Serena Booth

serena.l.booth@gmail.com
slbooth.com

I am a reinforcement learning researcher with expertise in learning from human feedback. I study the design of better AI and robot systems by giving people the tools to *change* agent behavior and by enabling agents to *explain* their behavior. I am currently working in the U.S. Senate for the Banking Committee to advance U.S. AI regulation in high-risk applications.

Education

2018–2023 Ph.D., Computer Science, CSAIL, Massachusetts Institute of Technology.
I was a member of the Interactive Robotics Group, advised by Prof. Julie Shah.
I was sponsored by an NSF GRFP and an MIT Jacobs Presidential Fellowship.
Thesis: *Building Blocks for Human-AI Interaction: Specify, Model, Inspect, and Revise*

2021–2022 Visiting Researcher, University of Texas, Austin.
With Dr. Brad Knox, Dr. Alessandro Allievi, Prof. Scott Niekum, and Prof. Peter Stone

2020 S.M., Computer Science, Massachusetts Institute of Technology, *GPA: 5.0/5.0.*
Thesis: *Explainable AI Foundations to Support Human-Robot Teaching and Learning*

2016 B.A., Computer Science with in-field Highest Honors, Harvard University, *GPA: 3.8/4.0.*
Awarded Thomas T. Hoopes thesis prize for excellence in undergraduate research.
Thesis: *Piggybacking Robots: Overtrust in Human-robot Security Dynamics*

Awards, Grants, & Recognition

2023–2024 AAAS Science and Technology Policy Fellow (STPF), with a focus on Artificial Intelligence.
I am spending a year working in the U.S. Senate on AI Policy through a highly-selective (≈ 3%) fellowship.
I am working for the Senate Committee on Banking, Housing, and Urban Affairs, under Sen. Brown (D-OH).

2018–2023 National Science Foundation Graduate Research Fellowship (NSF GRFP), ≈$150,000.

2023 HRI Pioneer, *Human-robot interaction "premier forum" for graduate students.*

2022 EECS Rising Star, *A selective and intensive workshop for underrepresented genders in EECS.*

2018 MIT Jacobs Presidential Fellowship, ≈$100,000.

— Finalist, Paul and Daisy Soros Fellowship for New Americans.

2016 Thomas T. Hoopes Prize for Excellence in Undergraduate Research, Harvard University.

2011 National Winner, Aspirations in Computing, NCWIT.

Conference & Journal Publications

TMLR '24 W. Bradley Knox, Stephane Hatgis-Kessell, **Serena Booth**, Scott Niekum, Peter Stone, & Alessandro Allievi, *Models of Human Preference for Learning Reward Functions,* arxiv.org/abs/2206.02231, Transactions on Machine Learning Research.

AAAI '24 W. Bradley Knox, Stephane Hatgis-Kessell, Sigurdur Orn Adalgeirsson, **Serena Booth**, Anca Dragan, Peter Stone, & Scott Niekum, *Learning Optimal Advantage from Preferences and Mistaking it for Reward,* arxiv.org/abs/2310.02456, AAAI Conference on Artificial Intelligence.

AAAI '24 Allen Chang, Matthew Fontaine, **Serena Booth**, Maja Mataric, & Stefanos Nikolaidis, *Quality-Diversity Generative Sampling for Learning with Synthetic Data*, AAAI Conference on Artificial Intelligence.

AAAI '23 **Serena Booth**, W. Bradley Knox, Julie Shah, Scott Niekum, Peter Stone, & Alessandro Allievi, *The Perils of Trial-and-Error Reward Design: Misdesign through Overfitting and Invalid Task Specifications*, AAAI Conference on Artificial Intelligence. Selected for oral presentation.
– Webpage: slbooth.com/Reward_Design_Perils

RLDM '22 **Serena Booth**, W. Bradley Knox, Julie Shah, Scott Niekum, Peter Stone, & Alessandro Allievi, *Extended Abstract: Graduate Student Descent Considered Harmful? A Proposal for Studying Overfitting in Reward Functions*, Multidisciplinary Conference on Reinforcement Learning and Decision Making.

RLDM '22 W. Bradley Knox, Stephane Hatgis-Kessell, **Serena Booth**, Scott Niekum, Peter Stone, & Alessandro Allievi, *Spotlight, Extended Abstract: Partial Return Poorly Explains Human Preferences*, Multidisciplinary Conference on Reinforcement Learning and Decision Making. Selected for oral presentation.

