a focus on diversity-aware batch-mode sampling.

Five Rings Capital New York City, NY
Quant Research Intern Q1 2019

- Analyzed market data for statistical arbitrage opportunities.

Dropbox San Francisco, CA
Network Reliability Engineering Intern Summer 2018

- Automated traffic draining for production routers.
- Hacked on `mypyc`, a compiler from typed Python to Python C extensions.

DigitalWoven San Mateo, CA
Software Engineering Intern Summer 2017

- Built on AWS the serverless backend for UTStamp, a blockchain notary service.
- Designed and implemented the UTStamp frontend in React.

Awards and Grants

Lightspeed Grant for AI safety research, 234000 USD	2023
Eric and Wendy Schmidt Center PhD Fellowship	2022 - 2023
MIT EECS Harold Hazen Teaching Award	2021
Undergraduate Teaching Assistant Award	2020
USA Computing Olympiad finalist (national top 24)	2013, 2015

Other projects

Roots of Random Polynomials Fall 2019
Term project for 18.821, Project Lab in Mathematics

- Proved roots of high-degree polynomials are roughly uniformly distributed over the unit circle in $\mathbb{C}$.
- Report: web.mit.edu/twang6/public/poly-roots.pdf

Statistical Inference Through the Lens of Information Geometry Spring 2019
Term paper for 18.424, Seminar in Information Theory

- Contains a proof of the Cramér-Rao bound via information geometry.
- Report: web.mit.edu/twang6/public/stats-info-geo.pdf

Voice Identification on the VoxCeleb Dataset Fall 2017
Term project for 6.867, Machine Learning

- Compared RNNs to CNNs for performing speaker identification.
- Report: web.mit.edu/twang6/public/rnn-voxceleb.pdf

Codeforces Round #336 Q4 2015

Competitive programming contest

- Main organizer and problem writer.
- Drew 3000+ participants.
- Particularly proud of authoring codeforces.com/contest/607/problem/C.

Other Activities

MIT AI Alignment 2022 – Present
Member, Advisor

MIT Club Tennis 2022 – Present
Member

MIT Anime Club 2016 – 2021
Member, President, Webmaster

MIT Chamber Music Society 2016 – 2020
Violinist

Peninsula Youth Orchestra 2011 – 2016
Violinist, Assistant Concertmaster