Dorsa Sadigh

Contact

246 Gates Computer Science Building
353 Jane Stanford Way, Stanford, CA 94305
+1 (949) 929-3559 https://dorsa.fyi
dorsa@cs.stanford.edu https://github.com/Stanford-ILIAD

Current Position

Stanford University **September 2017 - present**
Assistant Professor
Department of Computer Science and Department of Electrical Engineering

Education

University of California, Berkeley **2017**
Ph.D. in Electrical Engineering and Computer Sciences
Advisors: Sanjit Seshia and Shankar Sastry
Thesis: *Safe and Interactive Autonomy: Control, Learning, and Verification*

University of California, Berkeley **2012**
B.S. in Electrical Engineering and Computer Sciences

Awards

ONR Young Investigator Program Award **2022**

DARPA Young Faculty Award **2022**

IEEE RAS Early Career Award **2022**

Sloan Foundation Fellowship **2022**

Okawa Foundation Research Grant **2021**

MIT TR35 **2021**

JP Morgan Faculty Award **2021-22**

AFOSR Young Investigator Program Award **2020**

Best Paper Award **2020**
Conference on Robot Learning (CoRL) for *"Learning Latent Representations to Influence Multi-Agent Interaction"*

Best Student Paper Award (Finalist) **2020**
Robotics: Science and Systems (RSS) for *"Shared Autonomy with Learned Latent Actions"*

IEEE TC-CPS Early Career Award **2020**

Best Paper Award (Honorable Mention) **2020**
ACM/IEEE International Conference on Human-Robot Interaction (HRI) for *"When Humans Aren't Optimal: Robots that Collaborate with Risk-Aware Humans"*

National Science Foundation CAREER Award **2020**

Gilbreth Lecturer at National Academy of Engineering **2020**

Google Faculty Research Award **2020**

Amazon Research Award **2019**

Best Paper Award (Finalist) **2019**
European Control Conference (ECC), for *"Human-Robot Interaction for Truck Platooning Using Hierarchical Dynamic Games"*

Best Paper Award **2019**
ICML Workshop on Adaptive & Multitask Learning: Algorithms & Systems, for *"Continual Adaptation for Efficient Machine Communication"*

Best Cognitive Robotics Paper (Finalist) **2016**
IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) for *"Information Gathering Actions over Human Internal State"*

Leon O. Chua Award **2016**
for excellence in non-linear science, EECS Department, UC Berkeley 2016

Google Anita Borg Scholarship **2016**

National Defense Science and Engineering Graduate Fellowship **2013**

National Science Foundation Graduate Research Fellowship **2013**

CRA Outstanding Undergraduate Researcher Award **2012**

Arthur M. Hopkin Award **2010**
EECS Department, UC Berkeley

Teaching

CS 237B: Robot Autonomy II **Winter 2020 - 2023**
Instructor, Stanford University.

CS 221: Artificial Intelligence **Spring 2018, 2019, Fall 2019-2022**
Instructor, Stanford University.

CS 521: Seminar on AI Safety **Spring 2018, 2020**
Instructor, Stanford University.

CS 333: Safe and Interactive Robotics **Fall 2017, 2018, Winter 2022**
Instructor, Stanford University.

Advising & Mentoring

Current Graduate Students
Minae Kwon, Siddhartha Karamcheti (co-advised with Percy Liang), Andy Shih (co-advised with Stefano Ermon), Suneel Belkhale, Megha Srivastava (co-advised with Dan Boneh), Jennifer Grannen, Priya Sundaresan (co-advised with Jeannette Bohg), Joey Hejna, Suvir Mirchandani, Hengyuan Hu, Jensen Gao, Jonathan Yang (co-advised with Chelsea Finn)

Past Graduate Students
Erdem Bıyık – Thesis: Learning Preferences for Interactive Autonomy
(Starting as an Assistant Prof. in Computer Science at USC)

Zhangjie Cao – Thesis: Learning from Imperfect Demonstrations
(Quant Research at JQinvestments, China)

Mengxi Li – Thesis: Learning to Adapt for Intelligent Robot Behavior
(Quant Research at Citadel Securities)

Past Postdoctoral Students
Dylan Losey (Assistant Prof. in Mechanical Engineering at Virginia Tech)

Past Undergraduate Students
Nick Landolfi (Ph.D. student in CS at Stanford), Zhiyang He (Ph.D. student in EECS at UC Berkeley), Zheqing Zhu (Ph.D. student in MS&E at Stanford), Jovana Kondik (Ph.D. student in EECS at MIT), Songyuan Zhang (Ph.D. student in EECS at MIT), Albert Zhai (Ph.D. student in CS at UIUC), Woody Wang (Ph.D. student in CS at Stanford), Suvir Mirchandani (Ph.D. student in CS at Stanford), Vivek Myers (PhD. student in EECS at UC Berkeley), Zihan Wang (Ph.D. student in CS at UW).

Outreach

Stanford CS Mentorship Program **2018 - present**
I have organized the Stanford CS mentorship program, where we connect underrepresented minorities and female undergraduate students interested in AI with Ph.D. students at Stanford to meet monthly and discuss research and career choices.

Faculty Mentor for Stanford Robotics Club **2017 - 2020**
I mentor the Stanford undergraduate Robotics Club. Every year they work towards participating in a robotics competition. They have won the third place in the University Rover Challenge in 2019.

Faculty Mentor for Inclusion in AI **2018 - present**
I mentor the Stanford AI Lab graduate group "Inclusion in AI". The group holds regular social and networking events for Stanford AI Lab graduate students.

Talks at Women and Inclusion in STEM events and panels
AI4ALL summer program, Girls Who Code summer program, Gender in Robotics Workshop at Stanford, Berkeley-Stanford Meetup, Rising Stars (EECS) of 2018, Rising Stars (Mechanical Engineering) of 2019, Inclusion in AI.