HRI '22 **Serena Booth**, Sanjana Sharma, Sarah Chung, Julie Shah, & Elena L. Glassman, *Revisiting Human-Robot Teaching and Learning Through the Lens of Human Concept Learning Theory*, ACM/IEEE International Conference on Human-Robot Interaction (HRI).
- Webpage: slbooth.com/HRI_Concept_Learning
- Press: https://news.mit.edu/2022/humans-understand-robots-psychology-0302

AAAI '22 Yilun Zhou, **Serena Booth**, Marco Ribeiro, & Julie Shah, *Do Feature Attribution Methods Correctly Attribute Features?*, AAAI Conference on Artificial Intelligence.
- Webpage: yilunzhou.github.io/feature-attribution-evaluation
- Press: https://news.mit.edu/2022/test-machine-learning-models-work-0118

AAAI '21 **Serena Booth***, Yilun Zhou*, Ankit Shah, & Julie Shah, BAYES-TREX*: A Bayesian Sampling Approach to Model Transparency by Example*, AAAI Conference on Artificial Intelligence. *Equal Contribution.
- Webpage: slbooth.com/BayesTrEx
- Press: https://news.mit.edu/2021/more-transparency-understanding-machine-behaviors-bayes-trex-0322

CoRL '21 Yilun Zhou, **Serena Booth**, Nadia Figueroa, & Julie Shah, *RoCUS: Robot Controller Understanding via Sampling*, Conference on Robot Learning (CoRL).
- Webpage: yilunzhou.github.io/RoCUS

AIES '21 Aspen Hopkins* and **Serena Booth***, *Machine Learning Practice Outside Big Tech: How Resource Constraints Challenge Responsible Development*, AAAI/ACM Conference on Artificial Intelligence, Ethics, and Society (AIES). *Equal Contribution. Selected for oral presentation (9.6%).

Frontiers '21 Alan F T Winfield, **Serena Booth**, Louise A Dennis, Takashi Egawa, Helen Hastie, Naomi Jacobs, Roderick Muttram, Joanna I Olszewska, Fahimeh Rajabiyazdi, Andreas Theodorou, Mark Underwood, Robert H Wortham, and Eleanor Watson, *IEEE P7001: A Proposed Standard on Transparency*, Frontiers in Robotics and AI, Ethics in Robotics and Artificial Intelligence.
- Supplemental paper to our published IEEE standard: "IEEE Standard for Transparency of Autonomous Systems," in IEEE Std 7001-2021, pp. 1-54, 4 March 2022, doi: 10.1109/IEEESTD.2022.9726144.

IJCAI '19 **Serena Booth**, Christian Muise, & Julie Shah, *Evaluating the Interpretability of the Knowledge Compilation Map: Communicating Logical Statements Effectively*, International Joint Conference on AI (IJCAI).
- Webpage: slbooth.com/LogicInterpret

HRI '17 **Serena Booth**, James Tompkin, Hanspeter Pfister, Jim Waldo, Krzysztof Gajos, & Radhika Nagpal, *Piggybacking Robots: Human-robot Overtrust in University Dormitory Security*,
ACM/IEEE International Conference on Human-Robot Interaction (HRI).
- Webpage with video: slbooth.com/PiggybackingRobots
- Press: Vice, PhD Comics #2033, FujiTV (Japan), Soonish (Weinersmith and Weinersmith, 2017).

Workshop Publications

ICML '23 W. Bradley Knox, Stephane Hatgis-Kessell, Sigurdur Orn Adalgeirsson, **Serena Booth**, Anca Dragan, Peter Stone, & Scott Niekum, *Learning Optimal Advantage from Preferences and Mistaking it for Reward*, 2023 ICML Workshop on The Many Facets of Preference-based Learning (MFPL). Selected for oral presentation..

HRI '23 Tiffany Horter, Elena Glassman, Julie Shah, & **Serena Booth**, *Varying How We Teach: Adding Contrast Helps Humans Learn about Robot Motions*, 2023 HRI Workshop on Human-Interactive Robot Learning.

NAACL '22 Yiming Zheng, **Serena Booth**, Julie Shah, & Yilun Zhou, *The Irrationality of Neural Rationale Models*, 2022 NAACL Workshop on Trustworthy Natural Language Processing (TrustNLP).

NeurIPS '21 Yilun Zhou, **Serena Booth**, Marco Ribeiro, & Julie Shah, *Do Feature Attribution Methods Correctly Attribute Features?*, NeurIPS 2021 XAI4Debugging Workshop. Selected for oral presentation.

ICRA '21 **Serena Booth**, Sanjana Sharma, Sarah Chung, Julie Shah, & Elena Glassman, *How to Understand Your Robot: A Design Space Informed by Human Concept Learning*, ICRA 2021 Workshop on Social Intelligence in Humans and Robots (SIHR).

AAAI '20 **Serena Booth***, Ankit Shah*, Yilun Zhou*, Julie Shah, *Sampling Prediction-Matching Examples in Neural Networks: A Probabilistic Programming Approach*, AAAI 2020 Workshop on Statistical Relational Artificial Intelligence (StarAI). *Equal Contribution.

