Center for Human-Compatible Artificial Intelligence

PROGRESS REPORT | 2020

Prof. Stuart J. Russell, CHAI Faculty Director
and staff

September 30, 2020

TABLE OF CONTENTS

Research towards solving the problem of control

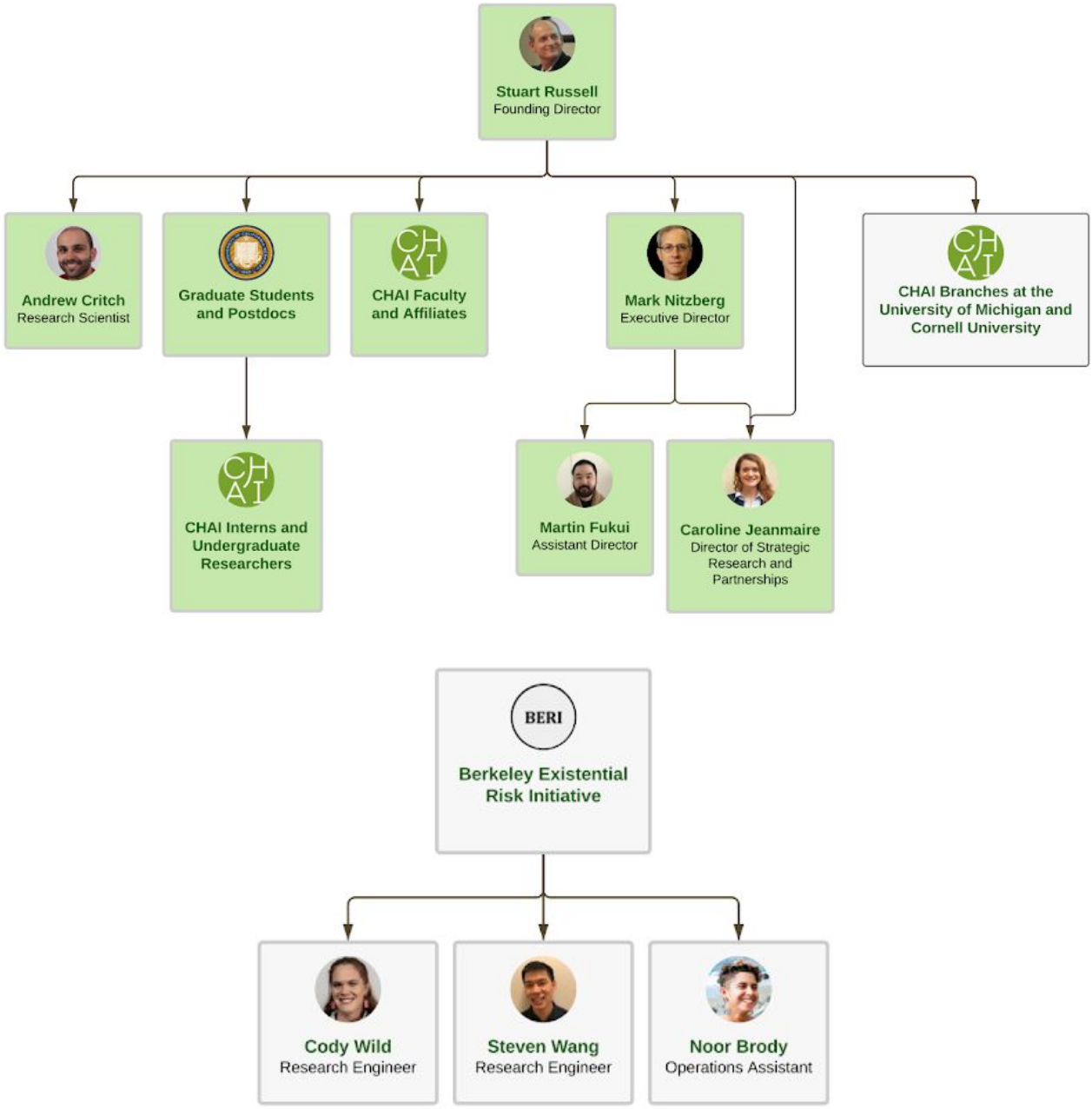
Founded in 2016, CHAI is a multi-site research center headquartered at UC Berkeley with branches at Michigan and Cornell. CHAI's aim is to reorient AI research towards provably beneficial systems, over which humans can retain control even as they approach or exceed human-level decision-making capabilities. This document reports on CHAI's activities and accomplishments in its first four years and its plans for the future.

CHAI currently has 9 faculty investigators, 18 affiliate faculty, around 30 additional graduate and postdoctoral researchers (including roughly 25 PhD students), many undergraduate researchers and interns, and a staff of 5. CHAI's primary support comes in the form of gifts from donor organizations and individuals. Its primary activities include research, academic outreach, and public policy engagement.

CHAI's research output includes foundational work to reframe AI on the basis of a new model that factors in uncertainty about human preferences, in contrast to the standard model for AI in which the objective is assumed to be known completely and correctly. Our work includes topics such as misspecified objectives, inverse reward design, assistance games with humans, obedience, preference learning methods, social aggregation theory, interpretability, and vulnerabilities of existing methods. Given the massive resources worldwide devoted to research within the standard model of AI, CHAI's undertaking also requires engaging with this research community to adopt and further develop AI based on this new model. In addition to academic outreach, CHAI strives to reach general audiences through publications and media. We also advise governments and international organizations on policies relevant to ensuring AI technologies will benefit society, and offer insight on a variety of individual-scale and societal-scale risks from AI, such as pertaining to autonomous weapons, the future of employment, and public health and safety.

The October 2019 release of the book *Human Compatible* explained the mission of the Center to a broad audience just before most of us were forced to work from home. The pandemic has made evident our dependence on AI technologies to understand and interact with each other and with the world outside our windows. The work of CHAI is crucial now, not just in some future in which AI is more powerful than it is today.

Organizational Chart



CHAI Research

Overview

We created the Center for Human-Compatible AI to understand and solve the problem of control in Artificial Intelligence. Although "this matters, not because AI is rapidly becoming a pervasive aspect of the present but because it is the dominant technology of the future,"[1] in fact it is clear that we are already in over our heads.

For instance, the content-selection systems of social media platforms and search engines choose what news articles, podcasts, videos and personal updates are viewed by half the population of the planet on a daily basis. These systems decide what people read and view, and to a degree, what they think and feel, based on AI algorithms that have fixed objectives - for example, the objective of maximizing "engagement" of users. This has driven a movement toward extremes that has eroded civility at best, and arguably threatens political stability.

The social media companies' struggle to implement piecemeal solutions with mixed results further illustrates the problem of how to control AI systems that are designed to achieve a fixed, known objective. This fixed-objective model is what we refer to as the "standard model" of AI.

AI systems built within this standard model present a significant control problem for both individuals and society; CHAI's strategy is to address this problem by reformulating the foundations of AI research and design.

Thus, CHAI has proposed and is developing a new model for AI, where (1) the machine's objective is to help humans in realizing the future we prefer; (2) the machine is explicitly uncertain about those human preferences; (3) human behavior provides evidence of human preferences. Machines designed in accordance with these principles behave cautiously and defer to humans; they allow themselves to be switched off; and, under some conditions, they are provably beneficial.

Characteristics of the new model

The key characteristics of the new model are the absence of a fixed, known objective — whether at design time or embedded in the agent itself — and the flow of preference information from human to machine at runtime.

The new model is strictly more general than the standard model, and at least as amenable to instantiation in a wide variety of forms. One particular formal instantiation is the *assistance game*

[1] Russell, Stuart. Human Compatible (p. xi)

— originally a cooperative inverse reinforcement learning or CIRL game [3ab].[2] In recent work [8], we have shown formally that many settings explored by other AI safety research groups can be understood within the assistance-game framework. We have explored several implications and extensions of the basic single-human single-robot assistance game, including showing that machines solving an assistance game allow themselves to be switched off [4a]; a more general analysis of complete and partial obedience [4b]; humans uncertain about their own preferences [3c]; humans giving noisy, partial rewards [5]; and the first forays into assistance games with real humans [6].

The simple, single-human/single-robot assistance game has yielded many important insights and also models the relationship between the human race and its machines, each construed monolithically. Additional complications arise, of course, when we consider the multiplicity of humans and machines. Decision making on behalf of multiple humans is the subject of millenia of research in moral philosophy and the social sciences and was the main subject of a graduate source co-taught by Prof. Russell with Economics and Philosophy professors in Spring 2020. Our initial results in this area include a strict generalization of Harsanyi's social aggregation theorem to handle heterogeneity in human beliefs (important for cross-cultural cooperation) [9] and some as-yet unpublished work on mechanism design to incentivize honest revelation of preferences by humans. Handling multiple "robots" is also extremely important, particularly when the robots are independently designed and not a priori cooperative. Here we have fundamental results on bounded formal reasoning leading to cooperation [11] and on global equilibria in symmetric games (under review).

We are in only the first phase of developing the new model as a practical and safe framework for AI. Many open problems remain, as outlined in the "Future Plans" section.

Promoting the new model

Given our belief that solutions are irrelevant if they are ignored, the principles of the new model have been disseminated in the form of a general-audience book [1], a revised textbook edition [2], numerous technical papers, many keynote talks at leading AI conferences, direct advice to national governments and international organizations, media articles, podcasts, invited talks at industry and general-interest conferences, TV and radio interviews, and documentary films.

In keeping with our view that the new model must become the normal approach within the mainstream community, rather than remaining confined to a relatively small and cloistered AI safety community, we have targeted most of our research papers at the most selective mainstream AI, machine learning, and robotics conferences including Neural Information processing Systems (NeurIPS), International Conference on Machine Learning (ICML), International Joint Conference on AI (IJCAI), Uncertainty in AI (UAI), International Conference on Learning Representations (ICLR), and Human-Robot Interaction (HRI). We believe these papers have helped to establish AI safety as a respectable field within mainstream AI.

[2] All numeric references point to items in the "Specific Outputs" section.

Other research outputs

Other work in CHAI overlaps with concerns in the broader AI community. We have shown surprising fragility in deep RL systems [12] and explored methods for increasing modularity and hence interpretability in deep networks (unpublished) [arXiv preprint].

We have also attempted a quasi-exhaustive analysis of existential risk from AI, including AI systems that are not necessarily superintelligent [10] [ARCHES]. One major, understudied category of risks to emerge from this analysis arises from unanalyzed interactions among multiple independent AI systems. This topic will form a significant part of the future research agenda of CHAI, as noted in the Future Plans section.

Specific outputs

Note: This does not include work from the two CHAI satellite groups at Michigan and Cornell.

1. **Human Compatible** — This book is aimed at the general intellectual reader, the policy community, and the AI community. It provides a thorough but nontechnical explanation of the standard model of AI, why it leads to societal-scale and existential risks, a new model based on the principles of provably beneficial AI, and the many important research questions that arise. It also covers fairness/bias, employment, surveillance/control, and autonomous weapons.