Talks at Graduate and Undergraduate Student Groups
Undergrad CS Women (WiCS), Grad Engineering Women (SWE), SAIL (Stanford AI Lab women), Women in Electrical Engineering, Women in Aero/Astro, Fire-Side chat with Stanford Undergrads.

EEGSA Outreach Member **2012 - 2017**
Visiting local K-12 schools and presenting engineering projects and demonstrations.

WICSE Outreach Coordinator **2014 - 2015**
Organizing events and outreach activities aiming young girls involvements in STEM.

Work Experience

Microsoft Research, Redmond **June - August 2015**
Internship at the Adaptive Systems and Interaction group with Ashish Kapoor and Eric Horvitz.

Stanford Research Institute, International **June - August 2013**
Internship at the Computer Science Laboratory in the formal methods group with Ashish Tiwari.

Professional Activities

Workshop Organizer **2022**
Conference on Robot Learning

Program Co-Chair **2018 - 2022**
Bay Area Robotics Symposium

Program Co-Chair **2022**
Workshop on Algorithmic Foundations of Robotics (WAFR)

Award Committee **2021**
Conference on Robot Learning

Publicity Chair **2021**
ACM International Conference on Hybrid Systems: Computation and Control

AAAI ACM SIGAI Dissertation Award Committee **2020, 2021**

Vice President for Publications in IEEE Robotics and Automation Society 2021

DARPA Information Science and Technology (ISAT) Study Group Committee Member **2021**

Center for AI Safety at Stanford **2018 - 2022**
Founding member of the Center for AI Safety at Stanford along with Mykel Kochenderfer, Clark Barrett, and David Dill. The center is focused on safety and verification issues for AI and machine learning systems.

Human-Centered AI Institute (HAI) **2018 - present**
Member of the design committee of Human-Centered AI Institute at Stanford. In addition I have been part of the HAI Ethical Review Board (ERB) committee.

Program Committee (Associate Editor, Area Chair)
RSS 2022, RA-L 2021, CoRL 2021-2020, RSS 2020, HRI 2020, HRI 2022, L4DC 2020, CAV 2019, HSCC 2019, CoRL 2018, ICRA 2018, HSCC Repeatability Eval 2016.

External Reviewer for Conferences, Journals, and Grant Panels
- *Robotics:* RA-L, RSS, CoRL, WAFR, ICRA, HRI, TASE, ACM TECS
- *Control Theory:* HSCC, CDC, ACC, TCST
- *Formal Methods:* CAV, FM, HVC, VMCAI
- *NSF and AFOSR Proposal Panelist and Reviewer*

Invited Talks

Princeton Robotics Seminar 2023

HRI Workshop on Lifelong Learning and Personalization in Long-Term Human-Robot Interaction 2023

CMU Robotics Institute Seminar 2023

AAAI Workshop on Reinforcement Learning Ready for Production 2023

NeurIPS Workshop on Language and Reinforcement Learning 2022

NeurIPS Workshop on Human in the Loop Learning 2022

NeurIPS Workshop on Foundation Models for Decision Making 2022

NeurIPS Workshop on Offline Reinforcement Learning 2022

NeurIPS Workshop on Machine Learning Safety 2022

NeurIPS Workshop on Trustworthy and Socially Responsible Machine Learning 2022

Johns Hopkins University – MINDS and CIS Seminar Series 2022

IROS Workshop on Artificial Intelligence for Social Robots Interacting with Humans in the Real World 2022

University of Maryland, Foundations of Deep Learning Seminar 2022

AI Distinguished Lecture - Argonne National Lab 2022

Keynote – International Symposium of Robotic Research (ISRR) 2022

ICML Tutorial on Learning for Interactive Agents 2022

CVPR Workshop on Artificial Social Intelligence Workshop 2022

Stockholm Workshop on Emerging Topics in Systems and Control 2022

ICRA Workshop on Shared Autonomy in Physical Human-Robot Interaction: Adaptability and Trust 2022

ICRA Workshop on Intelligent Control Methods and Machine Learning Algorithms for Human-Robot Interaction and Assistive Robotics 2022

Simons Institute Learning and Games Workshop 2022

Cooperative AI Seminar 2022

UC Berkeley, Semiautonomous Seminar 2022

University of Toronto, Robotics Seminar 2022

AAAI Symposium on closing the assessment loop: communicating proficiency and intent in human-robot teaming 2022

University of Waterloo, Artificial Intelligence Seminar 2022

HRI Workshop on Machine Learning in Human-Robot Collaboration 2022

HRI Workshop on Human Behavior Modeling 2022

University of Pennsylvania, GRASP Seminar 2021

NeurIPS Workshop on Robot Learning. 2021

NeurIPS Workshop on Cooperative AI. 2021

NeurIPS Workshop on Learning and Strategic Behavior. 2021

CDC Workshop on Aware Learning: How to Benefit from Priors. 2021

ETH/EPFL NCCR Automation Seminar. 2021

IROS Workshop on RL-CONFORM: Reinforcement Learning meets HRI, Control, and Formal Methods. 2021

IROS Workshop on Multi-Agent and Relational Reasoning. 2021

ACL Workshop on Interactive Learning for NLP. 2021

RSS Workshop on Robotics for People. 2021

Berkeley Seminar on Multi-Agent Reinforcement Learning. 2021

ICML Workshop on Human-AI Collaboration in Sequential Decision Making. 2021

CVPR Workshop on Autonomous Driving: Perception, Prediction and Planning. 2021

Center for Human-Compatible AI Workshop. 2021

ICRA Workshop on Robot-Assisted Systems for Medical Training. 2021

ICRA Workshop on Social Intelligence in Humans and Robots. 2021

SRI Summer School on Formal Techniques. 2021

ACC Workshop on Bridging the Gap in Autonomous Vehicle Controls in Mixed Traffic. 2021

ACC Workshop on Recent Advancement of Human Autonomy Interaction and Integration. 2021

