Anca Dragan

UC Berkeley EECS Department
2121 Berkeley Way Room #8042
Berkeley, CA 94720-1660
anca@berkeley.edu
www.ancadragan.com

Current Positions

Associate Professor, *UC Berkeley*, EECS & Psychology (by courtesy). 2021-present

Assistant Professor, *UC Berkeley*, EECS & Psychology (by courtesy). 2015-2021

Consultant (Staff Research Scientist), *Waymo*, Alphabet. 2017-present

Former

PhD, Robotics, *Carnegie Mellon University*, USA. 2009-2015
Advisor: Siddhartha Srinivasa. *"Legible Robot Motion Planning".*

B.Sc., Computer Science, *Jacobs University Bremen*, Germany. 2006-2009
Advisors: Michael Kohlhase and Herbert Jaeger.

Honors

IEEE RAS Early Career Award. 2021
Citation: "For pioneering algorithmic human-robot interaction"

ONR Young Investigator Award. 2020
"A Unified Framework for Inferring Rewards from Diverse Types of Human Feedback"

McEntyre Award for Excellence in Teaching. 2020

PECASE Winner. 2019
Presidential Early Career Award for Science and Engineering

IJCAI Early Career Spotlight. 2018

Sloan Research Fellow. 2018
Alfred P. Sloan Foundation

TR 35. 2017
MIT Tech Review 35 Innovators under 35

Okawa Foundation Award. 2017
Awarded to 9 faculty in the United States

NSF CAREER Award. 2017
"Towards Autonomously Generating Robot Behavior for Coordination with Humans – Accounting for Effects on Human Actions "

SCS Dissertation Award Honorable Mention. 2015
For "Legible Robot Motion Planning"

Rising Stars in EECS. 2014
Awarded to 40 EECS graduate and postdoctoral women.

Siebel Scholar. 2014
For academic excellence and demonstrated leadership.

Best Reviewer Award Finalist. 2014
Robotics: Science and Systems

Dan David Scholarship. 2014
For the Future Direction of Artificial Intelligence, 2014.

CITEC Award for Excellence in Doctoral HRI Research. 2014
At the International Conference on Human-Robot Interaction in 2014.

Intel PhD Fellow. 2013
I was one of the 14 students who were awarded the Intel PhD Fellowship.

Google Anita Borg Scholar. 2012
I was one of the 25 students in the U.S. who were awarded the Google Anita Borg Memorial Scholarship.

HRI Pioneer. 2011
I was selected to participate in the Human-Robot Interaction Pioneers Workshop, a highly selective workshop seeking to foster creativity, communication, and collaboration across HRI.

Paper Honors

Best Paper Finalist, *IEEE|ACM Human Robot Interaction (HRI).* 2021
"Feature-Expansive Reward Learning: Rethinking Human Input"

Best Paper Honorable Mention, *IEEE TRO* . 2020
"Quantifying Hypothesis Space Misspecification in Learning from Human-Robot Demonstrations and Physical Corrections"

Best Paper Award, *IEEE|ACM Human Robot Interaction (HRI).* 2020
"LESS is More: Rethinking Probabilistic Models of Human Behavior"

Best Paper Finalist, *IEEE|ACM Human Robot Interaction (HRI).* 2018
"Expressing Robot Incapability"

Best Bluesky Paper Finalist, *International Symposium on Robotics Research (ISRR).* 2018
"Pragmatic Pedagogic Value Alignment"

Best Cognitive Robotics Paper Finalist, *IEEE Intelligent Robots and Systems (IROS).* 2016
"Active Information Gathering over Human Internal State"

Best HRI Paper Finalist, *IEEE International Conference on Robotics and Automation (ICRA).* 2016
"Reducing Supervisor Burden in Online Learning from Demonstration"

Best Paper Finalist, *IEEE International Conference on Robotics and Automation (ICRA).* 2015
"Motion Primitives via Optimization"

Best Paper Award Finalist, *Robotics: Science and Systems (RSS).* 2013
"Generating Legible Motion"

Best Paper Award Finalist, *Robotics: Science and Systems (RSS).* 2012
"Formalizing Assistive Teleoperation"

Best Paper Award Nomination, *International Symposium on Human-Robot Communication (RoMan).* 2012

"Online Customization of Teleoperation Interfaces"

Alumni

Graduate Students, *Dorsa Sadigh (Faculty at Stanford), Sandy Huang (Research Scientist at Deepmind), Jaime Fisac (Faculty at Princeton), Dylan Hadfield-Menell (Faculty at MIT), Rohin Shah (Research Scientist at Deepmind), Andrea Bajcsy (Faculty at CMU), Smitha Milli (postodc at Cornell), Kush Bhatia (postdoc at Stanford), Sid Reddy (Facebook Reality Labs), Andreea Bobu (Faculty at MIT).*