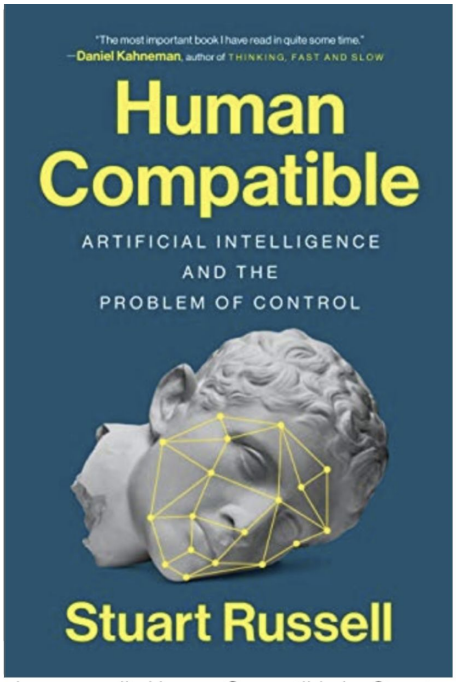


Image credit: Human Compatible by Stuart Russell (Penguin Random House, 2019)

The book had two primary goals. The first was to raise public awareness and understanding in a non-sensationalist way. The book was reviewed and excerpted in the New York Times, Wall Street Journal, Financial Times (4 times), Economist, Forbes (twice), Times (London), Sunday Times (twice), Daily Telegraph (twice), Guardian (3 times), Vox, Spectator, among others, and won "best book" awards from the Financial Times, Guardian, Daily Telegraph, and Forbes.

The second goal was to locate AI safety research squarely at the center of AI's intellectual tradition, and to begin the process of converting the AI community to a new way of thinking. Review comments from leading scientists include those from three Turing Award winners and one Nobel laureate:

> **Judea Pearl**, Professor of Computer Science, UCLA: "Human Compatible made me a convert to Russell's concerns with our ability to control our upcoming creation — super-intelligent machines. Unlike outside alarmists and futurists, Russell is a leading

authority on AI. His new book will educate the public about AI more than any book I can think of, and is a delightful and uplifting read."

Yoshua Bengio, Professor of Computer Science, U. Montreal: "This beautifully written book addresses a fundamental challenge for humanity: increasingly intelligent machines that do what we ask but not what we really intend. Essential reading if you care about our future."

Andy Yao, Dean of the Institute for Interdisciplinary Information Sciences at Tsinghua: "This is a fascinating masterpiece: both general readers and artificial intelligence experts will be inspired by it. Professor Russell has made the most profound and clearest analysis in the literature on superintelligence, the ultimate problem of artificial intelligence. More importantly, he proposed a novel solution – a new human-computer relationship – to solve the problem of superintelligence. This idea has opened up a new research direction in AI."

Daniel Kahneman, Professor of Psychology, Princeton: "This is the most important book I have read in quite some time. It lucidly explains how the coming age of artificial super-intelligence threatens human control. Crucially, it also introduces a novel solution and a reason for hope."

Similar comments from industry thought leaders include:

James Manyika, Chairman and Director, McKinsey Global Institute: "Stuart Russell, one of the most important AI scientists of the last 25 years, may have written the most important book about AI so far, on one of the most important questions of the 21st century: How to build AI to be compatible with us."

Tim O'Reilly, Founder, O'Reilly Media: "I just finished Stuart Russell's marvelous book on AI safety, Human Compatible, and I can't recommend it highly enough!"

In addition, *Human Compatible* was the theme for day-long, campus-wide symposia at UCLA (organized by the Department of Sociology) and UC San Diego (planned by the Institute for Practical Ethics for March 2020, postponed due to COVID). It was also the theme of a special session at the 2020 meeting of the American Political Science Association. Russell will deliver the inaugural *Forum Humanum* lecture at NYU, on the topic of the book, this fall.

a. *Human Compatible: AI and the Problem of Control* (2019). Stuart Russell. Viking. 352pp.

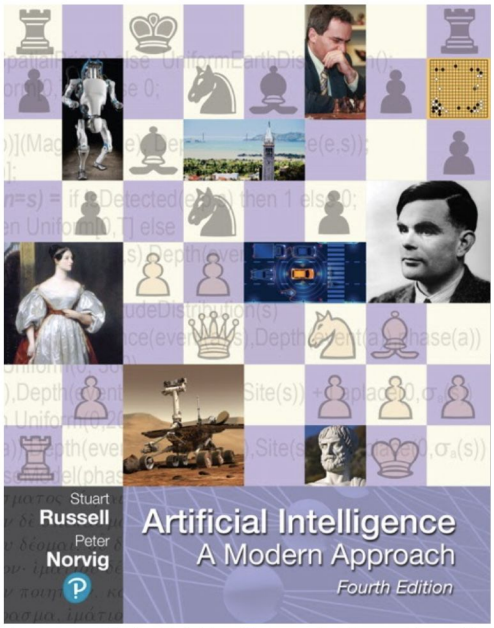


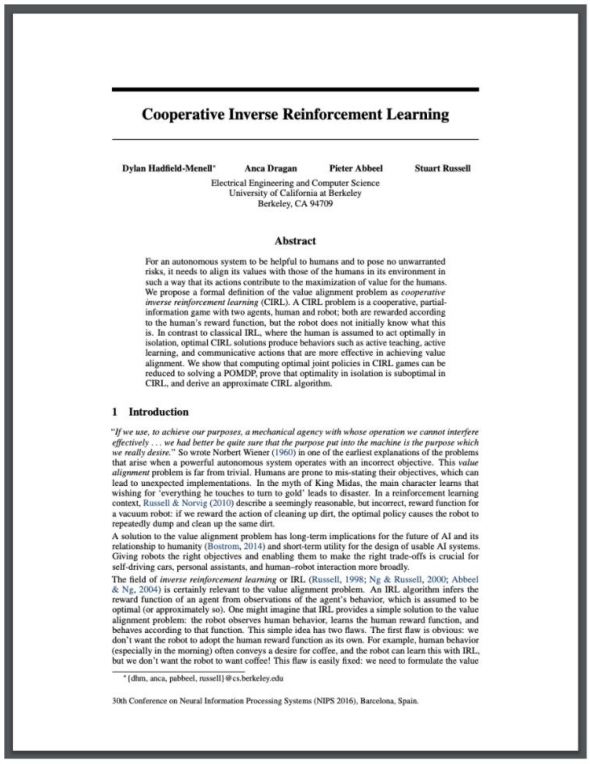
Image credit: Artificial Intelligence by Stuart Russell and Peter Norvig, Fourth Edition (Pearson, 2020)

2. **Artificial Intelligence: A Modern Approach (4th edition)** — AIMA is the standard text in AI, used in almost 1500 universities in 135 countries. According to a survey in *Nature* (539, 125-6, 2016; source), it is the most widely adopted of the roughly 80,000 textbooks in computer science. The 4th edition includes extensive descriptions

of the risks inherent in the standard model, the basic principles of the new model, and most of the technical material from the papers listed below. This ensures that current and future generations of AI students understand the importance of thinking about and mitigating potential risks, while promoting centrality of AI safety research.

   a. *Artificial Intelligence: A Modern Approach* (2020), Stuart Russell, Peter Norvig. Pearson. 1133pp.

3. **Cooperative Inverse Reinforcement Learning (CIRL)** — This sequence of papers formalizes the cooperative game-theoretic relationship between an AI assistant and a human user, instantiates the new model in a specific mathematical form, explores the provably beneficial nature of solutions, and derives efficient solution algorithms for handling preference uncertainty on the part of the human. This has made it easier for many researchers to begin talking about "the alignment problem" between a single AI system and a single human in greater technical detail than in prior work. It also enabled follow-up work on several important aspects of human/AI interaction (shut-down, overrides, rewards, and human models) as described in more detail in items 4, 5, and 6. CIRL is an important conceptual building block in our (explicit or implicit) understanding of how powerful AI technology should be integrated with society, without which any theory of societal-scale impact will be ungrounded.



Cooperative Inverse Reinforcement Learning

Dylan Hadfield-Menell* Anca Dragan Pieter Abbeel Stuart Russell

Electrical Engineering and Computer Science
University of California at Berkeley
Berkeley, CA 94709

Abstract

For an autonomous system to be helpful to humans and to pose no unwarranted risks, it needs to align its values with those of the humans in its environment in such a way that its actions contribute to the maximization of value for the humans. We propose a formal definition of the value alignment problem as *cooperative inverse reinforcement learning* (CIRL). A CIRL problem is a cooperative, partial-information game with two agents, human and robot; both are rewarded according to the human's reward function, but the robot does not initially know what this is. In contrast to classical IRL, where the human is assumed to act optimally in isolation, optimal CIRL solutions produce behaviors such as active teaching, active learning, and communicative actions that are more effective in achieving value alignment. We show that computing optimal joint policies in CIRL games can be reduced to solving a POMDP, prove that optimality in isolation is suboptimal in CIRL, and derive an approximate CIRL algorithm.

1 Introduction

"If we use, to achieve our purposes, a mechanical agency with whose operation we cannot interfere effectively . . . we had better be quite sure that the purpose put into the machine is the purpose which we really desire." So wrote Norbert Wiener (1960) in one of the earliest explanations of the problems that arise when a powerful autonomous system operates with an incorrect objective. This *value alignment* problem is far from trivial. Humans are prone to mis-stating their objectives, which can lead to unexpected implementations. In the myth of King Midas, the main character learns that wishing for 'everything he touches to turn to gold' leads to disaster. In a reinforcement learning context, Russell & Norvig (2010) describe a seemingly reasonable, but incorrect, reward function for a vacuum robot: if we reward the action of cleaning up dirt, the optimal policy causes the robot to repeatedly dump and clean up the same dirt.

A solution to the value alignment problem has long-term implications for the future of AI and its relationship to humanity (Bostrom, 2014) and short-term utility for the design of usable AI systems. Giving robots the right objectives and enabling them to make the right trade-offs is crucial for self-driving cars, personal assistants, and human–robot interaction more broadly.

The field of *inverse reinforcement learning* or IRL (Russell, 1998; Ng & Russell, 2000; Abbeel & Ng, 2004) is certainly relevant to the value alignment problem. An IRL algorithm infers the reward function of an agent from observations of the agent's behavior, which is assumed to be optimal (or approximately so). One might imagine that IRL provides a simple solution to the value alignment problem: the robot observes human behavior, learns the human reward function, and behaves according to that function. This simple idea has two flaws. The first flaw is obvious: we don't want the robot to adopt the human reward function as its own. For example, human behavior (especially in the morning) often conveys a desire for coffee, and the robot can learn this with IRL, but we don't want the robot to want coffee! This flaw is easily fixed: we need to formulate the value

*{dhm, anca, pabbeel, russell}@cs.berkeley.edu

30th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain.

Image credit: Cooperative Inverse Reinforcement Learning (arXiv:1606.03137)

   a. Cooperative inverse reinforcement learning (2016). Dylan Hadfield-Menell, Stuart Russell, Pieter Abbeel, and Anca Dragan. In *NeurIPS-16*.
   b. An Efficient, Generalized Bellman Update For Cooperative Inverse Reinforcement Learning (2018). Dhruv Malik, Malayandi Palaniappan, Jaime Fisac, Dylan Hadfield-Menell, Stuart Russell, and Anca Dragan. In *ICML-18*.
   c. The assistive multiarmed bandit (2019). Lawrence Chan, Dylan Hadfield-Menell, Siddhartha Srinivasa, and Anca Dragan. In *HRI-19*.