Computer Science Department Seminar, Yale. 2021

Computational Sensorimotor Learning Seminar, MIT. 2021

Center for Human-Compatible AI Seminar, UC Berkeley. 2021

ICLR Workshop on Responsible AI. 2021

RPI Department Seminar. 2021

Control Meets Learning Seminar, Caltech. 2021

AAAI Workshop on Plan, Activity, and Intent Recognition. 2021

Human-Centered AI Institute Seminar. 2021

Keynote – Conference on Robot Learning (CoRL)
Walking the Boundary of Learning and Interaction 2020

NeurIPS Workshop on Robot Learning. – " –. 2020

National Canadian Robotics Network (NCRN) Seminar. – " –. 2020

Distinguished Voices – National Academies of Sciences, Engineering, and Medicine
A Human-Centered Perspective on Interactive Robotics 2020

Panelist – AI Ethics Conference at Interdisciplinary Research Center in the Arts, Humanities, and Interpretive Social Sciences at Duke Kunshan University 2020

IPAM Workshop on Individual Vehicle Autonomy: Perception and Control.
Interaction-Aware Planning: A Human-Centered Approach toward Autonomous Driving. 2020

Keynote – 1st Colloquium on AI for Architecture, Engineering, and Construction 2020

ICML Workshop on Real-World Experiment Design & Active Learning.
Active Learning of Robot Reward Functions. 2020

RSS Workshop on Interaction and Decision-Making in Autonomous-Driving.
When our Human Modeling Assumptions Fail: Planning, learning, and prediction in near-accident driving scenarios. 2020

RSS Workshop on Power-On-and-Go Robots: Out-of-the-Box Systems for Real-World Applications.
To Ignore Humans or to Accept them with Open Arms: Challenges and Opportunities for Efficient, Robust, and Adaptive POGO Robots. 2020

RSS Workshop on AI & Its Alternatives in Assistive & Collaborative Robotics: Decoding Intent.
The Role of Learned Representations in Assistive Teleoperation. 2020

Keynote – 22nd ACM International Conference on Hybrid Systems: Computation and Control (HSCC).
Human-CPS from the Lens of Learning and Control. 2020

Keynote – Center for Human-Compatible AI Workshop. – " –. 2020

John Hopkins, Applied Physics Lab Seminar. – " –. 2020

ICRA Workshop on Long-Term Human Motion Prediction.
When our Human Modeling Assumptions Fail: The effects of risk, conventions, and non-stationarity on long-term human-robot interaction. 2020

NASA Formal Methods, AI Safety Workshop.
Risk-Aware Human Modeling. 2020

IPAM Workshop on Intersections between Control, Learning, and Optimization.
Beyond Theory of Mind: Learning & Influencing Conventions. 2020

Gilbreth Lecture, National Academy of Engineering.
Influencing Interactions in Autonomous Driving. 2020

Keynote – Formal Methods in Computer-Aided Design (FMCAD).
A journey about Safety of Autonomous Systems. 2019

Frontiers of Engineering, National Academy of Engineering.
Influencing Interactions in Autonomous Driving. 2019

RSS Workshop on Safe Autonomy. –"–. 2019

First Conference on Learning for Dynamics and Control.
Influencing Interactive Mixed-Autonomy Systems. 2019

ICML Workshop on AI for Autonomous Driving. –"–. 2019

MIT, Department Seminar.
Interactive Autonomy: Learning and Control for Human-Robot Systems. 2019

University of Washington, Department Seminar. –"–. 2019

Cornell, Department Seminar. –"–. 2019

CalTech, IST Seminar. –"–. 2019

USC, CPS Seminar. –"–. 2019

University of Maryland, Robotics Seminar. –"–. 2019

Theoretical Machine Learning Simons Foundation Workshop. –"–. 2019

Schloss Dagstuhl on Verification and Synthesis for Human-Robot Interaction.
Reward Functions and Specifications 2019

NeurIPS Workshop on Imitation Learning and its Challenges in Robotics.
Active Learning of Humans' Preferences. 2018

UAI Workshop on Safety, Risk and Uncertainty in RL. –"–. 2018

UC Berkeley, Center for Human Compatible AI. –"–. 2018

NeurIPS Workshop on Machine Learning for Intelligent Transportation Systems. Beating Congestion using Autonomous Cars. 2018

Halmstad University.
Reactive Synthesis and Human Modeling for Human-Robot Systems. 2018

University of Washington, Robotics Seminar. Safe and Interactive Robotics. 2018

UC Santa Barbara, Robotics Seminar. –"–. 2018

UC Santa Cruz, Robotics Seminar. –"–. 2018

Chinese University of Hong Kong in Shenzhen. –"–. 2018

Stanford University, Department Seminar.
Towards a Theory of Safe and Interactive Autonomy. 2017

MIT, Department Seminar. –"–. 2017

UC Berkeley, Department Seminar. –"–. 2017

CMU, Department Seminar. –"–. 2017

Princeton, Department Seminar. –"–. 2017

USC, Department Seminar. –"–. 2017

Cornell, Department Seminar. –"–. 2017

UC San Diego, Department Seminar. –"–. 2017

UC Los Angeles, Department Seminar. –"–. 2017

University of Michigan, Department Seminar. –"–. 2017

UT Austin, Department Seminar. –"–. 2017

Georgia Tech, Department Seminar. –"–. 2017

University of Pennsylvania, Department Seminar. –"–. 2017

Schloss Dagstuhl on Machine Learning and Formal Methods.
Planning for Cars that Coordinate with People. 2017

Schloss Dagstuhl on Non-Zero-Sum-Games and Control.
Correctness and Control for Human-Cyber-Physical Systems. 2015

Microsoft Research, Redmond.
Controller Synthesis for Human-in-the-Loop Systems 2014

Publications

[100] Minae Kwon, Sang Michael Xie, Kalesha Bullard, Dorsa Sadigh. Reward Design with Language Models. *International Conference on Learning Representations (ICLR), 2023.*