Selected Undergraduate Students, *Gaurav Ghosal (went on to PhD at CMU), Arjun Sripathy (went on to ML Scientist at Tesla), Micah Carroll (went on to PhD at Berkeley), Matthew Zurek (went on to PhD at U Wisconsin Madison), Gokul Swamy (went on to PhD at CMU), Sampada Deglurkar (went on to PhD at UC Berkeley), Ravi Panya (went on to PhD at CMU), Hong Jun Jeon (CRA finalist, went on to PhD at Stanford), Nick Landolfi (went on to PhD at Stanford), Allan Zhou (went on to PhD at Stanford), Andy Palaniappan (went on to PhD at Stanford), Jason Zhang (went on to PhD at CMU), Minae Kwon (went on to PhD at Stanford), Lawrence Chan (went on to PhD at Berkeley), Smitha Milli (CRA finalist, went on to PhD at Berkeley), Glen Chao (went on to PhD at U. Michigan), Rachel Holladay (went on to PhD at MIT), Kenton Lee (went on to PhD from UW).*

Teaching

Algorithmic Human-Robot Interaction (CS287H), *UC Berkeley.* 2015, 2016, 2017, 2020, 2021, 2023

Instructor.

Human-Compatible AI (CS294-125), *UC Berkeley.* 2016

Co-instructor.

Introduction to Artificial Intelligence (CS188), *UC Berkeley.* 2016, 2017,2018,2019, 2020, 2021

Instructor/Co-Instructor.

Manipulation Algorithms, *Carnegie Mellon University.* 2014

Co-instructor.

Mathematical Fundamentals for Robotics, *Carnegie Mellon University.* 2011

TA for Prof. Michael Erdmann.

Computability and Complexity, *Jacobs University Bremen.* 2009

TA for Prof. Herbert Jaeger.

General Computer Science I and II, *Jacobs University Bremen.* 2007-2009

TA for Prof. Michael Kohlhase.

General Electrical Engineering I and II, *Jacobs University Bremen.* 2007-2008

TA for Prof. Werner Bergholz.

Outreach

Berkeley AI4ALL Yearly Summer Camp, *Founded and ran a yearly week-long summer camp for high school students from underserved communities. The camp brings 23-25 students each summer to the Berkeley campus to teach them about human-centered AI. The camp is running yearly since 2016, with help from AI4ALL and Lawrence Hall of Science. .*

InterACT Summer Internship, *Founded a lab internship program offered yearly to one Bay Area underrepresented high-schooler; also hosting REU students from the BAIR and SUPERB REU programs.*

Lectures on Robotics, CS, and Math, *Carlmont High, SAILORS, Ellis School for Girls, Hawken School, Leap@CMU, Carnegie Science Center, Wilkinsburg Gifted Class.*

Talks at Women in STEM events and panels, *SWE, Fem Tech, Women in Tech SF Summit, WICSE.*

Research Team Leader, *OurCS: Opportunities for undergraduate research in Computer Science.*

Talks to Berkeley Undergraduates on Integrating Interaction into Robotics, *EECS Honors, HKN General Meeting, Transfer Students Breakfast with Faculty, etc..*

Lab Tours, *Tours and demos to the general public, particularly to children and young adults.*

Professional Activities - Robotics, HRI, and Machine Learning

Executive Board, Director: *Conference on Robot Learning*,2021-2024.

Program Chair: *Conference on Robot Learning*, 2018.

Chair: *Bay Area Robotics Symposium*, 2016 and 2017.

Associate Editor (or equivalent): ACM Transactions on HRI (Computational HRI track), AURO (special issue), CORL 2019, ICRA 2017, HRI 2016, 2018, 2019, WAFR 2016, IROS 2016, ARSO 2014.

Workshops Chair: *Robotics: Science and Systems*, 2017.

Workshop Organizer: *Robot Learning, Autonomous Driving, Interactive Learning from Human Feedback, Algorithms for Human-Robot Interaction, Human-Robot Collaboration, Planning for Human-Robot Interaction.* NeurIPS/ICML/RSS/HRI.

Professional Activities - Berkeley

BAIR Steering Committee.: I helped found and am on the steering committee of the Berkeley AI Research Lab (BAIR); `http://bair.berkeley.edu`.

Co-PI for the Center on Human-Compatible AI.: Our mission is AI that is built for scratch to be beneficial to humanity; `http://humancompatible.ai`

AI Admissions Chair: I managed a process involving 5 faculty and 40 graduate students to review 2000 PhD applications. 2017,2018,2019,2020.

Space Planning for College of Computing, Data, and Society: CDSS Gateway, 2023

CS Admissions Reform Chair: I helped coordinate the department and campus leadership through a solution for reforming L&S CS admissions.

AI Space Committee: I co-led the design of the Berkeley Way West AI floor, along with space assignments and occupancy policies. I am also on the AdHoc committee for the new division building.

AI Prelim Examiner: 2016, 2017,2019,2020,2021.

Berkeley Open Research Commons Board: I serve on the board and manage the collaboration between BAIR and Microsoft.