4. **The 'off-switch' and 'obedience' papers** — These two papers raise the issue of incorrigibility (an AI system resisting shutdown and repair) as previously defined by Soares et al. "The off-switch game" provides a partial solution to corrigibility in the form of epistemic humility on the part of the AI system (it allows itself to be shut down because it believes the human shutting it down knows best). This solution does not fully resolve the issue of incorrigibility, because if the AI system has a misspecified prior and comes to believe, incorrectly, that it already has perfect knowledge of human preferences, it can still resist shutdown. Nonetheless, we believe this paper has "taken a

bite out of" incorrigibility, and also makes incorrigibility easier to talk about in technical terms by pointing out which assumptions of the paper are valid or invalid. "Should robots be obedient?" illustrates how optimal performance from an AI system involves neither perfect obedience nor perfect obstinance. This principle was of course already well-known in application-specific cases (e.g., semi-autonomous vehicle control does not grant arbitrary overrides to the human driver), however, this paper makes the non-monotonic performance/obedience trade-off easier for technical researchers to begin talking and thinking about in general terms. This issue is crucial to discourse on how future AI technology will integrate with society: it means that economic pressures toward efficiency have a fundamental tendency to yield AI systems that sometimes ignore the instructions of their human users.

a. The off-switch game. (2017) Dylan Hadfield-Menell, Anca Dragan, Pieter Abbeel, and Stuart Russell. In *IJCAI-17*.
b. Should robots be obedient? (2017) Smitha Milli, Dylan Hadfield-Menell, Anca Dragan, and Stuart Russell. In *IJCAI-17*.

5. **Inverse Reward Design** — These papers present a version of assistance games where the human-supplied reward is viewed as noisy, partial evidence of the true reward function, probably approximately valid mainly on observed training trajectories but not necessarily in unseen parts of the state space. This enables the robot to have the right kind of uncertainty about the reward and leads to appropriately risk-averse behavior.
    a. Inverse Reward Design (2017). Dylan Hadfield-Menell, Smitha Milli, Pieter Abbeel, Stuart Russell J. Russell, Anca Dragan. In *NeurIPS-17*.
    b. Active Inverse Reward Design (2018). Sören Mindermann, Rohin Shah, Adam Gleave, Dylan Hadfield-Menell. In *Workshop on Goal Specifications for Reinforcement Learning*.

6. **Assistance games with humans** — These papers investigate solutions to assistance games where the "human" player is assumed to conform to an empirically motivated model of actual human decision making, rather than being a perfectly rational agent. In the first paper, "Pragmatic-Pedagogic Value Alignment", the human adopts a pedagogical approach to training the robot, and the robot interprets the human's instructions and demonstrations pragmatically. The paper received a second-place Blue Sky Ideas award at the 2017 International Symposium on Robotics Research (ISRR). Four other papers represent the first forays into assistance games with actual humans, confirming our assumptions that real humans are more complex than simple models allow for but showing that progress is nonetheless possible. The "LESS is more" paper won the Best Technical Paper Award at the 2020 ACM/IEEE International Conference on Human-Robot Interaction (HRI).
    a. Pragmatic-pedagogic value alignment. (2019) Jaime Fisac, Monica Gates, Jessica Hamrick, Chang Liu, Dylan Hadfield-Menell, et al. In *ISRR-19*.
    b. On the Utility of Learning about Humans for Human-AI Coordination (2019). Micah Carroll, Rohin Shah, Mark K. Ho, Tom Griffiths, Sanjit Seshia, Pieter Abbeel, and Anca Dragan. In *NeurIPS-19*.
    c. On the Feasibility of Learning, Rather than Assuming, Human Biases for Reward Inference (2019). Rohin Shah, Noah Gundotra, Pieter Abbeel, and Anca Dragan. In *ICML-19*.
    d. Where do you think you're going? Inferring beliefs about dynamics from behavior (2018). Sid Reddy, Anca Dragan, and Sergey Levine. In *NeurIPS-18*.

e. LESS is More: Rethinking Probabilistic Models of Human Behavior (2020). Andreea Bobu, Dexter Scobee, Jaime Fisac, S. Shankar Sastry, and Anca Dragan. In *HRI-20*.

7. **Misspecification** — In learning preferences and goals from human physical behavior, model misspecification is somewhat inevitable. These two papers investigate the issue both theoretically and empirically, considering both misspecification of the human's decision process and the human's space of possible objectives.
   a. Literal or Pedagogic Human? Analyzing Human Model Misspecification in Objective Learning (2020). Smitha Milli and Anca Dragan. In *UAI-19*.
   b. Quantifying Hypothesis Space Misspecification in Learning from Human-Robot Demonstrations and Physical Corrections (2020). Andreea Bobu, Andrea Bajcsy, Jaime Fisac, and Anca Dragan. *IEEE Transactions on Robotics* 36(3), 835-854.

8. **Theoretical unification of preference learning methods** — We show that two frameworks, reward-rational choice and assistance POMDPs (both of which are restrictions of general assistance games) capture a great many existing frameworks for reward and preference learning. In addition, they resolve many confusions within those frameworks and enable certain desirable classes of behavior to emerge naturally as solutions rather than having to be preprogrammed by human designers. Note: the Assistance POMDP paper is under review but not publicly available.
   a. Reward-rational (implicit) choice: A unifying formalism for reward learning (2020). Hong Jun Jeon, Smitha Milli, and Anca Dragan. Under review.

9. **Negotiable Reinforcement Learning** — This work aims to make it easier to negotiate over the policies of powerful AI systems, which makes it easier to share control of those systems and to avoid competitive arms races. Differences in beliefs across the parties are a major factor in negotiations, and the NRL line of work is the first to account for belief differences between principals in sequential decision making. The result generalizes Harsanyi's social aggregation theorem in a surprising way.
   a. Negotiable reinforcement learning for Pareto-optimal sequential decision-making (2018). Nishant Desai, Andrew Critch, and Stuart Russell. In *NeurIPS-18*.

10. **ARCHES** — This report by Andrew Critch and David Krueger attempts to explain the relevance of twenty-nine AI research directions to existential risk, and how they interrelate. It also introduces the concept of *prepotence*, a property weaker than superintelligence, which is more likely to occur before super intelligence, and sufficient to pose a substantial (arguably inevitable) existential threat.

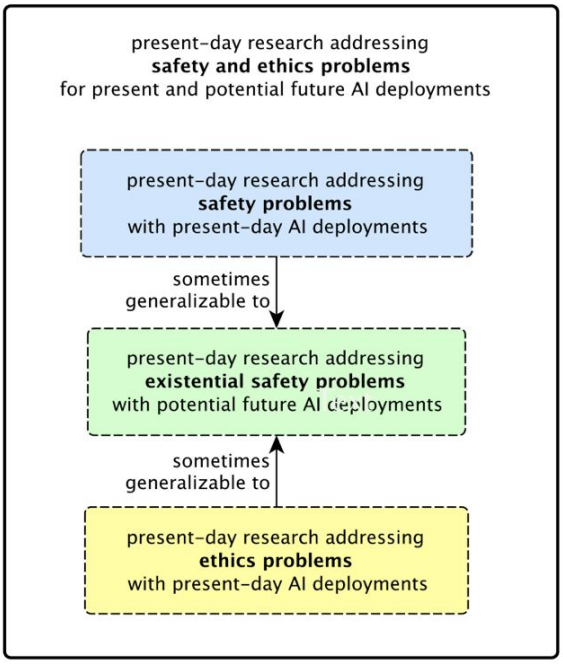


Image credit: ARCHES, p. 9

a. AI Research Considerations for Human Existential Safety (ARCHES) (2020). Andrew Critch and David Krueger. arXiv:2006.04948.

11. **Bounded Löbian Cooperation** — (Started at MIRI) Powerful AI technologies are likely to be created in multiple jurisdictions by multiple diverse stakeholders, in which case cooperation between these systems and their creators will be necessary to sustain global security. This paper shows definitively how two systems with bounded computational resources can achieve cooperation through transparency, by collapsing the infinite regress of metacognition between them (I'll cooperate if I know you will, which you'll only do if you know I will, which I'll only do if I know you will, which...) into a stable state of mutual trust. This result was conjectured at MIRI, but was difficult to formalize at a level of rigor acceptable for peer review. Andrew Critch's formalization was first developed at MIRI and carried through peer review at CHAI.
    a. A parametric, resource-bounded generalization of Löb's theorem, and a robust cooperation criterion for open-source game theory (2019). Andrew Critch. *Journal of Symbolic Logic*, 84(4), 1368-1381.

12. **Adversarial Policies: Attacking Deep Reinforcement Learning** — This paper demonstrates a serious failure mode in state-of-the-art continuous control policies which seem robust when tested using other evaluation methods. It was featured in MIT Technology Review and Two Minute Papers, and briefly in Science News and Nature News. This work will encourage the deep RL community to focus more on robustness and worst-case performance — long a focus in control theory and related communities. Additionally, it provides a clear empirical demonstration of a commonly held view at CHAI: that AI systems which seem reliable may harbor serious failure modes.
    a. Adversarial policies: Attacking deep reinforcement learning (2020). Adam Gleave, Michael Dennis, Cody Wild, Neel Kant, Sergey Levine, and Stuart Russell. In *ICLR-20*.

13. **Preferences implicit in the state of the world** — This paper shows that, contrary to naive interpretations of G.E. Moore's naturalistic fallacy, it *is* possible to infer human preferences from observation of a single world state, and not just from observations of human behavior. This is because the world state *results* (in part) from human behaviour. The state therefore provides evidence of what those preferences might be. This explains why the status quo bias ("doing nothing is a reasonably safe thing to do") is valid and provides a formal grounding for impact measures without requiring a separate "low-impact" principle.
    a. Preferences implicit in the state of the world (2019). Rohin Shah, Dmitrii Krasheninnikov, Jordan Alexander, Pieter Abbeel, and Anca Dragan. In *ICLR-19*.

14. **The Alignment Problem** — Brian Christian, author of *The Most Human Human* and *Algorithms To Live By*, has been affiliated with CHAI since early 2017 and regularly attends CHAI seminars and

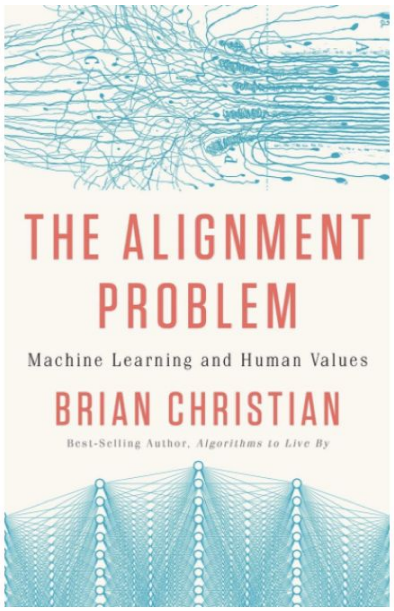


Image credit: The Alignment Problem by Brian Christian (Penguin Random House, 2020)

workshops. His new book narrates and documents the emergence of new perspectives on AI systems — at CHAI and elsewhere — that are moving beyond the standard model, toward aligning with human values.

a. *The Alignment Problem: Machine Learning and Human Values* (2020, October). Brian Christian. W. W. Norton.

CHAI alumni outputs

Below is a list of students who received significant training from CHAI or CHAI PIs, who we believe are well poised and in fact likely to make substantial contributions to the alignment of AI technology with human values and societal-scale safety. They have accepted top-tier and influential positions. Note: the students are ordered by seniority, not by merit.