[99] Lorenzo Shaikewitz, Yilin Wu, Suneel Belkhale, Jennifer Grannen, Priya Sundaresan, Dorsa Sadigh. In-Mouth Robotic Bite Transfer with Visual and Haptic Sensing. *International Conference on Robotics and Automation (ICRA), 2023.*

[98] Mengxi Li, Rika Antonova, Dorsa Sadigh, Jeannette Bohg . Learning Tool Morphology for Contact-Rich Manipulation Tasks with Differentiable Simulation. *International Conference on Robotics and Automation (ICRA), 2023.*

[97] Vivek Myers, Erdem Biyik, Dorsa Sadigh. Asking Preference Questions Online in Active Reward Learning. *International Conference on Robotics and Automation (ICRA), 2023.*

[96] Yuchen Cui, Sidd Karamcheti, Raj Palleti, Nidhya Shivakumar, Percy Liang, Dorsa Sadigh. "No, to the Right" – Online Language Corrections for Robotic Manipulation via Shared Autonomy. *ACM/IEEE International Conference on Human-Robot Interaction (HRI), 2023.*

[95] Megha Srivastava, Erdem Biyik, Suvir Mirchandani, Noah Goodman, Dorsa Sadigh. Assistive Teaching of Motor Control Tasks to Humans. *Conference on Neural Information Processing Systems (NeurIPS), 2022.*

[94] Andy Shih, Dorsa Sadigh, Stefano Ermon. Training and Inference on Any-Order Autoregressive Models the Right Way. *Conference on Neural Information Processing Systems (NeurIPS), 2022.***(Oral Presentation)**

[93] Priya Sundaresan, Suneel Belkhale, Dorsa Sadigh. Learning Visuo-Haptic Skewering Strategies for Robot-Assisted Feeding. *Conference on Robot Learning (CoRL), 2022.* **(Oral Presentation)**

[92] Donald Joseph Hejna III, Dorsa Sadigh. Few-Shot Preference Learning for Human-in-the-Loop RL. *Conference on Robot Learning (CoRL), 2022.*

[91] Kanishk Gandhi, Siddharth Karamcheti, Madeline Liao, Dorsa Sadigh. Eliciting Compatible Demonstrations for Multi-Human Imitation Learning. *Conference on Robot Learning (CoRL), 2022.*

[90] Jennifer Grannen, Yilin Wu, Suneel Belkhale, Dorsa Sadigh. Learning Bimanual Scooping Policies for Food Acquisition. *Conference on Robot Learning (CoRL), 2022.*

[89] Suneel Balkhale, Dorsa Sadigh. PLATO: Predicting Latent Affordances Through Object-Centric Play. *Conference on Robot Learning (CoRL), 2022.*

[88] Behrad Toghi, Rodolfo Valiente, Dorsa Sadigh, Ramtin Pedarsani, Yaser Fallah. Social Coordination and Altruism in Autonomous Driving. *IEEE Transactions on Intelligent Transportation Systems (T-ITS), 2022.*

[87] Mark Beliaev, Andy Shih, Stefano Ermon, Dorsa Sadigh, Ramtin Pedarsani. Imitation Learning by Estimating Expertise of Demonstrators. *The 39th International Conference on Machine Learning (ICML), 2022.*

[86] Sanjit Seshia, Dorsa Sadigh, Shankar Sastry. Towards Verified Artificial Intelligence. *Communications of the ACM, 2022.*

[85] Erik Brockbank, Haoliang Wang, Justin Yang, Suvir Mirchandani, Erdem Biyik, Dorsa Sadigh, Judith Fan. How do People Incorporate Advice from Artificial Agents when Making Physical Judgments? *Cognitive Science Society Coneference (CogSci), 2022.* **(Oral Presentation)**

[84] Suneel Belkhale, Ethan Kroll Gordon, Yuxiao Chen, Siddhartha Srinivasa, Tapomayukh Bhattacharjee, Dorsa Sadigh. Balancing Efficiency and Comfort in Robot-Assisted Bite Transfer. *International Conference on Robotics and Automation (ICRA), 2022.*

[83] Zhangjie Cao, Zihan Wang, Dorsa Sadigh. Learning from Imperfect Demonstrations via Adversarial Confidence Transfer. *International Conference on Robotics and*

Automation (ICRA), 2022.

[82] Zihan Wang, Zhangjie Cao, Yilun Hao, Dorsa Sadigh. Weakly Supervised Correspondence Learning. *International Conference on Robotics and Automation (ICRA), 2022.*

[81] Zhangjie Cao, Erdem Biyik, Guy Rosman, Dorsa Sadigh. Leveraging Smooth Attention Prior for Multi-Agent Trajectory Prediction. *International Conference on Robotics and Automation (ICRA), 2022.*

[80] Andy Shih, Stefano Ermon, Dorsa Sadigh. Conditional Imitation Learning for Multi-Agent Games. *ACM/IEEE International Conference on Human-Robot Interaction (HRI), 2022.*

[79] Erdem Biyik, Aditi Talati, Dorsa Sadigh. APReL: A Library for Active Preference-based Reward Learning Algorithms. *ACM/IEEE International Conference on Human-Robot Interaction, Short Contributions (HRI), 2022.*

[78] Erdem Biyik, Anusha Lalitha, Rajarshi Saha, Andrea Goldsmith, Dorsa Sadigh. Partner-Aware Algorithms in Decentralized Cooperative Bandit Teams. *AAAI Conference on Artificial Intelligence (AAAI), 2022.*

[77] Suvir Mirchandani, Siddharth Karamcheti, Dorsa Sadigh. ELLA: Exploration through Learned Language Abstraction. *Conference on Neural Information Processing Systems (NeurIPS), 2021.*

[76] Songyuan Zhang, Zhangjie Cao, Dorsa Sadigh, Yanan Sui. Confidence-Aware Imitation Learning from Demonstrations with Varying Optimality. *Conference on Neural Information Processing Systems (NeurIPS), 2021.*

