Dylan Hadfield-Menell

32 Vassar St, 32-330, Cambridge, MA
dhm@csail.mit.edu
https://people.csail.mit.edu/dhm

Education

University of California, Berkeley
- Ph.D. in Computer Science - Artificial Intelligence, 2021
- Dissertation*: The Principal-Agent Alignment Problem in Artificial Intelligence*
  - Advisors: Stuart Russell, Anca Dragan, Pieter Abbeel

Massachusetts Institute of Technology
- M.Eng. in Computer Science and Electrical Engineering, 2013
  - Thesis: *Execution Cost Optimization for Hierarchical Planning in the Now*
  - Advisors: Leslie Kaelbling, Tomás Lozano-Pérez
- S.B. in Computer Science and Computer Engineering, MIT 2012

Research and Academic Positions

Massachusetts Institute of Technology, Cambridge
- Bonnie and Marty (1964) Tenenbaum Career Development Assistant Professor of Artificial Intelligence and Decision-Making, July 2021 - current

Preamble AI
- Chief Research Officer & Co-Founder, March 2021 - Current

OpenAI, San Francisco
- Machine Learning Researcher, Spring 2018

University of California, Berkeley
- Ph.D. Student, September 2013 - August 2021

Teaching Experience

6.s979 - Values in AI: Accidents, Alignment, and Misuses, Fall 2023
6.4200 – Robotics: Science and Systems, Instructor, Spring 2023
6.1010 - Fundamentals of Programming, Instructor, Fall 2022
6.141 – Robotics: Science and Systems, Instructor, Spring 2022
6.009 - Fundamentals of Programming, Instructor, Fall 2021
CS 294 - Human-Compatible Artificial Intelligence, Instructor, Spring 2016
CS 188 - Introduction to A. I., Head Graduate Student Instructor, Fall 2014
6.141 - Robotics: Science and Systems, Head Lab Assistant, Fall 2012
6.004 - Computation Structures, Lab Assistant, Fall 2012

6.004 - Computation Structures, Lab Assistant, Fall 2011
6.004 - Computation Structures, Lab Assistant, Spring 2011

Fellowships, Awards & Honors

Berkeley Fellowship, 2013-2014
NSF Graduate Research Fellowship, 2013-2018
C.V. Ramamoorthy Distinguished Research Award, 2021
Bonnie and Marty (1864)Tenenbaum Career Development Chair, 2021-2024
Schmidt Futures AI2050 Early Career Fellow

Publications

Journal Papers

Open Problems and Fundamental Limitations of Reinforcement Learning from Human Feedback. *Stephen Casper, Xander Davies, Claudia Shi, Thomas Krendl Gilbert, Jérémy Scheurer, Javier Rando, Rachel Freedman, Tomasz Korbak, David Lindner, Pedro Freire, Tony Wang, Samuel Marks, Charbel-Raphaël Segerie, Micah Carroll, Andi Peng, Phillip Christoffersen, Mehul Damani, Stewart Slocum, Usman Anwar, Anand Siththaranjan, Max Nadeau, Eric J. Michaud, Jacob Pfau, Dmitrii Krasheninnikov, Xin Chen, Lauro Langosco, Peter Hase, Erdem Bıyık, Anca Dragan, David Krueger, Dorsa Sadigh, Dylan Hadfield-Menell.* Transactions on Machine Learning Research (TMLR). 2023.

Building Human Values into Recommender Systems: An Interdisciplinary Synthesis. *Jonathan Stray, Alon Halevy, Parisa Assar, Dylan Hadfield-Menell, Craig Boutilier, Amar Ashar, Lex Beattie, Michael Ekstrand, Claire Leibowicz, Connie Moon Sehat, Sara Johansen, Lianne Kerlin, David Vickrey, Spandana Singh, Sanne Vrijenhoek, Amy Zhang, McKane Andrus, Natali Helberger, Polina Proutskova, Tanushree Mitra, Nina Vasan. ACM Transactions on Recommender Systems, 2023.*

Judging facts, judging norms: Training machine learning models to judge humans requires a modified approach to labeling data. *Aparna Balagopalan, David Madras, David H Yang, Dylan Hadfield-Menell, Gillian K Hadfield, Marzyeh Ghassemi.* Science Advances. 2023.