1. **Prof. Dorsa Sadigh — Stanford** (PhD 2017, 62 publications, 1,201 citations, 76 highly influential, h-index 17) is Assistant Professor of Computer Science and of Electrical Engineering at Stanford University. As a PhD student, she connected with CHAI from Shankar Sastry's lab, through a collaboration with Anca Dragan seeking to formally incorporate humans into her work. Dorsa is now co-director of the AI Safety Center at Stanford, and is focused on research highly relevant to (and arguably necessary for) AI alignment. Her many awards include an invitation to give the Gilbreth Lecture at National Academy of Engineering. Her active reward learning work has also been featured on NPR and in the Wall Street Journal and the Atlantic magazine. She was a plenary speaker at the 2020 CHAI workshop. Since July 2016, 33 of her 44 papers have been on topics directly related to CHAI research goals. Many involve inferring human goals and preferences in assistance-game-like settings, particularly in the context of autonomous and semi-autonomous vehicles and assistive robotics. Here are four examples:

Image credit: Dorsa Sadigh, 2020

   - **When Humans Aren't Optimal: Robots that Collaborate with Risk-Aware Humans (2020)**. Minae Kwon, Erdem Bıyık, Aditi Talati, Karan Bhasin, Dylan P. Losey, Dorsa Sadigh. *ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, March 2020. This paper applies a well-known Risk-Aware human model from behavioral economics called Cumulative Prospect Theory to human-robot interaction (HRI). User studies offer supporting evidence that the Risk-Aware model more accurately predicts suboptimal human behavior, resulting in safer and more efficient human-robot collaboration. It extends existing rational human models so that collaborative robots can anticipate and plan around suboptimal human behavior in HRI.
   - **Shared Autonomy with Learned Latent Actions (2020)**. Hong Jun Jeon, Dylan Losey, Dorsa Sadigh. In *Proceedings of Robotics: Science and Systems (RSS), July 2020*. This paper demonstrates that combining intuitive embeddings from learned latent actions with

robotic assistance from shared autonomy enables precise assistive manipulation in robot assistance of persons with disabilities in everyday tasks. They adopt learned latent actions for shared autonomy by proposing a new model structure that changes the meaning of the human's input based on the robot's confidence of the goal. They show convergence bounds on the robot's distance to the most likely goal, and develop a training procedure to learn a controller that is able to move between goals even in the presence of shared autonomy.

- **Learning Reward Functions from Diverse Sources of Human Feedback: Optimally Integrating Demonstrations and Preferences (2020).** Erdem Bıyık, Dylan P. Losey, Malayandi Palan, Nicholas C. Landolfi, Gleb Shevchuk, Dorsa Sadigh. arXiv 2006.14091. *Submitted to The International Journal of Robotics Research (IJRR).* This paper (an extended version of an RSS 2019 paper) moves reward learning from humans away from the usual single-fixed-protocol approach and shows that (1) there are multiple ways in which reward information can flow from humans to machines and (2) combining them can give better results in practical applications. We view this as an important step towards the general capability for machines to extract information about human preferences from the environment, which includes structures, artefacts, arrangements, documents, media, etc., as well as direct observation of human choice behavior.
- **Altruistic Autonomy: Beating Congestion on Shared Roads (2018).** Erdem Bıyık, Daniel A. Lazar, Ramtin Pedarsani, Dorsa Sadigh. In *Proceedings of the 13th International Workshop on Algorithmic Foundations of Robotics (WAFR).* This paper is one of several that investigates the composition of many human-driven and autonomous vehicles and studies mechanism design problems to create socially optimal solutions for humans. It highlights the importance of the altruistic element in human preferences of socially optimal solutions are to be achieved. We view this paper as an early example of analyzing a problem that will become ubiquitous: the interaction of many humans and many AI systems. By learning how to design incentive mechanisms that avoid negative interactions, we hope to avoid potentially catastrophic outcomes from unanticipated interactions among many uncoordinated AI systems (as noted in the ARCHES paper).

In addition, she has examined the general challenge of creating formally verified AI systems, which is particularly important for CHAI's planned research area in embedded agents.

- Formalizing and Guaranteeing Human-Robot Interaction (2020). Hadas Kress-Gazit, Kerstin Eder, Guy Hoffman, Henny Admoni, Brenna Argall, Ruediger Ehlers, Christoffer Heckman, Nils Jansen, Ross Knepper, Jan Křetínský, Shelly Levy-Tzedek, Jamy Li, Todd Murphey, Laurel Riek, Dorsa Sadigh. arXiv 2006.16732.
- Towards Verified Artificial Intelligence (2016). Sanjit A. Seshia, Dorsa Sadigh, S. Shankar Sastry. arXiv 1606.08514.

2. **Prof. Jaime Fernandez Fisac — Princeton** (PhD 2019, 36 publications, 697 citations, 37 highly influential, h-index 14) just completed a year as a research scientist at Waymo, and has begun his post as Assistant Professor of Electrical Engineering at Princeton. Jaime is developing tools to safely deploy robotic & AI systems in the physical world, with the goal to ensure that autonomous

systems such as self-driving cars, delivery drones, or home robots can operate and learn in open spaces with humans while satisfying safety constraints at all times. Learning human rewards, preferences, and constraints is central to his work and he has made significant contributions to the new model. Jaime's interest in CHAI's work and mission began when he attended Russell's 2016 course on human-compatible AI, and continued through his regular attendance of the CHAI seminar and significant intellectual contributions to ARCHES. Since July 2016, 12 of his 21 papers have been on topics directly related to CHAI research goals. (Most of the other papers deal with more classical notions of safe AI systems.) He has contributed to core technical papers on assistance games (the first two listed below), to joint work with Dorsa Sadigh on autonomous and semi-autonomous driving in the presence of other human drivers, and to formal modelling (and mismodelling) of humans in assistance games.

Image credit: Jaime Fernández Fisac, 2019

- An Efficient, Generalized Bellman Update For Cooperative Inverse Reinforcement Learning (2018). Dhruv Malik, Malayandi Palaniappan, Jaime Fisac, Dylan Hadfield-Menell, Stuart Russell, and Anca Dragan,.' In *ICML-18*. See item [3] in "Specific Outputs" above.
- Pragmatic-pedagogic value alignment. (2019) Jaime Fisac, Monica Gates, Jessica Hamrick, Chang Liu, Dylan Hadfield-Menell, et al. In *ISRR-19*. See item [6] in "Specific Outputs" above.
- Quantifying Hypothesis Space Misspecification in Learning from Human-Robot Demonstrations and Physical Corrections (2020). Andreea Bobu, Andrea Bajcsy, Jaime Fisac, and Anca Dragan. *IEEE Transactions on Robotics* 36(3), 835-854. See item [7] in "Specific Outputs" above.

Image credit: Dylan Hadfield-Menell, 2018

Prof. Dylan Hadfield-Menell — MIT (PhD expected 2020, 33 publications, 519 citations, 27 highly influential, h-index 10) developed important results early in the exploration of the assistance game as a path to provable safety, notably CIRL and inverse reward design, under the advisement of Stuart Russell, Anca Dragan, and Pieter Abbeel. He will join MIT as Assistant Professor of EECS as of Fall 2021, after spending the upcoming year as a Research Scientist at Facebook. Dylan is, as far as we know, the first faculty member hired at a major university whose job talk focused on misalignment risks and the new model.

Rohin Shah — Deepmind (PhD expected 2020, 7 publications, 72 citations, 5 highly influential, h-index 4... not including more than 100 issues of the Alignment Newsletter sent to over 1700 subscribers) has accepted an offer to join DeepMind as a research scientist. Originally a programming language theory student working with Ras Bodik, Rohin switched to AI safety and joined CHAI in Fall 2017. Rohin's focus is on

intent alignment and human-machine cooperation. His research contributions, particularly "Preferences implicit in the state of the world," are likely to be considered foundational.

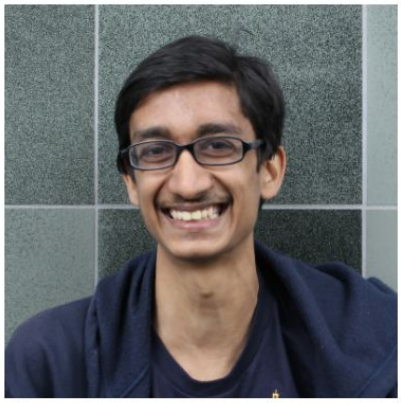


Image credit: Rohin Shah, 2016

How CHAI contributes to student training in general

The UC Berkeley AI Research Lab (BAIR) admissions committee, which includes several CHAI PIs, has observed an increasing number of applicants specifically mentioning CHAI in their statements. The existence of CHAI helps attract strong students interested in human-centered / human-compatible AI. This semester (fall 2020), a record 6 new PhD students joined CHAI to study with Stuart Russell and Anca Dragan.

Every student engaged with CHAI, including undergraduates, graduate students, and interns, has access to four types of formal CHAI training activities:

1. **Advising** by one or more of Stuart Russell, Anca Dragan, Pieter Abbeel, Andrew Critch, other CHAI affiliate faculty, and CHAI graduate students
2. **Weekly seminar meetings** (one technical and one interdisciplinary);
3. **Taking or assistant-teaching CHAI-specific graduate courses**. Graduate student instructors for the courses consistently report that designing, running and attending the new courses pushed them to better grasp the field. (The classes have also helped to interest more students in CHAI's work, e.g. Smitha Milli, Jaime Fisac, Neel Alex, and others).
   a. CS 294-125, Human-Compatible AI. Spring 2016, by Stuart Russell and Anca Dragan.
   b. CS 294-149, Safety and Control for Artificial General Intelligence. Fall 2018, by Andrew Critch and Stuart Russell.
   c. CS 294-166, Foundations for Beneficial AI. Spring 2020, by Lara Buchak, Wesley Holliday (Philosophy), Shachar Kariv (Economics), and Stuart Russell
4. **Selecting and mentoring interns**; mentors consistently report the experience as helpful both in developing their management skills and in advancing their work.

The CHAI graduate students particularly consider the less formal aspects of CHAI to be even more important for their training and development than the formal activities, including students mentoring one another; interacting with the faculty and other students at BAIR, Earth's no. 1 public AI lab, where CHAI is intentionally embedded (until COVID-19) in their new consolidated 14,000 square foot facility at Berkeley Way West; fostering high-value disagreement; and a sense of shared purpose and freedom to pursue morally-driven research.

Students are also exposed to a diversity of views and approaches to AI safety, via speakers in our technical seminars including David Duvenaud (Toronto), Eric Drexler (FHI), Catherine Olsson (then of OpenAI), El Mahdi El Mhamdi (EPFL), Jacob Steinhardt (UC Berkeley), Michael Littman

(Brown), Andreas Stuhlmüller (Stanford / Ought.org), Aleksander Madry and Natasha Jaques (MIT), Cynthia Rudin (Duke), and Chelsea Finn (Stanford).

Other impacts on the AI research community

Our "other impacts" on the AI community are in pursuit of four goals:

Goal 1: **Values-based community building.** This means connecting with researchers in and outside of AI who are beginning to share our moral concerns regarding existential and societal-scale risks. Community building helps to build and sustain motivation, moral support, and for some researchers, a sense of belonging.

Goal 2: **Intellectual recruitment**. This means stimulating and retaining serious intellectual interest in beneficial AI and societal-scale safety, across disciplines and within AI.

Goal 3: **Legitimacy building.** By being open about our risk-reduction motivation with each other and in our published research and public presentations, we increase its legitimacy among AI researchers, making it easier for others to engage, publish papers, and write proposals.

Goal 4: **Engagement with other societal-scale risks**. This is important not just because these other risks are important, but also because it may lead to new ideas coming into the AI risk field and because it develops a stronger sense of shared commitment to humanity.

CHAI Workshops — These bring together professors, graduate students and researchers that share a strong interest in reducing existential risks from advanced AI, along with some newcomers each year. The list of participants is highly curated from recommendations by past participants. Participation has increased consistently, from 30 in 2017, 50 in 2018, 90 in 2019, and 150 in 2020 (held online due to COVID-19). A little over 50% of participants in 2020 reported making significant changes to their work as a result of the workshop.