Invited Talk Highlights

University of Washington, *The Lytle Lecture*. 2022

Robotics Today, *Optimizing Intended Reward Functions*. 2020

Lex Fridman AI Podcast, *Human-Robot Interaction, Reward Engineering, and IRL*. 2020

WIRED 25 Summit, *A Glimpse at the Challenges in Human-Robot Interaction*. 2019

Baylearn Keynote, *Learning Intended Rewards: Extracting all the right information from all the right places*. 2019

ICAPS Keynote, *Planning for Human-Robot Interaction*. 2019

Apple ML Summit Keynote, *An Optimization-Based Theory of Mind for Human-Robot Interaction*. 2019

Microsoft Research AI Distinguished Lectures, -"-. 2019

IROS Keynote, *Optimal Robot Action for and around People*. 2018

IJCAI Early Career Spotlight, -"-. 2018

CoRL Keynote, *Putting Humans into the Robot Equation*. 2017

Peer-Reviewed Publications (Conferences and Journals)

[1] J. Hong, S. Levine, and A.D. Dragan. Learning to influence human behavior with offline reinforcement learning. In *Neural Information Processing Systems (NeurIPS)*, 2023.

[2] J. Gao, S. Reddy, G. Berseth, A.D. Dragan, and S. Levine. Bootstrapping adaptive human-machine interfaces with offline reinforcement learning. In *International Conference on Intelligent Robots and Systems (IROS)*, 2023.

[3] V. Myers, A. He, K. Fang, H. Walke, P. Hansen-Estruch, C. Cheng, M. Jalobeanu, A. Kolobov, A.D. Dragan, and S. Levine. Goal representations for instruction following: A semi-supervised language interface to control. In *Conference on Robot Learning (CoRL)*, 2023.

[4] E. Jones, A.D. Dragan, A. Raghunathan, and J. Steinhardt. Automatically auditing large language models via discrete optimization. In *International Conference on Machine Learning (ICML)*, 2023.

[5] J. Hong, K. Bhatia, and A.D. Dragan. On the sensitivity of reward inference to misspecified human models. In *International Conference on Learning Representations (ICLR)*, 2023.

[6] J. Tien, J.Z.Y. He, Z. Erickson, A.D. Dragan, and D.S. Brown. Causal confusion and reward misidentification in preference-based reward learning. In *International Conference on Learning Representations (ICLR)*, 2023.

[7] A. Bobu, Y. Liu, R. Shah, D. S. Brown, and A. D. Dragan. Similarity-based implicit representation learning. In *International Conference on Human-Robot Interaction (HRI)*, 2023.

[8] R. Tian, M. Tomizuka, A.D. Dragan, and A. Bajcsy. Towards modeling and influencing the dynamics of human learning. In *International Conference on Human-Robot Interaction (HRI)*, 2023.

[9] G.R. Ghosal, M. Zurek, D.S. Brown, and A.D. Dragan. The effect of modeling human rationality level on learning rewards from multiple feedback types. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2023. **(oral)**.

[10] D. Shin, A. D. Dragan, and D. S. Brown. Benchmarks and algorithms for offline preference-based reward learning. *Transactions on Machine Learning Research (TMLR)*, 2023.

[11] J.Z.Y. He, A. Raghunathan, D.S. Brown, Z. Erickson, and A.D. Dragan. Learning representations that enable generalization in assistive tasks. In *Conference on Robot Learning (CORL)*, 2022.

[12] J. Lin R. Georgescu M. Sun D. Bignell S. Milani K. Hofmann M. Hausknecht A.D. Dragan S. Devlin M. Carroll, O. Paradise. Uni[mask]: Unified inference in sequential decision problems. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2022. **(oral)**.

[13] S. Reddy, S. Levine, and A.D. Dragan. First contact: Unsupervised human-machine co-adaptation via mutual information maximization. In *Neural Information Processing Systems (NeurIPS)*, 2022.

[14] A. Sripathy, A. Bobu, Z. Li, K. Sreenath, D.S. Brown, and A.D. Dragan. Teaching robots to span the space of functional expressive motion. In *International Conference on Intelligent Robots and Systems (IROS)*, 2022.

[15] M. Carroll, D. Hadfield-Menell, S. Russell, and A.D. Dragan. Estimating and penalizing induced preference shifts in recommender systems. In *International Conference on Machine Learning (ICML)*, 2022.

[16] R. Tian, L. Sun, A. Bajcsy, M. Tomizuka, and A.D. Dragan. Safety assurances for human-robot interaction via confidence-aware game-theoretic human models. In *International Conference on Robotics and Automation (ICRA)*, 2022.

[17] S. Chen*, J. Gao*, S. Reddy, G. Berseth, A.D. Dragan, and S. Levine. Asha: Assistive teleoperation via human-in-the-loop reinforcement learning. In *International Conference on Robotics and Automation (ICRA)*, 2022.

[18] J. Lin, D. Fried, D. Klein, and A.D. Dragan. Inferring rewards from language in context. In *Association for Computational Linguistics (ACL)*, 2022.

[19] C. Laidlaw and A.D. Dragan. The boltzmann policy distribution: Accounting for systematic suboptimality in human models. In *International Conference on Learning Representations (ICLR)*, 2022.