Spurious Normativity Enhances Learning of Compliance and Enforcement Behavior in Artificial Agents. *Raphael Koster, Dylan Hadfield-Menell, Richard Everett, Laura Weidinger, Gillian K. Hadfield, and Joel Z. Leibo.* Proceedings of the National Academy of Sciences of the United States (PNAS). 2022.

When Curation Becomes Creation. *Leqi Liu, Dylan Hadfield-Menell, and Zachary C. Lipton.* Communications of the ACM 64.12: 44-47. 2021.

Conference Papers

Distributional Preference Learning: Understanding and Accounting for Hidden Context in RLHF. *Anand Siththaranjan, Cassidy Laidlaw, Dylan Hadfield-Menell. Proceedings of the 12th International Conference on Learning Representations, 2024.*

Cognitive Dissonance: Why Do Language Model Outputs Disagree with Internal Representations of Truthfulness? *Kevin Liu, Stephen Casper, Dylan Hadfield-Menell, Jacob Andreas. Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, 2023.*

Red Teaming Deep Neural Networks with Feature Synthesis Tools. *Stephen Casper, Yuxiao Li, Jiawei Li, Tong Bu, Kevin Zhang, Kaivalya Hariharan, Dylan Hadfield-Menell. Proceedings of the 37th Annual Conference on Neural Information Processing Systems, 2023.*

Toward Transparent AI: A Survey on Interpreting the Inner Structures of Deep Neural Networks. *Tilman Räuker, Anson Ho, Stephen Casper, Dylan Hadfield-Menell. Proceedings of the 2023 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML), 2023.*

How to talk so AI will learn: Instructions, descriptions, and autonomy. *Theodore Sumers, Robert Hawkins, Mark K Ho, Tom Griffiths, Dylan Hadfield-Menell. Proceedings of the 36th Annual Conference on Neural Information Processing Systems, 2022.*

Robust Feature-Level Adversaries are Interpretability Tools. *Stephen Casper, Max Nadeau, Dylan Hadfield-Menell, Gabriel Kreiman. Proceedings of the 36th Annual Conference on Neural Information Processing Systems, 2022.*

Estimating and Penalizing Induced Preference Shifts in Recommender Systems. *Micah D. Carroll, Anca Dragan, Stuart Russell, Dylan Hadfield-Menell. Proceedings of the 39th International Conference on Machine Learning, PMLR 162:2686-2708, 2022.*

A Penalty Default Approach to Preemptive Harm Disclosure and Mitigation for AI Systems. *Rui-Jie Yew, Dylan Hadfield-Menell.* Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES), 2022.

Guided Imitation of Task and Motion Planning. *Michael McDonald, Dylan Hadfield-Menell.* Proceedings of the 5th Conference on Robot Learning (CoRL). 2021.

The Consequences of Misaligned AI. *Simon Zhuang, Dylan Hadfield-Menell.* Proceedings of the 34th Conference on Neural Information Processing Systems (NeurIPS). 2020.

Conservative Agency via Attainable Utility Preservation. *Alexander M. Turner, Dylan Hadfield-Menell, Prasad Tadepalli,* Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES), 2020.

The Assistive Multi-Armed Bandit. *Lawrence Chan, Dylan Hadfield-Menell, Siddartha S. Srinivasa, Anca D. Dragan* ACM/IEEE International Conference on Human-Robot Interaction (HRI), 354:363. 2019

Human-AI Learning Performance in Multi-Armed Bandits. *Ravi Pandya, Sandy H. Huang, Dylan Hadfield-Menell, Anca D. Dragan*, Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES), 369-375. 2019

Legible Normativity for AI Alignment: The Value of Silly Rules. *Dylan Hadfield-Menell, McKane Andrus, Gillian K. Hadfield* Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES), 115-121. 2019.