The AI Alignment Newsletter — The Alignment Newsletter is a weekly publication, started by Rohin Shah, that contains recent content relevant to AI alignment around the world. It features summaries and analysis of prominent new papers in the field. The Newsletter reaches over 1700 email subscribers, is available in English and Mandarin, and is curated by a team of 12 people. It is also posted on the Alignment Forum and LessWrong. To date, the team has written summaries for almost 1500 technical AI safety papers, all accessible via a spreadsheet.

Organizing AI safety and ethics workshops — Stuart Russell has been part of the core organization of many AI conferences since CHAI's inception. These include co-chairing the Beneficial AI workshop (Puerto Rico 2015, Asilomar 2017, Puerto Rico 2019) and the Hastings Institute workshop series on Control and Responsible Innovation in the Development of Autonomous Machines, and serving on the organizing, steering, and/or program committee of the UN 2018 Conference on AI for Global Good, the First International Workshop on AI Safety Engineering, the 2018 and 2019 AAAI/ACM Conference on AI, Ethics, and Society, the 2019 Global Forum on AI for Humanity, and other AI safety/ethics workshops at IJCAI 2016, AAAI 2016, AAAI 2019, and CogSci 2017. In addition, Adam Gleave co-organized the Human Aligned AI Social at NeurIPS 2019; Dylan Hadfield-Menell organized the workshops Reliable Machine Learning in the Wild (NeurIPS 2016, ICML 2017) and Aligned AI (NeurIPS 2018); and Rachel Freedman contributed to the workshop for the AI Safety Landscape initiative, and served on the program committee for AISafety 2020 at IJCAI.

Mentorship and advising outside CHAI — We make extensive efforts to encourage early-stage researchers. Several CHAI students and staff serve as "ambassadors" for AI safety-interested attendees of the Effective Altruism Global conferences, as well as MIRI's AIRCS workshop attendees. In addition, PhD student Adam Gleave volunteered with 80,000 Hours, an organization giving career advice to promising individuals to have a large social impact. CHAI Director of Strategic Research and Partnerships Caroline Jeanmaire was an "EAG Ambassador" to six Fellows during EAGx 2020. PhD student Dylan Hadfield-Menell has provided remote advising and collaboration to help interested graduate students at other schools become active in AI safety. Michael Dennis was one of the speakers at the Human-Aligned AI Summer School in Prague 25th – 28th July 2019. Finally, Rachel Freedman, Rohin Shah, and Stuart Russell gave detailed technical feedback on Brian Christian's book *The Alignment Problem*.

Involvement in other AI Safety-Related Organizations — Stuart Russell has served on the Advisory Boards of the Center for the Study of Existential Risk (University of Cambridge), the Machine Intelligence Research Institute (Berkeley), and the Future of Life Institute (MIT/Harvard); on the AAAI Committee on Ethics and Social Impact of AI; on the Advisory Board of the Berggruen Institute "AI and the Human" program; and on the advisory Committee of the UC Berkeley Center for Long-Term Security program on AI and Security. He co-chairs the UC Presidential Working Group on AI, developing AI policy for the largest US university system.

Image credit: TechCrunch, 2019

Invited talks, podcast interviews — Core CHAI faculty have given keynote lectures at major conferences and meetings. The frequency and high level of the invitations suggest that the AI community (and the broader intellectual community) is hearing the message. Since 2016, CHAI-related talks by the PIs have run into the hundreds. High-profile keynotes include the

International Joint Conference on Artificial Intelligence, the Association for the Advancement of Artificial Intelligence Conference (twice), the Conference on Uncertainty in Artificial Intelligence, the European Conference on Artificial Intelligence, the Conference on Robot Learning, the International Conference on Intelligent Robots and Systems, the International Conference on Automated Planning and Scheduling, and the Turing Lecture (UK). Podcasts on CHAI themes include those with the World Economic Forum, Financial Times, Sean Carroll, Sam Harris, Lex Fridman, the Future of Life Institute, and 80,000 hours.

Seeding the potential for more AI risk oriented research centers — Through staying in touch with new and existing faculty at other universities, we hope to provide moral and intellectual support for more AI research centers oriented on existential and societal-scale risks.

- At the University of Toronto: Gillian Hadfield (CHAI Affiliate, Director of the Schwartz Reisman Institute for Technology and Society) and Sheila McIlraith, David Duvenaud, and Roger Grosse (CS). All of them attended the 2020 CHAI workshop and are in the process of submitting a large ($12-24M) institute proposal for AI safety and governance.
- At Stanford University: Dorsa Sadigh and Stefano Ermon, both CS faculty, have attended and spoken at CHAI workshops.
- At Princeton University: Tom Griffiths and Tania Lombrozo (CHAI PIs), Lara Buchak (Philosophy, CHAI Affiliate), and Jaime Fernández Fisac (EE, CHAI alumnus).
- At MIT: Dylan Hadfield-Menell (assistant professor and CHAI alumnus), Josh Tenenbaum (Brain and Cognitive Sciences), and Max Tegmark (Physics).

Contributing to responsible AI development within industry — Members of our team engage with industry on the topics of advanced risks from artificial intelligence as well as other societal-scale risks; for example, the Partnership on AI initiative to establish responsible publishing norms for AI researchers, the Partnership on AI working group on recommender systems, and the 2019 DeepMind AGI Safety Workshop.

Other impacts

Our contributions beyond the AI community include contributions to public awareness of the risks from AI systems, contributions to world leaders' awareness, and connecting with China.

Contributions to Public Awareness of AI Existential Risk

CHAI's interest in public awareness as a vehicle for reducing societal-scale and existential risk from AI is well captured by the following 2008 quote from Nobel Prize winner Paul Berg in *Nature* (source). Berg was one of 5 co-organizers of the 1975 Asilomar Conference on Recombinant DNA Molecules:

> "... there is a lesson in Asilomar for all of science: the best way to respond to concerns created by emerging knowledge or early-stage technologies is for scientists from publicly funded institutions to find common cause with the wider public about the best way to regulate—as early as possible. Once scientists from corporations begin to dominate the research enterprise, it will simply be too late."

CHAI's team has been raising awareness of the risks from advanced AI systems through talks and conferences. Stuart Russell appeared on dozens of prominent worldwide media (see the list). He also appeared in documentaries on this topic aimed at a large audience, including Do You Trust This Computer? (2018), iHUMAN (2019), and We Need to Talk About AI (2020).

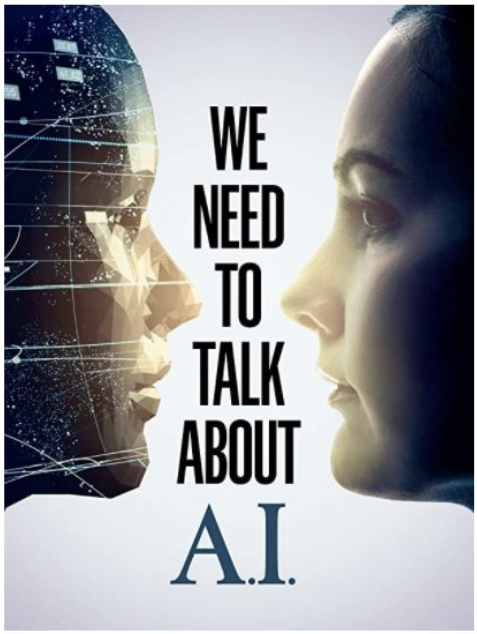


Image credit: We Need to Talk About AI, 2020

Russell has also been raising awareness of lethal autonomous weapons through speaking, writing, and media interviews (see the list). He also created the award-winning short film *Slaughterbots* (2017) (over 75 million views per Jaan Tallinn), and appeared in the New York Times documentary *Killing in the Age of Algorithms* (2019) on the future of AI and warfare. The connection to existential risk is two-fold: first, if we cannot set the precedent of restricting AI systems that can decide to kill humans, it may prove more difficult to restrict other kinds of AI systems; and second, if we do lose control over poorly or maliciously designed AI systems, the availability of large numbers of computer-controlled lethal weapons can only make the problem worse.

Contributions to World Leaders' Awareness

World leaders need to understand both individual-scale and societal-scale risks posed by artificial intelligence, because of their involvement in policy decisions.

Image credit: TED, 2017

CHAI faculty have been invited to give keynote talks at numerous global fora on CHAI-related matters. These include: National Academy meetings on AI (4), a Royal Society meeting on AI, the annual meetings of the World Economic Forum (4), the Nobel Foundation (3), the OECD (2), the World Government Summit (2), the UN AI Global Summit, the Global Forum on AI for Humanity, the American Association for the Advancement of Science, the American Physical Society Annual Meeting, the Nanjing Forum, and the World Conference on AI (Shanghai). CHAI is co-organizing a workshop series at the World Economic Forum in San Francisco, bringing together economists, science fiction writers, and computer scientists from April to December 2020 to imagine a future of shared prosperity with advanced AI systems.

Over the course of CHAI's existence, AI-related policy making and governance efforts have shifted from creating "principles" to creating policies. CHAI contributions to policy include: providing feedback on the EU Trustworthy AI Assessment List (2019) (link to our contribution), and on the US Guidance for Regulation of AI applications (2020) (link to our contribution).

Andrew Critch has given presentations in Singapore, advising the Prime Minister's Office on the potential impacts of AI on Singaporean society. Stuart Russell has also provided direct advice to leaders of national and international organizations as follows:

- **United Nations:** Secretary General, High Representative for Disarmament, Interregional Crime and Justice Research Institute, International Criminal Court
- **World Economic Forum**: Global Agenda Council on AI and Robotics (vice-chair), Global AI Council (member), Global Security Group (member)
- **EU:** EU Commission (advisor), Centre for European Policy Studies (Scientific Advisory Board)
- **US:** Assistant Secretary of State for Democracy, Human Rights, and Labor, Secretary of Defense Office of Net Assessment, Chief of Naval Operations Strategic Studies, National Intelligence Council, DARPA Director, IARPA, CIA, Chief of Staff of the Army, Strategic Studies Group, Army War College, US/China Track II arms control negotiations, Department of State Speaker Program in China, Federal Communication Commission Technical Advisory Council, JASONs.
- **Japan:** Ministry of Economy, Trade and Industry

- **UK:** Office of the Prime Minister, Department of Digital, Culture, Media, and Sport, House of Lords, Center for Data Ethics and Innovation
- **France:** President of France, members of the National Assembly and the Senate, International Scientific Board for AI (member)
- **Singapore:** Office of the Prime Minister
- **UAE:** Minister for AI

Connecting with China

Our aim here is primarily to help policymakers avoid an arms race that might precipitate unsafe AI systems deployment, and to encourage further development of the new model for AI in China.

- Stuart Russell spoke at the World Peace Forum organized by China in June 2020 in a panel that notably included Ya-Qin Zhang (Dean of the Institute for AI Industry Research of Tsinghua University) and Xue Lan (Dean of the Schwarzman College of Tsinghua University). He also led an extended meeting with Madam Fu Ying, who is former Vice-Minister of Foreign Affairs and current chairperson of the National People's Congress Foreign Affairs Committee.
- *Human Compatible* will be published in Mandarin in China in October 2020.
- CHAI has been intentional in its inclusion of Chinese collaborators. In 2020, we invited 24 Chinese participants to the CHAI workshop. CHAI hosted the Tianxia Fellowship for a virtual meeting in April 2020.