[20] A. Bobu, M. Wiggert, C. Tomlin, and A. D. Dragan. Inducing structure in reward learning by learning features. *International Journal of Robotics Research*, 2022.

[21] R. Shah, C. Wild, S. H. Wang, N. Alex, B. Houghton, W. Guss, S. Mohanty, A. Kanervisto, S. Milani, N. Topin, P. Abbeel, S. Russell, and A. Dragan. The minerl basalt competition on learning from human feedback. In *Neural Information Processing Systems, Competition Track (NeurIPS)*, 2021.

[22] S. Reddy, A.D. Dragan, and S. Levine. Pragmatic image compression for human-in-the-loop decision-making. In *Neural Information Processing Systems (NeurIPS)*, 2021. **(spotlight talk)**.

[23] K. Lee, L. Smith, A.D. Dragan, and P. Abbeel. B-pref: Benchmarking preference-based reinforcement learning. In *Neural Information Processing Systems (NeurIPS)*, 2021.

[24] D. Losey, A. Bajcsy, M. O'Malley, and A.D. Dragan. Physical interaction as communication: Learning robot objectives online from human corrections. *International Journal of Robotics Research (IJRR)*, 2021.

[25] A.D. Dragan J.Z. He. Assisted robust reward design. In *Conference on Robot Learning (CoRL)*, 2021.

[26] Z. Javed, D.S. Brown, S. Sharma, J Zhu, A. Balakrishna, M. Petrik, A.D. Dragan, and K. Goldberg. Policy gradient bayesian robust optimization for imitation learning. In *International Conference on Machine Learning (ICML)*, 2021.

[27] D.S. Brown, J. Schneider, A.D. Dragan, and S. Niekum. Value alignment verification. In *International Conference on Machine Learning (ICML)*, 2021.

[28] L. Sun, X. Jia, and A.D. Dragan. On complementing end-to-end human behavior predictors with planning. In *Robotics: Science and Systems (RSS)*, 2021.

[29] O. Watkins, S. Huang, J. Frost, K. Bhatia, E. Weiner, P. Abbeel, T. Darrell, B. Plummer, K. Saenko, and A.D. Dragan. Explaining robot policies. *Applied AI Letters (AAIL)*, 2021.

[30] A. Jain, L. Chan, D.S. Brown, and A.D. Dragan. Optimal cost design for model predictive control. In *Learning for Dynamics and Control (L4DC)*, 2021.

[31] A. Bajcsy, A. Siththaranjan, C.J. Tomlin, and A.D. Dragan. Analyzing human models that adapt online. In *International Conference on Robotics and Automation (ICRA)*, 2021.

[32] A. Sripathy, A. Bobu, D.S. Brown, and A.D. Dragan. Dynamically switching human prediction models for efficient planning. In *International Conference on Robotics and Automation (ICRA)*, 2021.

[33] M. Zurek, A. Bobu, D.S. Brown, and A.D. Dragan. Situational confidence assistance for lifelong shared autonomy. In *International Conference on Robotics and Automation (ICRA)*, 2021.

[34] S. Devlin K. Ciosek K. Hofmann A.D. Dragan R. Shah P. Knott, M. Carroll. Evaluating the robustness of collaborative agents. In *Autonomous Agents and Multiagent Systems (AAMAS)*, 2021.

[35] J. Gao, S. Reddy, G. Berseth, N. Hardy, N. Natraj, K. Ganguly, A. D. Dragan, and S. Levine. X2t: Training an x-to-text typing interface with online learning from user feedback. In *International Conference on Learning Representations (ICLR)*, 2021.

[36] E. Ratner, A. Bajcsy, T. Fong, C. J. Tomlin, and A. D. Dragan. Efficient dynamics estimation with adaptive model sets. *IEEE Robotics and Automation Letters (RA-L)*, 2021.

[37] A. Bajcsy, S. Bansal, E. Ratner, C.J. Tomlin, and A.D. Dragan. A robust control framework for human motion prediction. *IEEE Robotics and Automation Letters (RA-L)*, 2021.

[38] D. Lindner, R. Shah, P. Abbeel, and A.D. Dragan. Learning what to do by simulating the past. In *International Conference on Learning Representations (ICLR)*, 2021.

[39] A. Bobu, M. Wiggert, C. Tomlin, and A.D. Dragan. Feature expansive reward learning: Rethinking human input. In *International Conference on Human-Robot Interaction (HRI)*, 2021. **(best paper finalist)**.

[40] K. Bhatia, P.L. Bartlett, A.D. Dragan, and J. Steinhardt. Agnostic learning with unknown utilities. In *Innovations in Theoretical Computer Science (ITCS)*, 2021.

[41] H.J. Jeon, S. Milli, and A.D. Dragan. Reward-rational (implicit) choice: a unifying formalism for reward learning. In *Neural Information Processing Systems (NeurIPS)*, 2020.

