Megha Srivastava

megha@cs.stanford.edu • www.cs.stanford.edu/~megha
735 Campus Dr., Stanford, CA 94305

Education

Stanford University • Stanford, CA
present
PhD in Computer Science, co-advised by Dorsa Sadigh & Dan Boneh
Supported by the NSF Graduate Research Fellowship (2018 – 2023) and IBM PhD Fellowship (2023-2024)
Current Research Interests: AI and education, human-AI interaction, machine learning safety/security

Massachusetts Institute of Technology • Cambridge, MA
Fall 2023 – Winter 2024
Visiting PhD student at the Computer Science & Artificial Intelligence Laboratory, hosted by Jacob Andreas

Stanford University • Stanford, CA
Fall 2014 – Spring 2019
BS in Computer Science with Honors and Distinction, Minor in Creative Writing
MS in Computer Science (Artificial Intelligence), advised by Percy Liang & Tatsunori Hashimoto
Ben Wegbreit Prize for Best Undergraduate Honors Thesis in Computer Science
Bing Overseas Studies Program at Oxford University (Logic & Computability) in Spring 2017

Experience

Member of Technical Staff (Contractor) – Anthropic
present
San Francisco, CA

- Support research projects and writing.

AI Resident – Microsoft Research
Fall 2019 – Fall 2020
Redmond, WA

- Quantifying effects of nondeterminism in neural network training (Mentors: Besmira Nushi, Eric Horvitz).

Research Intern – Google Research
Summer 2019
Los Angeles, CA

- Improving robustness of image understanding models with the Research & Machine Intelligence team.

Research Intern – ETH Zürich Learning & Adaptive Systems Group
Summer 2018
Zürich, Switzerland

- Active learning and adaptive hypothesis testing (Mentors: Hoda Heidari, Andreas Krause)
- Supported by the ETH Zürich Student Summer Research Fellowship.

Research Assistant – Vision & Perception Neuroscience Group
Summer 2016 – Spring 2017
Stanford, CA

- Generalization and perceptual invariances in both human and artificial vision models (Mentor: Kalanit Grill-Spector)
- Supported by the Bio-X Undergraduate Research Award.

Selected Awards & Fellowships

- Human Robot Interaction (HRI) Pioneers, 2025
- Rising Star in Machine Learning (University of Maryland), 2023
- IBM PhD Fellowship, 2023
- Women in National Security Scholar, 2023 (*Project with Gordian Knot Center on developing AI Literacy*)
- American Association for the Advancement of Science (AAAS) Mass Media Fellowship Finalist, 2019
- International Conference on Machine Learning (ICML) Best Paper Runner-Up Award, 2018
- National Science Foundation Graduate Research Fellowship, 2018
- Ben Wegbreit Prize for Best Undergraduate Honors Thesis in Computer Science (Stanford), 2018
- Tau Beta Pi, 2018
- Bio-X Research Award (Stanford), 2016
- Lunsford Award for Oral Presentation of Research Finalist (Stanford), 2016

Service

- Science Small Groups Mentor (2024)
- Stanford AI Lab Blog Editor, 2020-present
- Student Program Chair for Women in Machine Learning (WiML) 2023, co-located with NeurIPS 2023
- Reviewer for NeurIPS (2020 - curr.), ICLR (2021 - curr.), ICML (2021 - curr.), RA-L 2022, CoRL 2023, L4DC 2024, CHI 2025, HRI-LBR 2026

Preprints

Alexis Ross*, **Megha Srivastava***, Jeremiah Blanchard, Jacob Andreas. *"Modeling Student Learning with 3.8 Million Program Traces"*, in submission.
AI for Education·NLP

Megha Srivastava, Cédric Colas, Dorsa Sadigh, Jacob Andreas. *"Policy Learning with a Language Bottleneck"*, in submission
Spotlight Talk *at Training Agents with Foundation Models Workshop (RLC 2024).*
NLP·Embodied AI ·Human-AI Interaction

Publications (See Google Scholar for full list, * denotes equal contribution)

Megha Srivastava*, Reihaneh Iranmanesh*, Yuchen Cui, Deepak Gopinath, Emily Sumner, Andrew Silva, Laporsha Dees, Guy Rosman, Dorsa Sadigh. *"Shared Autonomy for Proximal Teaching"*, ACM/IEEE International Conference on Human-Robot Interaction (HRI), 2025.
AI for Education ·Embodied AI