Image credit: Human Compatible by Stuart Russell (CITIC Press Group, 2020)

Future plans

CHAI will expand its research on the new model, particularly the multihuman and multirobot versions of assistance games (AGs); work on a foundational theory of agent design and embedded agency; and begin turning the new model into a practical replacement for standard-model technology. We will strengthen connections to the social and human sciences on topics such as aggregation of preferences across multiple individuals, formal safety models of the sociotechnical context of AI systems, and plasticity of human preferences. At the end of this section we outline plans for expanded training, field-building, policy and thought leadership.

Basic AG theory

The standard model of AI (search, planning, MDPs, POMDPs, RL, etc.) builds on long-established concepts and results such as the Markov property, Bellman's optimality principle for MDPs, and Astrom's separation principle for POMDPs. (The latter justifies optimal agents composed of perception, state update, and decision elements). We have just begun to carry out this research for AGs (see [3a,c] above) and there are many open questions:

- Is there a general separation principle for agents in partially observable AGs?
- If so, are improvements in perception and state-update elements always beneficial?
- Can we design model-free (policy and Q-function) AG agents and the associated RL algorithms (policy-search and Q-learning)?
- What behavior profiles can, in principle, be exhibited by agents in the new model but not the standard model?
- Can the theory of bounded optimality for single agents be extended to encompass AG agents? How should AG agents operate when they are (1) much less capable (2) roughly as capable (3) much more capable than a human in the same task environment?
- Can some form of universal prior on preferences allow AG agents to avoid misspecification problems? If so, can this be made computationally effective? Are there fundamental tradeoffs between utility (strong prior) and safety (weak prior)?
- When is it better for an AI system to learn preferences rather than imitate behavior?

This last question points to a gaping hole at the center of AI research: we have no solid theory to explain why it is (or seems to be) a good idea to know things, to reason, to plan, etc., as opposed to simply learning a history-dependent policy (as in a recurrent neural network). This is a complex question involving tradeoffs among decision optimality, speed of learning, and speed of decision making. We believe real progress can be made by considering abstract formal models of environments with large state spaces but simple (in the Kolmogorov sense) dynamics, leading to fundamental theorems concerning the architecture of intelligence. A solid mathematical foundation should improve our ability to design safe AI systems with semantically clear and distinct components.

Theory of embedded agency

As noted by Orseau and Ring and in recent work from MIRI, real-world AI systems are, unlike their idealized cousins in MDPs and POMDPs, embedded in their environments: their own computations are part of the environment and external events can modify their computations. This creates opportunities for wireheading (the agent can take actions that interfere with its sensors to create illusory rewards), incentives for legibility (if the agent can modify its internals to become more legible, it is more likely to be trusted by humans), capabilities for metareasoning, and myriad other complications. A full working theory of provably beneficial AI — one with meaningful formal guarantees — needs a better theory of embedded agency to begin addressing these issues. We hope to collaborate with MIRI on this topic.

Cooperation with multiple AI systems

The analysis in the ARCHES paper listed above surfaced a significant risk from heterogeneous AI systems, possibly optimizing for different subsets of humans (e.g., shareholders of different companies), that might produce unanticipated interactions. We have begun and expect to continue with research on zero-shot cooperation. Andrew Critch's work on open-source game theory is one line of work that can be made practical by building on the technology of proof-carrying code, whereby agents advertise formally checkable properties that allow rigorous cooperative contract formation. Another approach is Michael Dennis's work on policy-conditioned beliefs, which enriches standard Bayesian agent design with formal concepts from epistemic game theory and leads to a wide class of agents that naturally cooperate.

Making the new model practical

We believe it is unlikely that the broad AI community will abandon the standard model unless and until there is convincing evidence that the new model can replace it in practice.

On the foundations side, we need to rebuild many branches of AI — including search, game-playing, constraint satisfaction, logical planning, MDPs, POMDPs, RL — to allow for uncertain objectives. This includes finding the "natural" form of partial preference information, the corresponding "protocol" whereby preference information flows from the human, and new, efficient interactive algorithms. Supervised learning of *policies* is a key topic. We will analyze the implicit transfer of human preference information via human-labeled training data and extend algorithms to handle uncertainty in the loss function optimized by the learning algorithm.

We believe it will be persuasive to the field as a whole — and the AI industry in particular — to show that the new model can lead to better systems in practice. Possible targets (perhaps with industry collaboration) include recommender systems (the subject of a NeurIPS 2020 workshop proposal by Dylan Hadfield-Menell); personal digital assistants; personal robotics; personal financial advisors; and interactive design systems (architecture, site layout, etc.).

Social and human sciences: many humans, real humans

Philosophy and the social sciences have for centuries studied the problem of acting on behalf of multiple humans and identified many "extreme failure modes" for simplistic solutions. We have initiated and will expand collaborations with these disciplines, partly facilitated by Russell's receipt of the Andrew Carnegie Fellowship (one of the most prestigious awards in the social sciences and humanities). Issues include interpersonal comparisons of preferences, decisions that affect population size, and human preferences that are altruistic, sadistic, or relativized to the wellbeing/status of others. We believe many critiques of consequentialism can be overcome by nuanced formulations from which Kantian principles emerge as logical consequences.

Many harmful AI outcomes in the real world result from the combined failure of the algorithm and the sociotechnical context in which it is embedded. Racially biased training data is a well-known example (resulting also from objective misspecification); other, more complex failure modes include classifiers whose decisions affect their own future input data (as analyzed recently by Moritz Hardt) as well as moral hazard and adverse selection in insurance. Formal models of the sociotechnical context and the embedded AI system could be of enormous value in revealing new failure modes and providing guidance for safe design and use of AI systems.

We have begun to explore relevant properties of real humans [7, 8 above]. We hope to model the real preferences of populations of humans in restricted settings (e.g., recommender systems) and to deepen our work on hierarchically structured models of human activity, which we feel is the most central aspect of human cognition as it relates to the mapping from preferences to behavior. We also hope to improve on the Boltzmann model of approximate rationality, which ignores the fact that humans are more accurate when making easy decisions.

The final topic is plasticity of human preferences. "Version 0" of the new model assumes stable preferences, which could lead to AG agents that freely modify human preferences. This is a philosophically challenging problem, but recent approaches (e.g., by Pettigrew) are promising.

Training, field-building, policy, thought leadership

We aim for a modest expansion in the numbers of CHAI PhD students and interns. We expect graduate courses to include core technical AI safety, interdisciplinary AI/social-science, and new-model multiagent systems and computational economics.

To help tie together this growing field, we are planning a global (or at least North American) online seminar series involving all the active research centers, and will expand the in-person workshop accordingly. Some CHAI students also plan to develop a podcast series.

On the policy side, we will continue our engagement with the EU policy process, which is well ahead of processes in the US and China. Increasingly, commentators view GDPR as a de facto global standard, which may also be the case for AI-safety-related EU regulations.

Appendix: Publications

Ittai Abraham, Danny Dolev, Joseph Y. Halpern. "Distributed Protocols for Leader Election: A Game-Theoretic Perspective." ACM Transactions on Economics and Computation 7(1) (2019).

Ittai Abraham, Danny Dolev, Ivan Geffner, Joseph Y. Halpern. "Implementing Mediators with Asynchronous Cheap Talk." PODC 2019.

Mayank Agrawal, Joshua C. Peterson, Thomas L. Griffiths. "Using Machine Learning to Guide Cognitive Modeling: A Case Study in Moral Reasoning." CogSci 2019.

Mayank Agrawal, Joshua C. Peterson, Thomas L. Griffiths. "Scaling up psychology via Scientific Regret Minimization." PNAS 2020.

Stefano V. Albrechta, Peter Stone, Michael P. Wellman. "Special issue on autonomous agents modelling other agents: Guest editorial." Artificial Intelligence 285 (2020).

Natasha Alechina, Joseph Y. Halpern, Brian Logan. "Causality, Responsibility and Blame in Team Plans." AAMAS 2017.

Natasha Alechina, Joe Halpern, Ian Kash, Brian Logan. "Incentivising Monitoring in Open Normative Systems." AAAI 2017.

Natasha Alechina, Joseph Y. Halpern, Ian A. Kash, Brian Logan. "Incentive-Compatible Mechanisms for Norm Monitoring in Open Multi-Agent Systems." JAIR 2018.

Gadi Aleksandrowicz, Hana Chockler, Joseph Y. Halpern, Alexander Ivrii. "The Computational Complexity of Structure-Based Causality." JAIR 2017.

Kareem Amin, Nan Jiang, Satinder Singh. "Repeated Inverse Reinforcement Learning." NIPS 2017.

Jacob Andreas, Anca Dragan, Dan Klein. "Translating Neuralese." ACL 2017.

McKane Andrus, Thomas Krendl Gilbert. "Towards a Just Theory of Measurement: A Principled Social Measurement Assurance Program for Machine Learning." AIES 2019.

Andrea Bajcsy, Dylan P. Losey, Marcia K. O'Malley, Anca D. Dragan. "Learning from Physical Human Corrections, One Feature at a Time." HRI 2018.

Somil Bansal, Andrea Bajcsy, Ellis Ratner, Anca D. Dragan, Claire J. Tomlin. "A Hamilton-Jacobi Reachability-Based Framework for Predicting and Analyzing Human Motion for Safe Planning." ICRA 2020.

Jialu Bao, Kun He, Xiaodong Xin, Bart Selman, John E. Hopcroft. "Hidden Community Detection on Two-layer Stochastic Models: a Theoretical Perspective." (Preprint, submitted to TAMC 2020).

Chandrayee Basu, Qian Yang, David Hungerman, Mukesh Singhal, Anca Dragan. "Do You Want Your Autonomous Car to Drive Like You?." HRI 2017.

Chandrayee Basu, Mukesh Singhal, Anca D. Dragan. "Learning from Richer Human Guidance: Augmenting Comparison-Based Learning with Feature Queries." HRI 2018.

Ruairidh M. Battleday, Joshua C. Peterson, Thomas L. Griffiths. "Capturing human categorization of natural images at scale by combining deep networks and cognitive models." (Preprint) 2019.

Sander Beckers, Joseph Y. Halpern. "Abstracting causal models." AAAI 2019.

Sander Beckers, Frederick Eberhardt, Joseph Y. Halpern. "Approximate Causal Abstraction." UAI 2019.

Aaron Bestick, Ruzena Bajcsy, Anca Dragan. "Implicitly Assisting Humans to Choose Good Grasps in Robot to Human Handovers,." International Symposium on Experimental Robotics 2016.

Aaron Bestick, Ravi Pandya, Ruzena Bajcsy, Anca D. Dragan. "Learning Human Ergonomic Preferences for Handovers." ICRA 2018.

Kush Bhatia, Yi-An Ma, Anca D. Dragan, Peter L. Bartlett, Michael I. Jordan. "Bayesian Robustness: A Nonasymptotic Viewpoint." (Preprint) 2019.

Adam Bjorndahl, Joseph Y. Halpern, Rafael Pass. "Reasoning about Rationality." Games and Economic Behavior 104, 146-164 (2017).

Andreea Bobu, Andrea Bajcsy, Jaime F. Fisac, Sampada Deglurkar, Anca D. Dragan. "Quantifying Hypothesis Space Misspecification in Learning from Human-Robot Demonstrations and Physical Corrections." IEEE Transactions on Robotics 2019.

Andreea Bobu, Dexter R.R. Scobee, Jaime F. Fisac, S. Shankar Sastry, Anca D. Dragan. "LESS is More: Rethinking Probabilistic Models of Human Behavior." HRI 2020.