[42] Y. Du, S. Tiomkin, E. Kiciman, D. Polani, P. Abbeel, and A.D. Dragan. Ave: Assistance via empowerment. In *Neural Information Processing Systems (NeurIPS)*, 2020.

[43] K. Bhatia, A. Pananjady, P.L. Bartlett, A.D. Dragan, and M.J. Wainwright. Preference learning along multiple criteria: A game-theoretic perspective. In *Neural Information Processing Systems (NeurIPS)*, 2020.

[44] S. Reddy, S. Levine, and A.D. Dragan. Assisted perception: Optimizing observations to communicate state. In *Conference on Robot Learning (CoRL)*, 2020.

[45] S. Reddy, A.D. Dragan, S. Levine, S. Legg, and J. Leike. Learning human objectives by evaluating hypothetical behavior. In *International Conference on Machine Learning (ICML)*, 2020.

[46] A. Bobu, A. Bajcsy, J. Fisac, and A.D.Dragan. Quantifying hypothesis space misspecification in learning from human-robot demonstrations and physical corrections. *IEEE Transactions on Robotics (TRO)*, 2020. **(best paper honorable mention)**.

[47] V. Gates, T. Griffiths, and A.D. Dragan. How to be helpful to multiple people at once. In *Cognitive Science*, 2020.

[48] G. Swamy, S. Reddy, S. Levine, and A.D. Dragan. Scaled autonomy: Enabling human operators to control robot fleets. In *International Conference on Robotics and Automation (ICRA)*, 2020.

[49] D. Fridovich-Keil, E. Ratner, A.D. Dragan, and C. Tomlin. Efficient iterative linear-quadratic approximations for nonlinear multi-player general-sum games. In *International Conference on Robotics and Automation (ICRA)*, 2020.

[50] A. Bajcsy, S. Bansal, E. Ratner, A.D. Dragan, and C. Tomlin. A hamilton-jacobi reachability-based framework for predicting and analyzing human motion. In *International Conference on Robotics and Automation (ICRA)*, 2020.

[51] A. Bobu, D. Scobee, S. Satry, and A.D. Dragan. Less is more: Rethinking probabilistic models of human behavior. In *International Conference on Human-Robot Interaction (HRI)*, 2020. **(best paper award)**.

[52] S. Reddy, A.D. Dragan, and S. Levine. Sqil: Imitation learning via reinforcement learning with sparse rewards. In *International Conference on Learning Representations (ICLR)*, 2020.

[53] M. Carroll, R. Shah, M. Ho, T. Griffiths, S. Sheshia, P. Abbeel, and A.D. Dragan. On the utility of learning about humans for human-ai coordination. In *Neural Information Processing Systems (NeurIPS)*, 2019.

[54] I. Huang, S. Huang, R. Pandya, and A.D. Dragan. Nonverbal robot feedback for human teachers. In *Conference on Robot Learning (CoRL)*, 2019. **(oral)**.

[55] S. Milli and A.D. Dragan. Literal or pedagogic human? analyzing human model misspecification in objective learning. In *Uncertainty in Artificial Intelligence (UAI)*, 2019. **(oral)**.

[56] R. Shah, N. Gundotra, P. Abbeel, and A.D. Dragan. Inferring reward functions from demonstrators with unknown biases. In *International Conference on Machine Learning (ICML)*, 2019.

[57] K. Xu, E. Ratner, A.D. Dragan, S. Levine, and C. Finn. Learning a prior over intent via meta-inverse reinforcement learning. In *International Conference on Machine Learning (ICML)*, 2019.

[58] R. Shah and. Krasheninnikov, J. Alexander, P. Abbeel, and A.D. Dragan. Preferences implicit in the state of the world. In *International Conference on Learning Representations (ICLR)*, 2019.

[59] J. Fisac, E. Bronstein, E. Stefansson andD. Sadigh, S. Sastry, and A.D. Dragan. Hierarchical game-theoretic planning for autonomous vehicles. In *International Conference on Robotics and Automation (ICRA)*, 2019.

[60] J. Zhang and A.D. Dragan. Learning from extrapolated corrections. In *International Conference on Robotics and Automation (ICRA)*, 2019.

[61] D. Fridovich, A. Bajcsy, J. Fisac, S. Herbert, S. Wang, A.D. Dragan, and C. Tomlin. Confidence-aware motion prediction for real-time collision avoidance. In *International Journal of Robotics Research (IJRR)*, 2019.

[62] A. Bajcsy, S. Herbert, D. Fridovich, J. Fisac, S. Deglurkar, A.D. Dragan, and C. Tomlin. A scalable framework for real-time multi-robot and multi-human collision avoidance. In *International Conference on Robotics and Automation (ICRA)*, 2019.

[63] R. Choudhury, G. Swarmy, D. Hadfield-Menell, and A.D. Dragan. On the utility of model learning in hri. In *International Conference on Human-Robot Interaction (HRI)*, 2019.

[64] L. Chan, D. Hadfield-Menell, S. Srinivasa, and A.D. Dragan. The assistive multi-armed bandit. In *International Conference on Human-Robot Interaction (HRI)*, 2019.