[75] Andy Shih, Dorsa Sadigh, Stefano Ermon. HyperSPNs: Compact and Expressive Probabilistic Circuits. *Conference on Neural Information Processing Systems (NeurIPS), 2021.*

[74] Shushman Choudhury, Jayesh Gupta, Mykel Kochenderfer, Dorsa Sadigh, Jeannette Bohg. Dynamic Multi-Robot Task Allocation under Uncertainty and Temporal Constraints. *Journal of Autonomous Robots (AURO), September 2021.*

[73] Woodrow Zhouyuan Wang, Andy Shih, Annie Xie, Dorsa Sadigh. Influencing Towards Stable Multi-Agent Interactions. *Conference on Robot Learning (CoRL), 2021.* **(Oral Presentation)**

[72] Zhangjie Cao, Yilun Hao, Mengxi Li, Dorsa Sadigh. Learning Feasibility to Imitate Demonstrators with Different Dynamics. *Conference on Robot Learning (CoRL), 2021.*

[71] Nils Wilde, Erdem Biyik, Dorsa Sadigh, Stephen L. Smith. Learning Reward Functions from Scale Feedback. *Conference on Robot Learning (CoRL), 2021.*

[70] Vivek Myers, Erdem Biyik, Nima Anari, Dorsa Sadigh. Learning Multimodal Rewards from Rankings. *Conference on Robot Learning (CoRL), 2021.* **(Oral Presentation)**

[69] Siddharth Karamcheti, Megha Srivastava, Percy Liang, Dorsa Sadigh. LILA: Language-Informed Latent Actions. *Conference on Robot Learning (CoRL), 2021.*

[68] Julia White, Gabriel Poesia, Robert Hawkins, Dorsa Sadigh and Noah Goodman. Open-domain clarification question generation without question examples. *Conference on Empirical Methods in Natural Language Processing (EMNLP), 2021.*

[67] Minae Kwon, Mengxi Li, Dorsa Sadigh. Influencing Leading and Following in Human-Robot Teams. *Journal of Autonomous Robots (AURO), August 2021.*

[66] Erdem Bıyık, Dylan Losey, Malayandi Palan, Nick Landolfi, Gleb Shevchuk, Dorsa Sadigh. Learning Reward Functions from Diverse Sources of Human Feedback: Optimally Integrating Demonstrations and Preferences. *The International Journal of Robotics Research (IJRR), 2021.*

[65] Dylan Losey, Hong Jun Jeon, Mengxi Li, Krishnan Srinivasan, Ajay Mandlekar, Animesh Garg, Jeannette Bohg, Dorsa Sadigh. Learning Latent Actions to Control Assistive Robots. *Journal of Autonomous Robots (AURO), 2021.*

[64] Behrad Toghi, Rodolfo Valiente, Dorsa Sadigh, Ramtin Pedarsani, Yaser Fallah. Cooperative Autonomous Vehicles that Sympathize with Humans. *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2021.*

[63] Daniel Lazar, Erdem Bıyık, Dorsa Sadigh, Ramtin Pedarsani. Learning How to Dynamically Route Autonomous Vehicles on Shared Roads. *Journal of Transportation Research Part C, 2021.*

[62] Minae Kwon, Siddharth Karamcheti, Mariano-Florentino Cuellar, Dorsa Sadigh. Targeted Data Acquisition for Evolving Negotiation Agents. *The 38th International Conference on Machine Learning (ICML), 2021.*

[61] Woodrow Wang, Mark Beliaev, Erdem Biyik, Daniel Lazar, Ramtin Pedarsani, Dorsa Sadigh. Emergent Prosociality in Multi-Agent Games Through Gifting. *The 30th International Joint Conference on Artificial Intelligence (IJCAI), 2021.*

[60] Andy Shih, Arjun Sawhney, Jovana Kondic, Stefano Ermon, Dorsa Sadigh. On the Critical Role of Conventions in Adaptive Human-AI Collaboration. *International Conference on Learning Representations (ICLR), 2021.*

[59] Mengxi Li, Alper Canberk, Dylan Losey, Dorsa Sadigh. Learning Human Objectives from Sequences of Physical Corrections. *International Conference on Robotics and Automation (ICRA), 2021.*

[58] Kejun Li, Maegan Tucker, Erdem Bıyık, Ellen Novoseller, Joel Burdick, Yanan Sui, Dorsa Sadigh, Yisong Yue, Aaron Ames. ROIAL: Region of Interest Active Learning for Characterizing Exoskeleton Gait Preference Landscapes. *International Conference on Robotics and Automation (ICRA), 2021.*

[57] Zhangjie Cao, Minae Kwon, Dorsa Sadigh. Transfer Reinforcement Learning across Homotopy Classes. *IEEE Robotics and Automation Letters (RAL), 2021.*

[56] Zhangjie Cao, Dorsa Sadigh. Learning from Imperfect Demonstrations with Varying Dynamics. *IEEE Robotics and Automation Letters (RAL), 2021.*

[55] Siddharth Karamcheti, Albert Zhai, Dylan Losey, Dorsa Sadigh. Learning Visually Guided Latent Actions for Assistive Teleoperation. *3rd Learning for Dynamics & Control Conference (L4DC), 2021.*

[54] Mark Beliaev, Erdem Bıyık, Daniel Lazar, Woodrow Wang, Dorsa Sadigh, Ramtin Pedarsani. Incentivizing Routing Choices for Safe and Efficient Transportation in the Face of the COVID-19 Pandemic. *12th ACM/IEEE International Conference on Cyber-Physical Systems (ICCPS), May 2021.*

[53] Erdem Bıyık, Daniel A. Lazar, Ramtin Pedarsani, Dorsa Sadigh. Incentivizing Efficient Equilibria in Traffic Networks with Mixed Autonomy. *IEEE Transactions on Control of Network Systems (TCNS), 2020.*