AAAI '20 Christian Muise, Salomon Wollenstein Betech, **Serena Booth**, Julie Shah, Yasaman Khazaeni, *Modeling Blackbox Agent Behaviour via Knowledge Compilation*, AAAI 2020 Workshop on Plan, Activity, and Intent Recognition (PAIR).

Teaching

2020–2022 Socially Responsible Computing (SERC) Scholar, Massachusetts Institute of Technology.
- I have contributed extensively to MIT's social and ethical responsibilities for computing curricula. Some of my contributions are available on OpenCourseware: Machine Learning (here), Software Studio (here).
- MIT OpenCourseware released a podcast describing some of my work: ChalkRadio Podcast
- Designed ethics curriculum for *6.036: Machine Learning* (Spring 2021, Spring 2022). MIT News article.
- Advised *6.031: Software Construction* (Fall 2021), and co-designed ethics curriculum for class.
- Embedded socially-responsible education materials into *6.170: Software Studio* (Fall 2020).
- Guest lecture in *6.170* (2020 & 2021) on Milo Phillips-Brown and Abby Jacques' Ethics Protocol.

2020–2021 Graduate Teaching Fellow, Massachusetts Institute of Technology.
- *24.133: Experiential Ethics*, Summer 2021 & Summer 2020
  A philosophy course to explore ethical and social dimensions of students' everyday experiences.
- *6.S897: Classics of Computer Science*, Spring 2020.
  Described by students as a "finishing school" for CS, reviews influential CS papers from 1920-1980.

2015–2016 Undergraduate Teaching Fellow, Harvard University.
- *CS189: Autonomous Robot Systems* (Spring 2016).
  An introduction to robotics. I migrated the course to TurtleBots and developed six courseworks.
- *CS121: Theory of Computation* (Fall 2015).
  An introduction to CS theory, covering computability and complexity theory.
- *CS1: Great Ideas in Computer Science* (Spring 2015).
  An introductory CS class (a CS50 alternative for non-CS majors); taught in Java.

Industry Experience

2021–2022 Bosch R&D, Reinforcement Learning for Autonomous Driving Group, Research Intern.
- Working on reinforcement learning and reward design.
- Collaborating with Dr. Brad Knox, Dr. Alessandro Allievi, Prof. Peter Stone, and Prof. Scott Niekum

2016–2018 Google, APM: Associate Product Manager (APM).
- Before graduate school, I worked full time at Google for two years as a product manager.
- APM II: ARCore, augmented reality SDK. PM'ed 6DoF motion tracking, depth, scene understanding algorithms, and quality, measured by in-lab benchmarking and user-collected telemetry. Scaled ARCore to reach 100M devices.
- APM I: Google Search. Verticals included cars, motorbikes, lottery, commutes, and dinosaurs.
- 20% Project: Human-Robot Interaction at Google[X] (2017)

2015 Apple, Software Engineering Intern, Engineering for Retail.
- I prototyped a pre-ARKit augmented reality "Try It On" feature for the Apple Store App.

2014 Intuit, Software Engineering Intern, Payments & Commerce Solutions.

2011–2012 Imagineer in Behavioral Research, Walt Disney Imagineering Research & Development.
- Ran field tests at Paris and Orlando Parks; studied hotel towel use patterns and guest distribution in parks.

Mentoring

2022–2023 [Human-Robot Interaction] Tiffany Horter, Wellesley.

2020–2023 [Explainable AI, co-supervised with Yilun Zhou] Yiming Zheng, MIT.

2021–2023 [Reinforcement Learning, co-supervised with UT Faculty] Stephane Hatgis-Kessell, UT Austin.

2023 [Algorithmic Fairness, co-supervised with USC Faculty] Allen Chang, USC.

2021 [Science Policy] Reagan Zimmerman, MIT.

2021 [Science Policy] Emily Levenson, MIT.

2020 [Explainable AI, co-supervised with Yilun Zhou] Rene Reyes, MIT.

2020 [Explainable AI, co-supervised with Yilun Zhou] Paul Calvetti, MIT.

2020 [Explainable AI, co-supervised with Yilun Zhou] Melissa Calvert, MIT.

Science Advocacy

2023–2024 AI Policy Fellow, AAAS Science and Technology Policy Fellowship.
- Senate Committee on Banking, Housing, and Urban Affairs, under Sen. Sherrod Brown (D-OH).
- Meeting with banks, advocates, and regulators to gather information about their use and approaches to AI.
- Researching existing regulations in these domains, and seeking to identify regulatory gaps.
- Writing letters from the Committee to apply public pressure for policy changes.
- Organizing hearings around AI and technology, particularly in banking and housing.

2019–2024 MIT Science Policy Initiative (SPI).