Incomplete contracting and AI alignment. *Dylan Hadfield-Menell, Gillian K. Hadfield,* Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES), 417-422. 2019.

An Efficient Generalized Bellman Update for Cooperative Inverse Reinforcement Learning. *Dhruv Malik, Malayandi Palaniappan, Jamie F. Fisac, Dylan Hadfield-Menell, Stuart J. Russell, Anca D. Dragan*, Proceedings of International Conference on Machine Learning Research (ICML), 3391-3399. 2018

Simplifying Reward Design through Divide-and-Conquer. Ellis Ratner, Dylan Hadfield-Menell, Anca D. Dragan, Proceedings of Robotics: Science and Systems (RSS), 2018.

Pragmatic-Pedagogic Value Alignment. Jaime F. Fisac, Monica A. Gates, Jessica B. Hamrick, Chang Liu, Dylan Hadfield-Menell, Malayandi Palaniappan, Dhruv Malik, S. Shankar Sastry, Thomas L. Griffiths, Anca D. Dragan, Proceedings of the International Symposium on Robotics Research (ISRR), 2017

Inverse Reward Design. *Dylan Hadfield-Menell, Smitha Milli, Pieter Abbeel, Stuart Russell, and Anca D. Dragan.* Proceedings of the 31st Conference on Neural Information Processing Systems (NeurIPS). 2017.

The Off-Switch Game. *Dylan Hadfield-Menell, Anca D. Dragan, Pieter Abbeel, and Stuart Russell.* In Proceedings of the 26th International Joint Conference on Artificial Intelligence (IJCAI), 220-227. 2017.

Should Robots be Obedient? *Smitha Milli, Dylan Hadfield-Menell, Anca D. Dragan, and Stuart Russell.* In Proceedings of the 26th International Joint Conference on Artificial Intelligence (IJCAI), 2017.

Expressive Robot Motion Timing. *Allan Zhou, Dylan Hadfield-Menell, and Anca D. Dragan.* In ACM/IEEE International Conference on Human Robot Interaction (HRI), 2017.

Cooperative Inverse Reinforcement Learning. *Dylan Hadfield-Menell, Anca D. Dragan, Pieter Abbeel, and Stuart Russell.* In Proceedings of the 30th Conference on Neural Information Processing Systems (NeurIPS), 2016.

Sequential Quadratic Programming for Task Plan Optimization. *Dylan Hadfield-Menell, Chris Lin, Rohan Chitnis, Pieter Abbeel, and Stuart Russell.* In IEEE/RSJ Conference on Intelligent Robots and Systems (IROS), 2016.

Guided Search for Task and Motion Plans Using Learned Heuristics. *Rohan Chitnis, Dylan Hadfield-Menell, Abhishek Gupta, Siddharth Srivastava, Edward Groshev, Christopher Lin, and Pieter Abbeel.* In IEEE Conference on Robotics and Automation (ICRA), 2016.

Modular Task and Motion Planning in Belief Space. *Dylan Hadfield-Menell, Edward Groshev, Rohan Chitnis, and Pieter Abbeel.* In IEEE/RSJ Conference on Intelligent Robots and Systems (IROS), 2015.

Multitasking: Efficient Optimal Planning for Bandit Superprocesses. *Dylan Hadfield-Menell, and Stuart Russell.* In Proceedings of the 31st Conference on Uncertainty in Artificial Intelligence (UAI), 2015.

Beyond Lowest-Warping Cost Action Selection in Trajectory Transfer. *Dylan Hadfield-Menell, Alex X. Lee, Chelsea Finn, Eric Tzeng, Sandy Huang, and Pieter Abbeel.* In IEEE Conference on Robotics and Automation (ICRA), 2015.

Unifying Scene Registration and Trajectory Optimization for Learning from Demonstrations with Application to Manipulation of Deformable Objects. *Alex X. Lee, Sandy H. Huang, Dylan Hadfield-Menell, Eric Tzeng, Pieter Abbeel.* In IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2014.