[52] Hadas Kress-Gazit, Kerstin Eder, Guy Hoffman, Henny Admoni, Brenna Argall, Ruediger Ehlers, Christoffer Heckman, Nils Jansen, Ross Knepper, Jan Kretinsky, Shelly Levy-Tzedek, Jamy Li, Todd Murphey, Laurel Riek, Dorsa Sadigh. Formalizing and Guaranteeing* Human-Robot Interaction. *Communications of the ACM, 2020.*

[51] Annie Xie, Dylan Losey, Ryan Tolsma, Chelsea Finn, Dorsa Sadigh. Learning Latent Representations to Influence Multi-Agent Interaction. *Proceedings of the 4th Conference on Robot Learning (CoRL), November 2020.* **(Oral Presentation, Best Paper Award)**

[50] Robert X. D. Hawkins, Minae Kwon, Dorsa Sadigh, Noah D. Goodman. Continual Adaptation for Efficient Machine Communication. *The 24th Conference on Computational Natural Language Learning (CoNLL), November 2020.*

[49] Mengxi Li, Dylan Losey, Jeannette Bohg, Dorsa Sadigh. Learning User-Preferred Mappings for Intuitive Robot Control. *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), October 2020.*

[48] Zheqing Zhu, Erdem Bıyık, Dorsa Sadigh. Multi-Agent Safe Planning with Gaussian Processes. *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), October 2020.*

[47] Jonathan Mern, Dorsa Sadigh, Mykel Kochenderfer. Object Exchangability in Reinforcement Learning. *2020 American Control Conference (ACC), July 2020.*

[46] Hong Jun Jeon, Dylan Losey, Dorsa Sadigh. Shared Autonomy with Learned Latent Actions. *Robotics: Science and Systems (RSS), June 2020.* **(Best Student Paper Award, Finalist)**

[45] Erdem Bıyık, Nicolas Huynh, Mykel Kochenderfer, Dorsa Sadigh. Active Preference-Based Gaussian Process Regression for Reward Learning. *Robotics: Science and Systems (RSS), June 2020.*

[44] Zhangjie Cao, Erdem Bıyık, Woodrow Wang, Allan Raventos, Adrien Gaidon, Guy Rosman, Dorsa Sadigh. Reinforcement Learning based Control of Imitative Policies for Near-Accident Driving. *Robotics: Science and Systems (RSS), June 2020.*

[43] Shushman Choudhury, Jayesh Gupta, Mykel Kochenderfer, Dorsa Sadigh, Jeannette Bohg. Dynamic Multi-Robot Task Allocation under Uncertainty and Temporal Constraints. *Robotics: Science and Systems (RSS), June 2020.*

[42] Malayandi Palan, Shane Barratt, Alex McCauley, Dorsa Sadigh, Vikas Sindhwani, Stephen P. Boyd. Fitting a Linear Control Policy to Demonstrations with a Kalman Constraint. *2nd Learning for Dynamics & Control Conference (L4DC), June 2020.*

[41] Minae Kwon, Erdem Bıyık, Aditi Talati, Karan Bhasin, Dylan P. Losey, Dorsa Sadigh. When Humans Aren't Optimal: Robots that Collaborate with Risk-Aware Humans. *ACM/IEEE International Conference on Human-Robot Interaction (HRI), 2020.* **(Best Paper Award, Honorable Mention)**

[40] Yuhang Che, Allison M. Okamura, Dorsa Sadigh. Efficient and Trustworthy Social Navigation Via Explicit and Implicit Robot-Human Communication. *IEEE Transactions on Robotics (TRO), 2019.*

[39] Dylan P. Losey, Krishnan Srinivasan, Ajay Mandlekar, Animesh Garg, Dorsa Sadigh. Controlling Assistive Robots with Learned Latent Actions. *International Conference on Robotics and Automation (ICRA), May 2020.*

[38] Dylan P. Losey, Mengxi Li, Jeannette Bohg, Dorsa Sadigh. Learning from My Partner's Actions: Roles in Decentralized Robot Teams. *Conference on Robot Learning (CoRL), 2019.* **(Oral Presentation)**

[37] Erdem Bıyık, Malayandi Palan, Nicholas Landolfi, Dylan P. Losey, Dorsa Sadigh. Asking Easy Questions: A User-Friendly Approach to Active Reward Learning. *Conference on Robot Learning (CoRL), 2019.*

[36] Dylan P. Losey, Dorsa Sadigh. Robots that Take Advantage of Human Trust. *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), November 2019.*

[35] Chandrayee Basu, Erdem Bıyık, Zhixun He, Mukesh Singhal, Dorsa Sadigh. Active Learning of Reward Dynamics from Hierarchical Queries. *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), November 2019.*

[34] Erdem Bıyık, Daniel A. Lazar, Dorsa Sadigh, Ramtin Pedarsani. The Green Choice: Learning and Influencing Human Decisions on Shared Roads. *Proceedings of the 58th IEEE Conference on Decision and Control (CDC), December 2019.*

[33] Minae Kwon, Mengxi Li, Alexandre Bucquet, Dorsa Sadigh. Influencing Leading and Following in Human-Robot Teams. *Robotics: Science and Systems (RSS), June 2019.*

[32] Malayandi Palan, Gleb Shevchuk, Nicholas C. Landolfi, Dorsa Sadigh. Learning Reward Functions by Integrating Human Demonstrations and Preferences. *Robotics: Science and Systems (RSS), June 2019.*

[31] Tianhe Yu, Gleb Shevchuk, Dorsa Sadigh, Chelsea Finn. Unsupervised Visuomotor Control through Distributional Planning Networks. *Robotics: Science and Systems (RSS), June 2019.*

[30] Erdem Bıyık, Jonathan Margoliash, Shahrouz Ryan Alimo, Dorsa Sadigh. Efficient and Safe Exploration in Deterministic Markov Decision Processes with Unknown Transition Models. *2019 American Control Conference (ACC), July 2019.*