[65] S. Milli, J. Miller, A.D. Dragan, and M. Hardt. The social cost of strategic classification. In *Conference on Fairness and Accountability and Transparency (FAT*)*, 2019.

[66] S. Milli, L. Schmidt, A.D. Dragan, and M. Hardt. Model reconstruction from model explanations. In *Conference on Fairness and Accountability and Transparency (FAT*)*, 2019.

[67] R. Shah, N. Gundotra, P. Abbeel, and A.D. Dragan. On the feasibility of learning and rather than assuming and human biases for reward inference. In *International Conference on Machine Learning (ICML)*, 2019.

[68] R. Pandya, S. Huang, D. Hadfield-Menell, and A.D. Dragan. Human-ai learning performance in multi-armed bandits. In *Artificial Intelligence and Ethics and Society (AIES)*, 2019.

[69] S. Reddy, A.D. Dragan, and S. Levine. Where do you think you're going? inferring beliefs about dynamics from behavior. In *Neural Information Processing Systems (NeurIPS)*, 2018.

[70] A. Bobu, A. Bajcsy, J. Fisac, and A.D.Dragan. Learning under misspecified objective spaces. In *Conference on Robot Learning (CoRL)*, 2018. **(invited to special issue)**.

[71] L. Sun, W. Zhan, M. Tomizuka, and A.D. Dragan. Courteous autonomous cars. In *International Conference on Intelligent Robots and Systems (IROS)*, 2018.

[72] A. Zhou and A.D. Dragan. Cost functions for robot motion style. In *International Conference on Intelligent Robots and Systems (IROS)*, 2018.

[73] N. Landolfi and A.D. Dragan. Social cohesion in autonomous driving. In *International Conference on Intelligent Robots and Systems (IROS)*, 2018.

[74] H.J. Jeon and A.D. Dragan. Configuration space metrics. In *International Conference on Intelligent Robots and Systems (IROS)*, 2018. **(best student paper award finalist)**.

[75] S. Huang, K. Bhatia, P. Abbeel, and A.D. Dragan. Establishing appropriate trust via critical states. In *International Conference on Intelligent Robots and Systems (IROS)*, 2018.

[76] D. Malik, M. Palaniappan, J. Fisac, D. Hadfield-Menell, S. Russell, and A. D. Dragan. An efficient and generalized bellman update for cooperative inverse reinforcement learning. In *International Conference on Machine Learning (ICML)*, 2018. **(oral)**.

[77] E. Ratner, D. Hadfield-Menell, and A.D. Dragan. Simplifying reward design through divide-and-conquer. In *Robotics: Science and Systems (RSS)*, 2018.

[78] S. Reddy, A.D. Dragan, and S. Levine. Shared autonomy via deep reinforcement learning. In *Robotics: Science and Systems (RSS)*, 2018.

[79] J. Fisac, A. Bajcsy, D. Fridovich, S. Herbert, S. Wang, C. Tomlin, and A.D. Dragan. Probabilistically safe robot planning with confidence-based human predictions. In *Robotics: Science and Systems (RSS)*, 2018. **(invited to special issue)**.

[80] A. Bestick, R. Panya, R. Bajcsy, and A.D. Dragan. Learning human ergonomic preferences for handovers. In *International Conference on Robotics and Automation (ICRA)*, 2018.

[81] D. Sadigh, B. Landolfi, S. Sastry, S. Seshia, and A.D. Dragan. Planning for cars that coordinate with people: Leveraging effects on human actions for planning and active information gathering over human internal state. In *Autonomous Robots (AURO)*, 2018.

[82] A. Bajcsy, D. Losey, M. O'Malley, and A.D. Dragan. Learning from physical human corrections and one feature at a time. In *International Conference on Human-Robot Interaction (HRI)*, 2018.

[83] M. Kwon, S. Huang, and A.D. Dragan. Expressing robot incapability. In *International Conference on Human-Robot Interaction (HRI)*, 2018. **(best paper award finalist)**.

[84] C. Basu, M, Singhal, and A.D. Dragan. Learning from richer human guidance: Augmenting comparison-based learning with feature queries. In *International Conference on Human-Robot Interaction (HRI)*, 2018.

[85] D. Hadfield-Menell, S. Milli, P. Abbeel, S. Russell, and A.D. Dragan. Inverse reward design. In *Neural Information Processing Systems (NIPS)*, 2017. **(oral, acceptance rate 1.2 percent)**.

[86] J. Fisac, M. Gates, J. Hammrick, C. Liu, D. Hadfield-Menell, S. Sastry, T. Griffiths, and A.D. Dragan. Pragmatic-pedagogic value alignment. In *International Symposium on Robotics Research (ISRR)*, 2017. **(best bluesky paper award finalist)**.

[87] M. Laskey, J. Mahler, A.D. Dragan, and K. Goldberg. Dart:optimizing noise injection in imitation learning. In *Conference on Robot Learning (CoRL)*, 2017.