Optimization in the Now: Dynamic Peephole Optimization for Hierarchical Planning. *Dylan Hadfield-Menell, Leslie Pack Kaelbling, and Tomás Lozano-Pérez.* In IEEE Conference on Robotics and Automation (ICRA), 2013.

Pre-Prints & Workshop Papers

Counterfactual Metrics for Auditing Black-Box Recommender Systems for Ethical Concerns. *Nil-Jana Akpinar, Liu Leqi, Dylan Hadfield-Menell, Zachary Lipton.* ICML Workshop on Responsible Decision Making in Dynamic Environments. 2022

Open Coding for Machine Learning. *Magdalena Price, Dylan Hadfield-Menell.* ICML DataPerf Workshop. 2022

Estimating and Penalizing Induced Preference Shifts in Recommender Systems. *Micah Carroll, Dylan Hadfield-Menell, Stuart Russell, Anca D. Dragan.* RecSys LBR Track, FAcctRec Workshop. 2021

Multi-Principal Assistance Games: Definition and Collegial Mechanisms. *Arnaud Fickinger, Simon Zhuang, Andrew Critch, Dylan Hadfield-Menell, Stuart Russell.* NeurIPS Workshop on Cooperative AI. 2020.

What are you optimizing for? Aligning Recommender Systems with Human Values. *Jonathan Stray, Ivan Vandrov, Jeremy Nixon, Steven Adler, Dylan Hadfield-Menell.* ICML Workshop on Participatory Machine Learning. 2020.

An Extensible Interactive Interface for Agent Design. *Matthew Rahtz, James Fang, Anca D. Dragan, Dylan Hadfield-Menell*

Adversarial Training with Voronoi Constraints. *Marc Khoury, Dylan Hadfield-Menell*

Talks and Panels

Designer in the Loop Reinforcement Learning, Workshop on Human in (and around) the Loop Reinforcement Learning, RLDM, June 2022

Aligning Recommendation Systems, Center for Human-Compatible AI Workshop, June 2022

Incompleteness, Overoptimization, and Goodhart's Law, RaD-AI Workshop, AAMAS, May 2022

Incompleteness, Artificial Intelligence, and the Problem with Proxies, Harvard-MIT Alignment Speaker Series, April 2022

Incompleteness, Artificial Intelligence, and the Problem with Proxies, IAFI Journal Club, April 2022

Silly Rules and Normative Structures: Using multi-agent learning to study 3rd party punishment in groups, MIT Media Lab Computational Social Science Seminar, February 2022

The Theory and History of Goodhart's Law, Existential and Societal-Scale Safety for Artificial Intelligence Colloquium, February 2022

Workshop on Safe Reinforcement Learning, Panel Discussion & Debate, NeurIPS, 2021

Incompleteness, Artificial Intelligence, and the Problem with Proxies, Deepmind Beneficial RL Seminar, 2021

Incompleteness, Artificial Intelligence, and the Problem with Proxies, University of Toronto AI Safety Reading Group, 2021

AI Needs Normative Infrastructure, Meta Governance Seminar Series, 2021

Human Values in AI, Wonderfest Science Fellow Speaker Series, 2019

Formalizing the Value Alignment Problem in AI, ICLR Workshop on Safe Machine Learning: Specification, Robustness, and Assurance, New Orleans 2019

Future of Life Institute - Beneficial AGI, Moderator, Technical safety student panel: *What does the next generation of researchers think that the action items should be?* Puerto Rico, 2019

The Assistive Multi-Armed Bandit, Human Robot Interaction, Daegu, Korea, 2019

On the Utility of Model Learning in HRI, Human Robot Interaction, Daegu, Korea, 2019

Value Alignment in Artificial Intelligence, Hastings Institute Control and Responsible Innovation in the Development of Autonomous Machines Experts Workshop, New York, 2018