Megha Srivastava, Simran Arora, Dan Boneh. *"Optimistic Verifiable Training by Controlling Hardware Nondeterminism"*, Advances in Neural Information Processing Systems 38 (NeurIPS) , 2024.
Robust ML·Security

Neil Perry*, **Megha Srivastava***, Deepak Kumar, Dan Boneh. *"Do Users Write More Insecure Code with AI Assistants?"* ACM Conference on Computer and Communications Security (CCS), 2023. **mlsec.org Top-100 Computer Security Papers** 🔗
Security ·Human-AI Interaction

Megha Srivastava, Noah Goodman, Dorsa Sadigh. *"Generating Language Corrections for Teaching Physical Control Tasks"* Proceedings of the 40th International Conference of Machine Learning (ICML) , 2023.
AI for Education ·Embodied AI

Mina Lee, **Megha Srivastava**, Amelia Hardy, John Thickstun, Esin Durmus, Ashwin Paranjape, Ines Gerard-Ursin, Xiang Lisa Li, Faisal Ladhak, Frieda Rong, Rose E. Wang, Minae Kwon, Joon Sung Park, Hancheng Cao, Tony Lee, Rishi Bommasani, Michael Bernstein, Percy Liang *"Evaluating Human-Language Model Interaction"* Transactions of Machine Learning Research (TMLR), 2023.

§ **Question Answering**: Megha Srivastava, John Thickstun, Rose Wang, Minae Kwon, Mina Lee
§ **Crossword Puzzles**: Megha Srivastava

Human-AI Interaction ·NLP

Megha Srivastava, Erdem Biyik, Suvir Mirchandani, Noah Goodman, Dorsa Sadigh. *"Assistive Teaching of Motor Control Tasks to Humans"* Advances in Neural Information Processing Systems 36 (NeurIPS) , 2022.
AI for Education ·Embodied AI

Siddharth Karamcheti*, **Megha Srivastava*** Percy Liang, Dorsa Sadigh. *"LILA: Language-Informed Latent Actions"* Conference on Robot Learning (CoRL), 2021.
Embodied AI·NLP

Megha Srivastava and Noah Goodman. *"Question Generation for Adaptive Education"* Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics (ACL), 2021.
AI for Education·NLP

Megha Srivastava, Tatsunori Hashimoto, Percy Liang. *"Robustness to Spurious Correlations via Human Annotations"* Proceedings of the 37th International Conference on Machine Learning (ICML), 2020.
Robust ML·NLP·Human-AI Interaction

Megha Srivastava, Besmira Nushi, Ece Kamar, Shital Shah, Eric Horvitz *"An Empirical Analysis of Backward Compatibility in Machine Learning Systems "*Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD), 2020.
Robust ML

Megha Srivastava, Hoda Heidari, Andreas Krause. *"Mathematical Notions vs. Human Perception of Fairness: A Descriptive Approach to Fairness for Machine Learning."*Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD), 2019.
Human-AI Interaction

Tatsunori Hashimoto, **Megha Srivastava**, Hongseok Namkoong, Percy Liang. *"Fairness Without Demographics in Repeated Loss Minimization."* Proceedings of the 35th International Conference on Machine Learning (ICML), 2018. **Best Paper Runner-Up Award**
Robust-ML

Other Works (Blog Posts, Abstracts, Technical Reports, etc.)

- **Megha Srivastava** and Claude Sonnet 3.5. *"Echo: A multi-agent AI system for patient-centered pharmacovigilance."* 1st Open Conference of AI Agents for Science, 2025. **Spotlight Talk**. *[Experimental conference exploring human-AI collaboration in research. More details in blog post on research process* 🔗*]*
- Sam Bowman, **Megha Srivastava**, Jon Kutasov, Rowan Wang, Trenton Bricken, Benjamin Wright, Ethan Perez, and Nicholas Carlini. *"Findings from a Pilot Anthropic—OpenAI Alignment Evaluation Exercise."* Anthropic, 2025. 🔗
- Rose Wang and **Megha Srivastava**. *"Productive Struggle: The Future of Human Learning in the Age of AI."* The Stanford AI Lab Blog, 2025. Featured and translated in Spanish by the Inter-American Development Bank. 🔗
- **Megha Srivastava** and John Thickstun. *"Observations from HALIE: A Closer Look at Human-LM Interactions in Information-Seeking Contexts."* Center for Research on Foundation Models Blog, 2023. 🔗
- Kristin Lauter*, Cathy Yuanchen Li*, Krystal Maughan*, Rachel Newton*, **Megha Srivastava***. *"Machine learning for modular multiplication."* Proceedings of the WIN6 Workshop at Banff International Research Station for Mathematical Innovation and Discovery, 2023.
- **Megha Srivastava**, David Remus, Kalanit Grill-Spector. *"The Role of Learning in Complex Object Recognition and Discrimination Across Spatial Transformations: An Experimental Comparison of Artificial CNNs and Human Subjects."* Vision Sciences Society (VSS), 2017.