[88] A. Bajcsy, D. Losey, M. O'Malley, and A.D. Dragan. Learning robot objectives from physical human interaction. In *Conference on Robot Learning (CoRL)*, 2017. **(oral, acceptance rate 10 percent)**.

[89] S. Huang, P. Abbeel, and A.D. Dragan. Enabling robots to communicate their objectives. In *Robotics: Science and Systems (RSS)*, 2017. **(invited to special issue)**.

[90] D. Sadigh, A.D. Dragan, S. Sastry, and S. Seshia. Active preference-based learning of reward functions. In *Robotics: Science and Systems (RSS)*, 2017.

[91] S. Milli, D. Hadfield-Menell, A.D. Dragan, P. Abbeell, and S. Russell. Should robots be obedient? In *International Joint Confernece on Artificial Intelligence (IJCAI)*, 2017.

[92] D. Hadfield-Menell, A.D. Dragan, P. Abbeell, and S. Russell. The off-switch game. In *International Joint Confernece on Artificial Intelligence (IJCAI)*, 2017.

[93] J. Andreas, A.D. Dragan, and D. Klein. Translating neuralese. In *Association for Computational Linguistics (ACL)*, 2017.

[94] M. Laskey, S. Krishnan, J. Mahler, K. Jamieson, A.D. Dragan, and K. Goldberg. Comparing human-centric and robot-centric sampling for robot learning from demonstration. In *International Conference on Robotics and Automation (ICRA)*, 2017.

[95] C. Basu, Q. Yang andD. Hungerman, M, Singhal, and A.D. Dragan. Do you want your autonomous car to drive like you? In *International Conference on Human-Robot Interaction (HRI)*, 2017.

[96] A. Zhou, D. Hadfield-Menell andA. Nagabaudi, and A.D. Dragan. Expressive robot motion timing. In *International Conference on Human-Robot Interaction (HRI)*, 2017.

[97] J. Fisac, C. Liu, J. Harick, K. Hedrick, S. Sastry, T. Griffiths, and A.D. Dragan. Generating plans that predict themselves. In *Workshop on the Algorithmic Foundations of Robotics (WAFR)*, 2016.

[98] D. Hadfield-Menell, A.D. Dragan, P. Abbeell, and S. Russell. Cooperative inverse reinforcement learning. In *Neural Information Processing Systems (NIPS)*, 2016.

[99] D. Sadigh, S. Sastry, S. Seshia, and A.D. Dragan. Information gathering actions over human internal state. In *International Conference on Intelligent Robots and Systems (IROS)*, 2016. **(best cognitive robotics paper award finalist)**.

[100] A. Bestick, R. Bajcsy, and A.D. Dragan. Implicitly assisting humans to choose good grasps in robot to human handovers. In *International Symposium on Experimental Robotics (ISER)*, 2016.

[101] M. Laskey, J. Lee, C. Chuck, D.V. Gealy, W. Hsieh, F.T. Pokorny, A.D. Dragan, and K. Goldberg. Using a hierarchy of supervisors in learning from demonstration. In *International Conference on Automation Science and Engineering (CASE)*, 2016.

[102] Z. Marinho, B. Boots, A.D. Dragan, A. Byravan, G.J. Gordon, and S.S. Srinivasa. Functional gradient motion planning in reproducing kernel hilbert spaces. In *Robotics: Science and Systems (R:SS)*, 2016.

[103] D. Sadigh, S. Sastry, S. Seshia, and A.D. Dragan. Planning for autonomous cars that leverages effects on human drivers. In *Robotics: Science and Systems (R:SS)*, 2016. **(invited to special issue)**.

[104] C. Liu, J. Harick, J. Fisac, A.D. Dragan, K. Hedrick, S. Sastry, and T. Griffiths. Goal inference improves objective and perceived performance in human-robot collaboration. In *Autonomous Agents and Multiagent Systems (AAMAS)*, 2016.

[105] S. Nikolaidis, A.D. Dragan, and S.S. Srinivasa. Viewpoint-based legibility optimization. In *International Conference on Human-Robot Interaction (HRI)*, 2016.

[106] M. Laskey, S. Staszak, W. Y. Hsieh, J. Mahler, F.T. Pokorny, A.D. Dragan, and K. Goldberg. Shiv: Reducing supervisor burden in dagger using support vectors for efficient learning from demonstrations in high dimensional state spaces. In *International Conference on Robotics and Automation (ICRA)*, 2016. **(best HRI paper award finalist)**.

[107] N. Mehr and A.D. Dragan. Inferring and assisting with constraints in shared autonomy. In *Conference on Decision and Control (CDC)*, 2016.

[108] A.D. Dragan, K. Muellin, J.A. Bagnell, and S.S. Srinivasa. Movement primitives via optimization. In *International Conference on Robotics and Automation (ICRA)*, 2015. **(best paper and best student paper award finalist)**.

[109] A.D. Dragan, S. Bauman, J. Forlizzi, and S.S. Srinivasa. Effects of robot motion on human-robot collaboration. In *International Conference on Human-Robot Interaction (HRI)*, 2015.

