

## EDUCATION

---

<b>Massachusetts Institute of Technology</b> Ph.D. in Computer Science, advised by Nir Shavit	Cambridge, MA 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
<ul style="list-style-type: none"><li>– <i>Master’s thesis</i>: <a href="#">Adversarial Examples in Simpler Settings</a>.</li><li>– <i>Selected CS coursework</i>: Machine Learning, Inference and Information, Robotic Manipulation, Formal Reasoning about Programs, Compilers, Performance Engineering, Randomized Algorithms, Quantum Computation.</li><li>– <i>Selected math coursework</i>: Measure Theoretic Probability, Complex Analysis, Functional Analysis, Differential Geometry, General Relativity, Abstract Algebra.</li></ul>	

## PUBLICATIONS

---

1. **Tony T. Wang\***, Adam Gleave\*, Tom Tseng, Kellin Pelrine, Nora Belrose, Joseph Miller, Michael D. Dennis, Yawen Duan, Viktor Pogrebnik, 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 Zablatchi, 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](#)”. Preprint, 2023.
5. 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

---

<b>Computational Connectomics Group, MIT</b> Research Assistant	Cambridge, MA Fall 2021 – Present
<ul style="list-style-type: none"><li>– Working on AI safety and connectomics. I am trying to understand the computational mechanisms of both artificial and biological neural networks.</li></ul>	
<b>Genesis Therapeutics</b> AI Engineer Intern	Burlingame, CA Summer 2021
<ul style="list-style-type: none"><li>– Worked to understand the behavior and improve the capabilities of deep neural networks for molecular property prediction.</li></ul>	
<b>Signals, Information, and Algorithms Laboratory, MIT</b> ML Researcher (M.Eng.)	Cambridge, MA Summer 2020 – Spring 2021
<ul style="list-style-type: none"><li>– Studied toy examples of adversarial examples to unify different aspects of the phenomenon.</li><li>– Collaborated with researchers at the Poggio Lab on neurosymbolic algorithms for solving the Abstraction and Reasoning Corpus.</li></ul>	

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

- Researched active learning for self-driving, with a focus on diversity-aware batch-mode sampling.
- Implemented and evaluated algorithms to promote batch diversity at scale (across hundreds of thousands of hi-res images).
- Developed t-SNE based visualization tools for batch sampling at scale.

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

- Analyzed market data for statistical arbitrage opportunities.

**Madry Lab, MIT** Cambridge, MA  
 ML Researcher (B.Sc.) 2018

- Explored alternative distance metrics for adversarial examples for deep vision networks.
- Demonstrated empirically that projected-gradient descent attacks generalize to metrics like SSIM and VGG-embedding similarity.

**Dropbox** San Francisco, CA  
 Network Reliability Engineering Intern Summer 2018

- Automated traffic draining for production routers.
- Added primitives to NRE’s distributed task scheduler.
- 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.
- Developed a distributed load-testing tool to stress-test services.
- Designed and implemented the UTStamp frontend in React.

## AWARDS

---

Eric and Wendy Schmidt Center PhD Fellowship	2022 - 2023
<a href="#">MIT EECS Harold Hazen Teaching Award</a>	2021
<a href="#">Undergraduate Teaching Assistant Award</a>	2020
6.035 Compiler Competition winning team	2018
USA Computing Olympiad finalist (national top 24)	<a href="#">2013</a> , <a href="#">2015</a>

## 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](http://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](http://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](http://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](https://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