Science Envoy: Science Slam, Computer History Museum, Palo Alto, 2018,

Inverse Reward Design, ISAT/DARPA Workshop on Diverse Ways of Inferring Missions, Washington DC, 2017

Inverse Reward Design, NeurIPS Oral, Long Beach, 2017

The Off-Switch Game, International Joint Conference on Artificial Intelligence, Melbourne, 2017

Value Alignment in Artificial Intelligence, Oxford Future Humanity Institute Workshop on Malicious Actors and A.I., Oxford, England 2017

The Off-Switch Machine Intelligence Research Institute & Future of Humanity Institute, Colloquium Series on Robust and Beneficial AI, 2016

Sequential Quadratic Programming for Task Plan Optimization, Planning & Robotics Workshop at International Conference on Planning and Scheduling Systems, 2016

Cooperative Inverse Reinforcement Learning, NeurIPS Spotlight, Barcelona, 2016

Cooperative Inverse Reinforcement Learning, Algorithmic HRI Workshop, Paris, 2016

Sequential quadratic programming for task plan optimization, IEEE/RSJ International Conference on Intelligent Robots and Systems, 2016

Multitasking: Optimal Planning for Bandit Superprocesses, AAAI Conference on Uncertainty in A.I., Amsterdam, Netherlands, 2015

Modular Task and Motion Planning in Belief Space, IEEE Conference on Robotics and Automation, Incheon, Korea, 2015

Learning to Select Expert Demonstrations, Robotics: Science and Systems Workshop on Information-based Grasp and Manipulation Planning, 2014

Optimization in the Now: Dynamic Peephole Optimization for Task and Motion Planning, IEEE Conference on Robotics and Automation, Karlsruhe, Germany, 2013

Professional Activities and Service

Workshops (Co-)Organized

Building and Evaluating Ethical Robot Systems. IEEE/RSJ International Conference on Intelligent Robots and Systems, Prague, Czech Republic 2021

RL-CONFORM. IEEE/RSJ International Conference on Intelligent Robots and Systems, Prague, Czech Republic 2021

Aligned AI. Neural Information Processing Systems, Long Beach, California 2017.

Reliable Machine Learning in the Wild. International Conference on Machine Learning, Sydney, Australia 2017.

Reliable Machine Learning in the Wild. Neural Information Processing Systems, Barcelona, Spain 2016.

Program Committees and Journal Refereeing

Journal of Machine Learning Research

Transactions on Human-Robot Interaction

International Journal of Robotics Research

Journal of AI Research

AI Magazine

AIES, Artificial Intelligence Ethics and Society, 2017, 2018, 2019, 2020, 2022

WAFR, International Workshop on Algorithmic Foundations of Robotics, 2022

CoRL, Conference on Robot Learning, 2020, 2021, 2022

RSS, Robotics: Science and Systems, 2019, 2020, 2021, 2022

HRI, Human-Robot Interaction, 2018, 2019, 2020, 2021, 2022

AIStats, Artificial Intelligence and Statistics Conference, 2021, 2022 (Top Reviewer)

NeurIPS, Neural Information Processing Systems, 2016, 2017, 2018, 2019, 2020

ICML, International Conference on Machine Learning, 2020, 2021, 2022

AAAI, Conference on Artificial Intelligence, 2018

IJCAI, International Joint Conference on Artificial Intelligence, 2016, 2017

IROS, IEEE Conference on Intelligent Robots and Systems, 2015, 2016

ICRA, IEEE Conference on Robotics and Automation, 2016, 2017, 2018

Public Service

Pan-African Robotics Competition, Judge, 2021

AI4ALL, Summer Camp for Underrepresented Talent, Berkeley, CA 2017, 2018, 2019

Victoria and Albert Museum, Technical Lead, *The Future Starts Here Exhibit*, London, England, 2018

Podcasts

Cooperative Inverse Reinforcement Learning, AI Alignment Podcast, Future of Life Institute, 2019

Inverse Reinforcement Learning and Inferring Human Preferences, AI Alignment Podcast, Future of Life Institute, 2018