I have been an executive member of SPI in many capacities:
- President, 2021–2022; Vice President 2020–2021, Acting Vice President 2022–2023.
  I launched State and Local Visit Days to teach students about careers in state and local government.
  I advocated for the COMPETES and USICA bills, and for the GRAD Caucus [grad-caucus.github.io].
- Executive Visit Days Chair, 2022–2023.
  I brought 20 students and postdocs to D.C. to learn about careers and policymaking in federal agencies.
- Alumni Relations Chair, 2023–2024.
  I continue to advise SPI as an alumni and former president, and am organizing the alumni community.
- Community Member, 2018–2020.
  I advocated in Congress for increased science funding & harassment monitoring in sciences.

2020–2021 Associate Editor, MIT Science Policy Review.

2019–2021 IEEE Standards Working Group P7001, Transparency of Autonomous Systems.

Diversity & Inclusion

2021 Panelist, MIT: Picture A Scientist Viewing and Panel.

I presented the current student perspective on DEI at MIT and in academia.

— Advisory Board Member to MIT EECS Committee on Diversity, Equity, and Inclusion.

— CRA-W Cohort for Graduate Women.

2019 Co-President, MIT GW6: Graduate Women of EECS.
- Co-organized the first GW6 Research Summit to foster fellowship and collaboration across EECS.
- Supported the community of EECS Grad Women through social, professional, and wellness events.

— Instructor, Beautiful Patterns: Intro to CS, Puebla, Mexico.

Taught a week-long course on foundations of computer science to 20 high school women in Mexico.

— Outreach Talk for MOSTEC High School Students.

2018 Panelist, WeCode (Women Engineers Code) Conference, Harvard University.

Participated on two panels discussing product management. Hosted mentoring lunch.

— Teaching Assistant, International Women's Day Android Things Workshop, Google.

— She Innovates Hackathon Mentor, University of Pittsburgh.

2017–2018 VEX Robotics Mentor: Space Cookies, Girl Scouts Teams supported by NASA Ames.

2017 Society of Women Leaders Retreat Panelist, Stanford University.

2014–2015 Student Volunteer, WeCode (Women Engineers Code) Conference, Harvard University.

2014 EngageCSEdu: Student Researcher, Google Research & NCWIT, www.engage-csedu.org.
Researched deterrents to women & minorities studying CS, aggregated inclusive CS course materials.

Additional Academic & Institute Service

2023 Workshop Organizer: Variable Autonomy for Human-Robot Teaming (VAT).
Workshop co-located with Human-Robot Interaction Conference (HRI) 2023.

2019–2023 Secretary (2021-2022), Treasurer (2020-2021), Events Chair (2019-2020), MIT European Club.
I help organize a career fair for 1000+ attendees and 100+ European companies and universities.

2020 Workshop Organizer: Virtual, Augmented, and Mixed Reality for Human-Robot Interaction.
Workshop co-located with Human-Robot Interaction Conference (HRI) 2020.

2020 Guest Talk, *Interpretability in Machine Learning*, MIT IAP AI Policy & Technology Workshop.

2017–2018 Botball Robotics Competition Judge, NASA Ames.

2017 Reality, Virtually, Hackathon Mentor, MIT Media Lab.

Reviewing TLMR; ICML; AAAI; HRI; CHI; Workshops at HRI, ICRA, RSS, NeurIPS, and many others.

Invited Talks, Lectures, & Panels

2023 Stanford ILEAD Lab Meeting, *Iterative Reward Design: Specify, Model, Inspect, and Revise.*

2022 USC Robotics Seminar, *Conceptual Model Formation for Human-Robot Interaction.*

2022 UC Berkeley InterACT Lab Meeting, *Conceptual Model Formation for Human-Robot Interaction.*

2022 Queens University Seminar, *Conceptual Model Formation for Human-Robot Interaction.*

2022 HRI Human-Interactive Robot Learning (HIRL) Workshop, *Human Concept Learning for HRI.*

2022 MIT Open Learning, with Prof. Elena Glassman, *How Humans Can Understand Robot Behaviors.*

2021 Brown University Robotics Symposium, *Humans Learning About Robots Learning About Humans.*

2021 Panelist, NeurIPS Meaningful Representations of Life Workshop.

2021 Guest Lecture, SP.250 Good Intentions $\rightarrow$ Good Outcomes, *Ethical Dilemmas in Big Tech.*

2021 Guest Lecture, 6.170 Software Studio, *Socially-Responsible Computing: The Ethics Protocol.*

2021 Panelist, MIT PKG Center, *Building Pathways Towards Tech for Good Careers.*

2020 Guest Lecture, 6.170 Software Studio, *Socially-Responsible Computing: The Ethics Protocol.*

2020 MIT IAP Policy & Technology Workshop, *Interpretability in Machine Learning.*