[110] A.D. Dragan, R. Holladay, and S.S. Srinivasa. From legibility to deception. In *Autonomous Robots (AURO)*, 2015.

[111] A.D. Dragan, R. Holladay, and S.S. Srinivasa. Deceptive robot motion: Synthesis and analysis and experiments. In *Autonomous Robots (AURO)*, 2015.

[112] R. Holladay, A.D. Dragan, and S.S. Srinivasa. Legible robot pointing. In *International Symposium on Human and Robot Communication (Ro-Man)*, 2014.

[113] A.D. Dragan, R. Holladay, and S.S. Srinivasa. An analysis of deceptive robot motion. In *Robotics: Science and Systems (R:SS)*, 2014.

[114] A.D. Dragan and S.S. Srinivasa. Integrating human observer inferences into robot motion planning. *Autonomous Robots (AURO)*, 2014.

[115] E. Cha, A.D. Dragan, and S.S. Srinivasa. Pre-school children's first encounter with a robot. In *International Conference on Human-Robot Interaction (HRI)*, 2014.

[116] E. Cha, A.D. Dragan, and S.S. Srinivasa. Effects of speech on perceived capability. In *International Conference on Human-Robot Interaction (HRI)*, 2014.

[117] A.D. Dragan and S.S. Srinivasa. Familiarization to robot motion. In *International Conference on Human-Robot Interaction (HRI)*, 2014.

[118] H. Admoni, A.D. Dragan, B. Scassellati, and S.S. Srinivasa. Deliberate delays during robot-to-human handovers improve compliance with gaze communication. In *International Conference on Human-Robot Interaction (HRI)*, 2014.

[119] A.D. Dragan and S.S. Srinivasa. A policy blending formalism for shared control. *International Journal of Robotics Research (IJRR)*, 2013.

[120] M. Zucker, N. Ratliff, A.D. Dragan, M. Pivtoraiko, M. Klingensmith, C. Dellin, J. Bagnell, and S.S. Srinivasa. CHOMP: Covariant Hamiltonian Optimization for Motion Planning. *International Journal of Robotics Research (IJRR)*, 2013.

[121] A.D. Dragan, K.T. Lee, and S.S. Srinivasa. Teleoperation with intelligent and customizable interfaces. *Journal of Human-Robot Interaction (JHRI)*, 2013.

[122] A.D. Dragan and S.S. Srinivasa. Generating legible motion. In *Robotics: Science and Systems (R:SS)*, 2013. **(best paper award finalist)**.

[123] E. Cha, A.D. Dragan, and S.S. Srinivasa. Effects of robot capability on user acceptance. In *International Conference on Human-Robot Interaction (HRI)*, 2013.

[124] K.T. Lee, A.D. Dragan, and S.S. Srinivasa. Legible user input for intent prediction. In *International Conference on Human-Robot Interaction (HRI)*, 2013.

[125] A.D. Dragan, K.T. Lee, and S.S. Srinivasa. Legibility and predictability of robot motion. In *International Conference on Human-Robot Interaction (HRI)*, 2013.

[126] K. Strabala, M.K. Lee, A.D. Dragan, J. Forlizzi, S.S. Srinivasa, M. Cakmak, and V. Micelli. Towards seamless human-robot handovers. *Journal of Human-Robot Interaction (JHRI)*, 2013.

[127] A.D. Dragan and S.S. Srinivasa. Formalizing assistive teleoperation. In *Robotics: Science and Systems (R:SS)*, 2012. **(best paper award finalist)**.

[128] A.D. Dragan and S.S. Srinivasa. Online customization of teleoperation interfaces. In *International Symposium on Human and Robot Communication (Ro-Man)*, 2012. **(best paper award finalist)**.

[129] K. Strabala, M.K. Lee, A.D. Dragan, J. Forlizzi, and S.S. Srinivasa. Learning the communication of intent prior to physical collaboration. In *International Symposium on Robot and Human Interactive Communication (Ro-Man)*, 2012.

[130] S.S. Srinivasa, D. Berenson, M. Cakmak, A. Collet, M.R. Dogar, A.D. Dragan, R.A. Knepper, T. Niemueller, K. Strabala, M. Vande Weghe, and J. Ziegler. HERB 20: Lessons learned from developing a mobile manipulator for the home. *Proc. of the IEEE and Special Issue on Quality of Life Technology*, 2012.

[131] A.D. Dragan, G. Gordon, and S.S. Srinivasa. Learning from experience in manipulation planning: Setting the right goals. In *International Symposium on Robotics Research (ISRR)*, 2011.

[132] A.D. Dragan, N. Ratliff, and S.S. Srinivasa. Manipulation planning with goal sets using constrained trajectory optimization. In *International Conference on Robotics and Automation (ICRA)*, 2011.

[133] I. Schuele, A.D. Dragan, A. Radev, M. Schroeder, and K.H. Kuffer. Multi-criteria optimization for regional timetable synchronization in public transport. *Operations Research Proceedings*, 2008.