David Bourgin, Falk Lieder, Daniel Reichman, Nimrod Talmon, Tom Griffiths. "The Structure of Goal Systems Predicts Human Performance." CogSci 2017.

Miles Brundage, Shahar Avin, Jasmine Wang, Haydn Belfield, Gretchen Krueger, Gillian Hadfield, Heidy Khlaaf, Jingying Yang, Helen Toner, Ruth Fong, Tegan Maharaj, Pang Wei Koh, Sara Hooker, Jade Leung, Andrew Trask, Emma Bluemke, Jonathan Lebensold, Cullen O'Keefe, Mark Koren, Théo Ryffel, JB Rubinovitz, Tamay Besiroglu, Federica Carugati, Jack Clark, Peter Eckersley, Sarah de Haas, Maritza Johnson, Ben Laurie, Alex Ingerman, Igor Krawczuk, Amanda Askell, Rosario Cammarota, Andrew Lohn, David Krueger, Charlotte Stix, Peter Henderson, Logan Graham, Carina Prunkl, Bianca Martin, Elizabeth Seger, Noa Zilberman, Seán Ó

hÉigeartaigh, Frens Kroeger, Girish Sastry, Rebecca Kagan, Adrian Weller, Brian Tse, Elizabeth Barnes, Allan Dafoe, Paul Scharre, Ariel Herbert-Voss, Martijn Rasser, Shagun Sodhani, Carrick Flynn, Thomas Krendl Gilbert, Lisa Dyer, Saif Khan, Yoshua Bengio, Markus Anderljung. "Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims." (Preprint) 2020.

Frederick Callaway, Tom Griffiths. "Attention in value-based choice as optimal sequential sampling." (Preprint) 2019.

Frederick Callaway, Antonio Rangel, Tom Griffiths. "Fixation patterns in simple choice are consistent with optimal use of cognitive resources." (Preprint) 2020.

Valerio Capraro, Joseph Y Halpern. "Translucent players: Explaining cooperative behavior in social dilemmas." Rationality and Society 31(4), 371-408 (2019).

Micah Carroll, Rohin Shah, Mark Ho, Thomas Griffiths, Sanjit Seshia, Pieter Abbeel, Anca Dragan. "On the Utility of Learning about Humans for Human-AI Coordination." NeurIPS 2019.

Lawrence Chan, Dylan Hadfield-Menell, Siddhartha Srinivasa, Anca Dragan. "The Assistive Multi-Armed Bandit." HRI 2019.

Margaret P. Chapman, Jonathan Lacotte, Aviv Tamar, Donggun Lee, Kevin M. Smith, Victoria Cheng, Jaime F. Fisac, Susmit Jha, Marco Pavone, Claire J. Tomlin. "A Risk-Sensitive Finite-Time Reachability Approach for Safety of Stochastic Dynamic Systems." American Control Conference (ACC) 2019.

Rohan Choudhury, Gokul Swamy, Dylan Hadfield-Menell, Anca D. Dragan. "On the Utility of Model Learning in HRI." HRI 2019.

Andrew Critch, Stuart Russell. "Servant of Many Masters: Shifting priorities in Pareto-optimal sequential decision-making." AIES 2019.

Andrew Critch. "A Parametric, Resource-Bounded Generalization of Löb's Theorem, and a Robust Cooperation Criterion for Open-Source Game Theory." The Journal of Symbolic Logic, Cambridge University Press 2019.

Andrew Critch, David Krueger. "AI Research Considerations for Human Existential Safety (ARCHES)." (Preprint) 2020.

Chris Cundy, Daniel Filan. "Exploring Hierarchy-Aware Inverse Reinforcement Learning." Unpublished (ICML Goals RL Workshop) 2018.

Nishant Desai, Andrew Critch, Stuart J. Russell. "Negotiable Reinforcement Learning for Pareto Optimal Sequential Decision-Making." NeurIPS 2018.

Roel Dobbe, Sarah Dean, Thomas Gilbert, Nitin Kohli. "A Broader View on Bias in Automated Decision-Making: Reflecting on Epistemology and Dynamics." FAT/ML 2018.

Roel Dobbe, Thomas Krendl Gilbert, Yonatan Mintz. "Hard Choices in Artificial Intelligence: Addressing Normative Uncertainty through Sociotechnical Commitments." NeurIPS 2019.

Ravit Dotan, Smitha Milli. "Value-laden Disciplinary Shifts in Machine Learning." (Preprint) 2020.

R. Dubey, T. L. Griffiths. "Reconciling novelty and complexity through a rational analysis of curiosity." Psychological Review, 127(3), 455–476 (2020).

Tom Everitt, Daniel Filan, Mayank Daswani, Marcus Hutter. "Self-Modification of Policy and Utility Function in Rational Agents." AGI 2016.

Daniel Filan, Shlomi Hod, Cody Wild, Andrew Critch, Stuart Russell. "Pruned Neural Networks are Surprisingly Modular." (Preprint, under review NeurIPS 2020).

Jaime F. Fisac, Anayo K. Akametalu, Melanie N. Zeilinger, Shahab Kaynama, Jeremy Gillula, Claire J. Tomlin. "A General Safety Framework for Learning-Based Control in Uncertain Robotic Systems." IEEE Transactions on Automatic Control 2016.

Jaime F. Fisac, Chang Liu, Jessica B. Hamrick, S. Shankar Sastry, J. Karl Hedrick, Thomas L. Griffiths, Anca D. Dragan. "Generating Plans that Predict Themselves." CDC 2016.

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. "Pragmatic-Pedagogic Value Alignment." ISRR 2017.

Jaime F. Fisac, Andrea Bajcsy, Sylvia L. Herbert, David Fridovich-Keil, Steven Wang, Claire J. Tomlin, Anca D. Dragan. "Probabilistically Safe Robot Planning with Confidence-Based Human Predictions." RSS 2018.

Jaime F. Fisac, Neil F. Lugovoy, Vicenç Rubies-Royo, Shromona Ghosh, Claire J. Tomlin. "Bridging Hamilton-Jacobi Safety Analysis and Reinforcement Learning." IEEE 2019.

Rachel Freedman, Jana Schaich Borg, Walter Sinnott-Armstrong, John P. Dickerson, Vincent Conitzer. "Adapting a kidney exchange algorithm to align with human values." Artificial Intelligence, 283 (2020).

David Fridovich-Keil, Andrea Bajcsy, Jaime F. Fisac, Sylvia L. Herbert, Steven Wang, Anca D. Dragan, Claire J. Tomlin. "Confidence-aware motion prediction for real-time collision avoidance." International Journal of Robotics Research 2018.

David Fridovich-Keil, Ellis Ratner, Lasse Peters, Anca D. Dragan, Claire J. Tomlin. "Efficient Iterative Linear-Quadratic Approximations for Nonlinear Multi-Player General-Sum Differential Games." ICRA 2020.

Meir Friedenberg, Joseph Y. Halpern. "Blameworthiness in Multi-Agent Settings." AAAI 2019.

Vael Gates, Thomas L. Griffiths, Anca D. Dragan. "How to Be Helpful to Multiple People at Once." Cognitive Science 44(6) (2020).

Ivan Geffner, Joseph Y. Halpern. "Security in Asynchronous Interactive Systems." (Preprint) 2019.

Thomas Krendl Gilbert, Yonatan Mintz. "Epistemic Therapy for Bias in Automated Decision-Making." AIES 2019.

Adam Gleave, Oliver Habryka. "Multi-task Maximum Entropy Inverse Reinforcement Learning." ICML Goals RL Workshop 2018.

Adam Gleave, Michael Dennis, Cody Wild, Neel Kant, Sergey Levine, Stuart Russell. "Adversarial Policies: Attacking Deep Reinforcement Learning." ICLR 2020.

Adam Gleave, Michael Dennis, Shane Legg, Stuart Russell, Jan Leike. "Quantifying Differences in Reward Functions." (Preprint, under review NeurIPS 2020).

Olaf Graf, Mark Nitzberg. "Solomon's Code: Humanity in a World with Thinking Machines." Pegasus Books 2018.

Thomas L. Griffiths, Frederick Callaway, Michael B. Chang, Erin Grant, Paul M. Krueger, Falk Lieder. "Doing more with less: meta-reasoning and meta-learning in humans and machines." Current Opinion in Behavioral Sciences 29: 24-30 (2019).

Dylan Hadfield-Menell, Anca Dragan, Pieter Abbeel, Stuart Russell. "Cooperative Inverse Reinforcement Learning." NeurIPS 2016.

Dylan Hadfield-Menell, Smitha Milli, Pieter Abbeel, Stuart Russell, Anca Dragan. "Inverse Reward Design." NeurIPS 2017.

Dylan Hadfield-Menell, Anca Dragan, Pieter Abbeel, Stuart Russell. "The Off-Switch Game." IJCAI 2017.

Dylan Hadfield-Menell, McKane Andrus, Gillian Hadfield. "Legible Normativity for AI Alignment: The Value of Silly Rules." AIES 2019.

Dylan Hadfield-Menell, Gillian K. Hadfield. "Incomplete Contracting and AI Alignment." AIES 2020.

Joseph Y. Halpern, Xavier Vilaca. "Rational Consensus (extended abstract)." ACM Symposium on Principles of Distributed Computing 2016.

Joseph Y. Halpern. "Sufficient Conditions for Causality to be Transitive." Philosophy of Science, 83, 213--226 (2016).

Joseph Y. Halpern. "Actual Causality (book)." MIT Press 2016.

Joseph Y. Halpern, Rafael Pass, Lior Seeman. "Computational Extensive-Form Games." EC 2016.

Joseph Y. Halpern. "A Note on the Existence of Ratifiable Acts." Review of Symbolic Logic 2018.

Joseph Y. Halpern, Rafael Pass. "Game Theory with Translucent Players." International Journal of Game Theory 2018.

Joseph Y. Halpern, Lior Seeman. "Is state-dependent valuation more adaptive than simpler rules?." Behavioural Processes 2018.

Joseph Y. Halpern, Rafael Pass. "A Conceptually Well-Founded Characterization of Iterated Admissibility Using an "All I Know" Operator." TARK 2019.

Joseph Y. Halpern, Evan Piermont. "Partial Awareness." AAAI 2019.

Joseph Y. Halpern, Rafael Pass. "Sequential equilibrium in computational games." ACM Transactions on Economics and Computation 2019.

Joseph Y. Halpern, Rafael Pass, Lior Seeman. "The truth behind the myth of the folk theorem." Games and Economic Behavior, 117 (2019).

Joseph Y. Halpern, Rafael Pass, Daniel Reichman. "On the Existence of Nash Equilibrium in Games with Resource-Bounded Players." SAGT 2019.

Mathew Hardy, Tom Griffiths. "Demonstrating the Impact of Prior Knowledge in Risky Choice." (Preprint) 2019.

Robert D. Hawkins, Noah D. Goodman, Adele E. Goldberg, Thomas L. Griffiths. "Generalizing meanings from partners to populations: Hierarchical inference supports convention formation on networks." CogSci 2020.

Donald J. Hejna III, Pieter Abbeel, Lerrel Pinto. "Hierarchically Decoupled Imitation for Morphological Transfer." (Preprint) 2019.

Mark K. Ho, David Abel, Tom Griffiths, Michael L. Littman. "The Value of Abstraction." Current Opinion in Behavioral Sciences, 29:111-116 (2019).

Mark K. Ho, David Abel, Jonathan D. Cohen, Michael L. Littman, Thomas L. Griffiths. "The Efficiency of Human Cognition Reflects Planned Information Processing." AAAI 2020.