[29] Elis Stefansson, Jaime Fisac, Dorsa Sadigh, Shankar Sastry, Karl H. Johansson. Human-Robot Interaction for Truck Platooning Using Hierarchical Dynamic Games. *European Control Conference (ECC), June 2019.* **(Best Paper Award, Finalist).**

[28] Ashwini Pokle, Roberto Martin-Martin, Patrick Goebel, Vincent Chow, Hans M. Ewald, Junwei Yang, Zenkai Wang, Amir Sadeghian, Dorsa Sadigh, Silvio Savarese, Marynel Vazquez. Deep Local Trajectory Planning and Control for Robot Navigation. *International Conference on Robotics and Automation (ICRA), May 2019.*

[27] Jaime F. Fisac, Eli Bronstein, Elis Stefansson, Dorsa Sadigh, S. Shankar Sastry, Anca D. Dragan. Hierarchical Game-Theoretic Planning for Autonomous Vehicles. *International Conference on Robotics and Automation (ICRA), May 2019.*

[26] Erdem Bıyık, Dorsa Sadigh. Batch Active Preference-Based Learning of Reward Functions. *Conference on Robot Learning (CoRL), 2018.* **(Oral Presentation)**

[25] Erdem Bıyık, Daniel A. Lazar, Ramtin Pedarsani, Dorsa Sadigh. Altruistic Autonomy: Beating Congestion on Shared Roads . *International Workshop on Algorithmic Foundations of Robotics (WAFR), 2018.*

[24] Daniel Lazar, Kabir Chandrasekher, Ramtin Pedarsani, Dorsa Sadigh. Maximizing Road Capacity Using Cars that Influence People. *IEEE Conference on Decision and Control (CDC), 2018.*

[23] Jiaming Song, Hongyu Ren, Dorsa Sadigh, Stefano Ermon. Multi-Agent Generative Adversarial Imitation Learning. *Conference on Neural Information Processing Systems (NeurIPS), 2018.*

[22] Dorsa Sadigh, S. Shankar Sastry, Sanjit Seshia. Verifying Robustness of Human-Aware Autonomous Cars . *IFAC conference on Cyber-Physical and Human Systems (CPHS), 2018.*

[21] Dorsa Sadigh, Nick Landolfi, S. Shankar Sastry, Sanjit A. Seshia, Anca Dragan. Planning for Autonomous Cars that Leverages Effects on Human Actions. *Journal of Autonomous Robots (AURO), 2018.*

[20] Susmit Jha, Vasumathi Raman, Dorsa Sadigh, Sanjit A. Seshia. Safe Autonomy Under Perception Uncertainty Using Chance-Constrained Temporal Logic . *Journal of Automatic Reasoning (JAR), 2018.*

[19] Dorsa Sadigh. Safe and Interactive Autonomy: Control, Learning, and Verification. *Ph.D. Dissertation. EECS Department, University of California, Berkeley, August 2017.*

[18] Dorsa Sadigh, S. Shankar Sastry, Sanjit Seshia, Anca Dragan. Active Preference-Based Learning of Reward Functions. *Robotics: Science and Systems Conference (RSS), July 2017.*

[17] Negar Mehr, Dorsa Sadigh, Roberto Horowitz, S. Shankar Sastry, Sanjit Seshia. Stochastic Predictive Freeway Ramp Metering from Signal Temporal Logic Specifications. *American Control Conference (ACC), May 2017.*

[16] Dorsa Sadigh, S. Shankar Sastry, Sanjit Seshia, Anca Dragan. Information Gathering Actions over Human Internal State. *International Conference on Intelligent Robots*

and Systems (IROS), 2016. **(Best Paper in Cognitive Robotics Award, Finalist).**

[15] Tara Rezvani, Katherine Driggs-Campbell, Dorsa Sadigh, S. Shankar Sastry, Sanjit Seshia, Ruzena Bajcsy. Towards Trustworthy Automation: User Interfaces that Convey Internal and External Awareness. *IEEE Intelligent Transportation Systems Conference (ITSC), November 2016.*

[14] Dorsa Sadigh, S. Shankar Sastry, Sanjit Seshia, Anca Dragan Planning for Autonomous Cars that Leverages Effects on Human Actions . *Robotics: Science and Systems Conference (RSS), 2016.*

[13] Dorsa Sadigh, Ashish Kapoor. Safe Control under Uncertainty with Probabilistic Signal Temporal Logic. *Robotics: Science and Systems Conference (RSS), 2016.*

[12] Shromona Ghosh, Dorsa Sadigh, Pierluigi Nuzzo, Vasumathi Raman, Alexandre Donze, Alberto Sangiovanni-Vincentelli, S. Shankar Sastry, Sanjit Seshia. Diagnosis and Repair for Synthesis from Signal Temporal Logic Specifications. *Conference on Hybrid Systems: Computation and Control (HSCC), 2016.*

[11] Sanjit A. Seshia, Dorsa Sadigh, S. Shankar Sastry. Formal Methods for Semi-autonomous Driving. *Design and Automation Conference (DAC), 2015.*

[10] Vasumathi Raman, Alexandre Donze, Dorsa Sadigh, Richard M. Murray, Sanjit Seshia. Reactive Synthesis from Signal Temporal Logic Specifications. *Conference on Hybrid Systems: Computation and Control (HSCC), 2015.*

[9] Dorsa Sadigh, Eric S. Kim, Samuel Coogan, S. Shankar Sastry, Sanjit Seshia. A Learning Based Approach to Control Synthesis of Markov Decision Processes for Linear Temporal Logic Specifications. *IEEE Conference on Decision and Control (CDC), 2014.*

[8] Dorsa Sadigh, Henrik Ohlsson, S. Shankar Sastry, Sanjit Seshia. Robust Subspace System Identification via Weighted Nuclear Norm Optimization. *International Federation of Automatic Control (IFAC), 2014.*