Teaching and Mentoring

- Course Assistant for CS 255 (Introduction to Cryptography, 2025)
- Course Assistant for CS 221 (Introduction to Artificial Intelligence, 2018)
- Instructor for Stanford Splash (Demystifying Hot Topics in Computer Science , 2017)
- Mentored research of undergraduate and MS students: *Reihaneh Iranmanesh (Amherst College, co-author on HRI'25 paper), Zhiyin Lin (Stanford CURIS), Mallika Parulekar (Stanford), Jonathan Ouyang (UCLA), Benita Wong (Stanford)*

Guest Lectures

1. "Security of Modern Machine Learning Systems" *AI4All, June 2025*
2. Beyond Instruction Following: Language and Human-Robot Interaction *Stanford University (Guest Lecture for CS 329X: Human-Centered NLP), October 2024*
3. "Do Users Write More Insecure Code with AI Assistants?" *University of Maryland College Park (Guest Lecture for Large Language Models, Security, and Privacy seminar), October 2023*

Invited Talks

1. "Modeling Student Learning with 3.8 Million Program Traces" *Deep Learning: Classics and Trends (ML Collective), November 2025*
2. "Security of Modern Machine Learning" *Accenture Responsible AI Course, September 2025*
3. "Optimistic Verifiable Training by Controlling Hardware Nondeterminism" *Google Privacy in ML Seminar, February 2025*
4. "Security of Modern Machine Learning" *DaVita x Stanford Human-Centered AI, February 2025*
5. "Optimistic Verifiable Training by Controlling Hardware Nondeterminism" *AI and Cryptography Workshop, Joint Mathematics Meeting, January 2025*
6. "Optimistic Verifiable Training by Controlling Hardware Nondeterminism" *University of Washington Security Seminar, January 2025*
7. "New Challenges of Trust with Large-Scale AI Systems" *University of Chicago, November 2024*.
8. "Security of Modern Machine Learning" *Stanford Cybersecurity & Privacy Festival, October 2024*
9. "Policy Learning with a Language Bottleneck" ***Spotlight Talk*** *at Training Agents with Foundation Models Workshop (RLC 2024), August 2024*.
10. "Optimistic Verifiable Training by Controlling Hardware Nondeterminism" *University of Toronto, July 2024*
11. "Security of Modern Machine Learning" *Accenture Responsible AI Course, July 2024*
12. "Optimistic Verifiable Training by Controlling Hardware Nondeterminism" *Stanford Security Forum, April 2024*
13. "Challenges in Human-AI Interaction for Information-Seeking Tasks" *Rising Stars in Machine Learning Workshop, November 2023*

14. “Robustnesss to Spurious Correlations via Human Annotations” *Two Sigma PhD Symposium*, *June 2023*.
15. “Do Users Write More Insecure Code with AI Assistants?” *UC Berkeley Security Seminar* , *May 2023*
16. “Assistive Teaching of Motor Control Tasks to Humans” *Stanford Collaborative Haptics in Robotics and Medicine Lab*, *May 2023*.
17. “Assistive Teaching of Motor Control Tasks to Humans” *Simons Institute Workshop on AI & Humanity* , *July 2022*.
18. “Assistive Teaching of Motor Control Tasks to Humans” *SystemX Alliance Fall Conference* , *November 2022*.
19. “Mathematical Notions vs. Human Perception of Fairness: A Descriptive Approach to Fairness for Machine Learning” *Oxford University Algorithms at Work Reading Group*, *May 2021*.
20. “Mathematical Notions vs. Human Perception of Fairness: A Descriptive Approach to Fairness for Machine Learning” *Microsoft Research AI & Society Reading Group*, *October 2019*.
21. “Fairness & Robustness with Missing Information” *Stanford Causality & Cognition Lab*, *December 2020*.