[7] Dorsa Sadigh, Katherine Driggs-Campbell, Alberto Puggelli, Wenchao Li, Victor Shia, Ruzena Bajcsy, Alberto Sangiovanni-Vincentelli, Shankar Sastry, and Sanjit Seshia. Data-driven probabilistic modeling and verification of human driver behavior. *Formal Verification and Modeling in Human-Machine Systems (AAAI Spring Symposium), 2014.*

[6] Dorsa Sadigh, Katherine Driggs Campbell, Ruzena Bajcsy, S. Shankar Sastry, Sanjit Seshia. User Interface Design and Verification for Semi-autonomous Driving. *Conference on High Confidence Networked Systems, 2014.*

[5] Ashish Tiwari, Bruno Dutertre, Dejan Jovanovic, Thomas de Candia, Dorsa Sadigh, Sanjit Seshia. Safety Envelop in Security. *Conference on High Confidence Networked Systems (HiCoNS), 2014.*

[4] Wenchao Li, Dorsa Sadigh, S. Shankar Sastry, Sanjit Seshia. Synthesis for Human-in-the-Loop Control Systems. *Tools and Algorithms for the Construction and Analysis of Systems (TACAS), 2014.*

[3] Dorsa Sadigh, Sanjit Seshia and Mona Gupta. Automating Exercise Generation: A Step towards Meeting the MOOC Challenge for Embedded Systems. *Workshop on*

Embedded Systems Education, 2012.

[2] Orna Kupferman, Dorsa Sadigh, and Sanjit A. Seshia. Synthesis with Clairvoyance. *Haifa Verification Conference (HVC), 2011.*

[1] Jonathan Kotker, Dorsa Sadigh, and Sanjit A. Seshia. Timing Analysis of Interrupt-Driven Programs under Context Bounds. *Formal Methods in Computer Aided Design (FMCAD), 2011.*

Technical Reports & Workshop Papers

[14] Bommasani et al. On the Opportunities and Risks of Foundation Models.

[13] Bidipta Sarkar*, Aditi Talati*, Andy Shih*, Dorsa Sadigh. PantheonRL: A MARL Library for Dynamic Training Interactions. *Proceedings of the 36th AAAI Conference on Artificial Intelligence (Demo Track), February 2022.*

[12] Erdem Biyik, Anusha Lalitha, Rajarshi Saha, Andrea Goldsmith, Dorsa Sadigh. Partner-Aware Algorithms in Decentralized Cooperative Bandit Teams. *Artificial Intelligence for Human-Robot Interaction (AI-HRI) at AAAI Fall Symposium Series, November 2021.*

[11] Erdem Biyik, Aditi Talati, Dorsa Sadigh. APReL: A Library for Active Preference-based Reward Learning Algorithms. *Artificial Intelligence for Human-Robot Interaction (AI-HRI) at AAAI Fall Symposium Series, November 2021.*

[10] Nicholas Roy, Ingmar Posner, Tim Barfoot, Philippe Beaudoin, Yoshua Bengio, Jeannette Bohg, Oliver Brock, Isabelle Depatie, Dieter Fox, Dan Koditschek, Tomas Lozano-Perez, Vikash Mansinghka, Christopher Pal, Blake Richards, Dorsa Sadigh, Stefan Schaal, Gaurav Sukhatme, Denis Therien, Marc Toussaint, Michiel Van de Panne. From Machine Learning to Robotics: Challenges and Opportunities for Embodied Intelligence. *arXiv, November 2021.*

[9] Andy Shih, Dorsa Sadigh, Stefano Ermon. HyperSPNs: Compact and Expressive Probabilistic Circuits. *The 4th Workshop on Tractable Probabilistic Modeling, 2021.*

[8] Behrad Toghi, Rodolfo Valiente, Dorsa Sadigh, Ramtin Pedarsani, Yaser Fallah. Altruistic Maneuver Planning for Cooperative Autonomous Vehicles Using Multi-agent Advantage Actor-Critic. *CVPR Workshop on Autonomous Driving: Perception, Prediction and Planning, 2021.*

[7] Mark Beliaev, Woodrow Z. Wang, Daniel A. Lazar, Erdem Bıyık, Dorsa Sadigh, Ramtin Pedarsani. Emergent Correlated Equilibrium through Synchronized Exploration. *RSS 2020 Workshop on Emergent Behaviors in Human-Robot Systems, 2020.*

[6] Kawin Ethayarajh, Dorsa Sadigh. BLEU Neighbors: A Reference-less Approach to Automatic Evaluation. *1st Workshop on Evaluation and Comparison for NLP systems (Eval4NLP), 2020.*

[5] Siddharth Karamcheti, Dorsa Sadigh and Percy Liang. Continual Learning Adaptive Language Interfaces through Decomposition. *Proceedings of the EMNLP Workshop on Interactive and Executable Semantic Parsing, 2020.*

[4] Robert X. D. Hawkins, Minae Kwon, Dorsa Sadigh, Noah D. Goodman. Continual Adaptation for Efficient Machine Communication. *Proceedings of the ICML Workshop on Adaptive & Multitask Learning: Algorithms & Systems, June 2019.* **(Best Paper Award).**

[3] Jiaming Song, Hongyu, Ren, Dorsa Sadigh, Stefano Ermon. Multi-Agent Generative Adversarial Imitation Learning. *International Conference on Learning Representations (ICLR), Workshop Track, April 2018.*

[2] Sanjit Seshia, Dorsa Sadigh, S. Shankar Sastry. Towards Verified Artificial Intelligence. *Technical Report, July 2016.*

[1] Debadeepta Dey, Dorsa Sadigh, Ashish Kapoor. Fast Safe Mission Plans for Autonomous Vehicles. *Proceedings of Robotics: Science and Systems Workshop, June 2016.*

Dissertation

Dorsa Sadigh. Safe and Interactive Autonomy: Control, Learning, and Verification. *Ph.D. Dissertation; EECS Department, University of California, Berkeley, August 2017.*