Anne S. Hsu, Jay B. Martin, Adam N. Sanborn, Thomas L. Griffiths. "Identifying category representations for complex stimuli using discrete Markov chain Monte Carlo with people." Behavior Research Methods 51:1706–1716 (2019).

Sandy H. Huang, David Held, Pieter Abbeel, Anca Dragan. "Enabling Robots to Communicate their Objectives." RSS 2017.

Sandy H. Huang, Kush Bhatia, Pieter Abbeel, Anca D. Dragan. "Establishing Appropriate Trust via Critical States." IROS 2018.

Sandy H. Huang, Isabella Huang, Ravi Pandya, Anca D. Dragan. "Nonverbal Robot Feedback for Human Teachers." CoRL 2019.

Sushil Jajodia, George Cybenko, V. S. Subrahmanian, Vipin Swarup, Cliff Wang, Michael Wellman. "Adaptive Autonomous Secure Cyber Systems." Springer/Nature Books 2020.

Hong Jun Jeon, Smitha Milli, Anca D. Dragan. "Reward-rational (implicit) choice: A unifying formalism for reward learning." (Preprint) 2019.

Aditi Jha, Joshua Peterson, Thomas L. Griffiths. "Extracting low-dimensional psychological representations from convolutional neural networks." CogSci 2020.

Mark K. Ho, Joanna Korman, Thomas L. Griffiths. "The Computational Structure of Unintentional Meaning." CogSci 2019.

Marc Khoury, Dylan Hadfield-Menell. "Adversarial Training with Voronoi Constraints." (Preprint) 2019.

Marc Khoury, Dylan Hadfield-Menell. "On the Geometry of Adversarial Examples." (Preprint) 2020.

Raphael Köster, Dylan Hadfield-Menell, Gillian K. Hadfield, Joel Z. Leibo. "Silly rules improve the capacity of agents to learn stable enforcement and compliance behaviors." AAMAS 2020.

Dmitrii Krasheninnikov, Rohin Shah, Herke van Hoof. "Combining reward information from multiple sources." NeurIPS Learning with Rich Experience Workshop 2019.

Paul Krueger, Falk Lieder, Tom Griffiths. "Enhancing metacognitive reinforcement learning using reward structures and feedback." CogSci 2017.

Anagha Kulkarni, Siddharth Srivastava, Subbarao Kambhampati. "A unified framework for planning in adversarial and cooperative environments." AAAI 2019.

Thanard Kurutach, Aviv Tamar, Ge Yang, Stuart Russell, Pieter Abbeel. "Learning Plannable Representations with Causal InfoGAN." ICML 2018 Workshop on Planning and Learning.

Minae Kwon, Sandy H. Huang, Anca D. Dragan. "Expressing Robot Incapability." HRI 2018.

Nicholas C. Landolfi, Anca D. Dragan. "Social Cohesion in Autonomous Driving." IROS 2018.

Antonia Langenhoff, Alex Wiegmann, Joseph Y. Halpern, Joshua B. Tenenbaum, Tobias Gerstenberg. "Predicting responsibility judgments from dispositional inferences and causal attributions." (Preprint) 2020.

Michael Laskey, Caleb Chuck, Jonathan Lee, Jeffrey Mahler, Sanjay Krishnan, Kevin Jamieson, Anca Dragan, Ken Goldberg. "Comparing Human-Centric and Robot-Centric Sampling for Robot Deep Learning from Demonstrations." ICRA 2017.

Shihui Li, Yi Wu, Xinyue Cui, Honghua Dong, Fei Fang, Stuart Russell. "Robust multi-agent reinforcement learning via minimax deep deterministic policy gradient." AAAI 2019.

Zun Li, Michael P. Wellman. "Structure Learning for Approximate Solution of Many-Player Games." AAAI 2020.

Falk Lieder, Paul Krueger, Tom Griffiths. "An automatic method for discovering rational heuristics for risky choice." CogSci 2017.

Falk Lieder, Thomas L. Griffiths, Quentin J. M. Huys, Noah D. Goodman. "Empirical evidence for resource-rational anchoring and adjustment." Psychonomic Bulletin & Review 2018.

Falk Lieder, Thomas L. Griffiths, Ming Hsu. "Overrepresentation of extreme events in decision making reflects rational use of cognitive resources." Psychological Review 2018.

Falk Lieder, Amitai Shenhav, Sebastian Musslick, Thomas L. Griffiths. "Rational metareasoning and the plasticity of cognitive control." PLoS Comp. Biol. 2018.

Falk Lieder, Thomas L. Griffiths, Quentin J. M. Huys, Noah D. Goodman. "The anchoring bias reflects rational use of cognitive resources." Psychonomic Bulletin & Review 2018.

Falk Lieder, Thomas L. Griffiths. "Resource-rational analysis: understanding human cognition as the optimal use of limited computational resources." Behavioral and Brain Sciences, 43, E1 (2019).

Falk Lieder, Thomas L. Griffiths. "Advancing rational analysis to the algorithmic level." Behavioral and Brain Sciences, 43, E27 (2020).

Falk Lieder, Owen X. Chen, Paul M. Krueger, Thomas L. Griffiths. "Cognitive prostheses for goal achievement." Nature Human Behaviour 3:1096–1106 (2019).

Chang Liu, Jessica B. Hamrick, Jaime F. Fisac, Anca D. Dragan, J. Karl Hedrick, S. Shankar Sastry, Thomas L. Griffiths. "Goal Inference Improves Objective and Perceived Performance in Human-Robot Collaboration." AAMAS 2017.

Arnon Lotem, Joseph Y. Halpern, Shimon Edelman, Oren Kolodny. "The evolution of cognitive mechanisms in response to cultural innovations." PNAS 2017.

Dhruv Malik, Malayandi Palaniappan, Jaime F. Fisac, Dylan Hadfield-Menell, Stuart Russell, Anca D. Dragan. "An Efficient, Generalized Bellman Update For Cooperative Inverse Reinforcement Learning." ICML 2018.

Negar Mehr, Roberto Horowitz, Anca Dragan. "Inferring and Assisting with Constraints in Shared Autonomy." CDC 2016.

John Miller, Smitha Milli, Moritz Hardt. "Strategic Classification is Causal Modeling in Disguise." FAT* 2019.

Smitha Milli, Dylan Hadfield-Menell, Anca Dragan, Stuart Russell. "Should Robots be Obedient?." IJCAI 2017.

Smitha Milli, Falk Lieder, Tom Griffiths. "When Does Bounded-Optimal Metareasoning Favor Few Cognitive Systems?." AAAI 2017.

Smitha Milli, Anca D. Dragan. "Literal or Pedagogic Human? Analyzing Human Model Misspecification in Objective Learning." UAI 2019.

Smitha Milli, Ludwig Schmidt, Anca D. Dragan, Moritz Hardt. "Model Reconstruction from Model Explanations." FAT* 2019.

Smitha Milli, John Miller, Anca D. Dragan, Moritz Hardt. "The Social Cost of Strategic Classification." FAT* 2019.

Smitha Milli, Falk Lieder, Tom Griffiths. "A Rational Reinterpretation of Dual-Process Theories." UAI 2020.

Smitha Milli, Pieter Abbeel, Igor Mordatch. "Interpretable and Pedagogical Examples." (Preprint) 2020.

Sören Mindermann, Rohin Shah, Adam Gleave, Dylan Hadfield-Menell. "Active Inverse Reward Design." ICML 2018 GoalsRL workshop.

Karthika Mohan, Felix Thoemmes, Judea Pearl. "Estimation with Incomplete Data: The Linear Case." IJCAI 2018.

Karthika Mohan. "On Handling Self-masking and Other Hard Missing Data Problems." AAAI 2018.

Karthika Mohan, Judea Pearl. "Graphical Models for Processing Missing Data." JASA 2019.

Thomas J. H. Morgan, Jordan W. Suchow, Thomas L. Griffiths. "What the Baldwin Effect affects depends on the nature of plasticity." Cognition, 197 (2020).

Thanh H. Nguyen, Yongzhao Wang, Arunesh Sinha, Michael P. Wellman. "Deception in finitely repeated security games." AAAI 2019.

Junhyuk Oh, Yijie Guo, Satinder Singh, Honglak Lee. "Self-Imitation Learning." ICML 2018.

Xinlei Pan, Weiyao Wang, Xiaoshuai Zhang, Bo Li, Jinfeng Yi, Dawn Song. "How You Act Tells a Lot: Privacy-Leaking Attack on Deep Reinforcement Learning." AAMAS 2019.

Nicolas Papernot, Patrick McDaniel, Arunesh Sinha, Michael P. Wellman. "SoK: Security and Privacy in Machine Learning." IEEE European Symposium on Security and Privacy 2018.

Xue Bin Peng, Angjoo Kanazawa, Sam Toyer, Pieter Abbeel, Sergey Levine. "Variational Discriminator Bottleneck: Improving Imitation Learning, Inverse RL, and GANs by Constraining Information Flow." ICLR 2019.

Joshua Peterson, David Bourgin, Daniel Reichman, Thomas Griffiths, Stuart Russell. "Cognitive model priors for predicting human decisions." ICML 2019.

Ori Plonsky, Reut Apel, Eyal Ert, Moshe Tennenholtz, David Bourgin, Joshua C. Peterson, Daniel Reichman, Thomas L. Griffiths, Stuart J. Russell, Evan C. Carter, James F. Cavanagh, Ido Erev. "Predicting human decisions with behavioral theories and machine learning." (Preprint) 2019.

Matthew Rahtz, James Fang, Anca D. Dragan, Dylan Hadfield-Menell. "An Extensible Interactive Interface for Agent Design." ICML Human-in-the-Loop Learning Workshop 2019.

Ellis Ratner, Dylan Hadfield-Menell, Anca D. Dragan. "Simplifying Reward Design through Divide-and-Conquer." RSS 2018.

Siddharth Reddy, Anca D. Dragan, Sergey Levine. "Shared Autonomy via Deep Reinforcement Learning." RSS 2018.

Siddharth Reddy, Anca D. Dragan, Sergey Levine. "Where Do You Think You're Going?: Inferring Beliefs about Dynamics from Behavior." NeurIPS 2018.

Siddharth Reddy, Anca D. Dragan, Sergey Levine, Shane Legg, Jan Leike. "Learning Human Objectives by Evaluating Hypothetical Behavior." (Preprint) 2019.

Siddharth Reddy, Anca D. Dragan, Sergey Levine. "SQIL: Imitation Learning via Regularized Behavioral Cloning.." ICLR 2020.

Nan Rong, Joseph Y. Halpern, Ashutosh Saxena. "MDPs with Unawareness in Robotics." UAI 2016.

Stuart Russell. "The new weapons of mass destruction?." The Security Times 2018.

Stuart Russell. "Human Compatible: Artificial Intelligence and The Problem of Control." Penguin Random House 2019.

Stuart Russell. "Artificial Intelligence: A Modern Approach (Textbook, 4th Edition)." Pearson 2020.

Dorsa Sadigh, S. Shankar Sastry, Sanjit A. Seshia, Anca Dragan. "Information Gathering Actions Over Human Internal State." IROS 2016.

Dorsa Sadigh, Shankar Sastry, Sanjit Seshia, Anca Dragan. "Planning for Autonomous Cars that Leverage Effects on Human Actions." RSS 2016.

Dorsa Sadigh, Anca Dragan, S. Shankar Sastry, Sanjit Seshia. "Active Preference-Based Learning of Reward Functions." RSS 2017.

Dorsa Sadigh, Nick Landolfi, Shankar S. Sastry, Sanjit A. Seshia, Anca D. Dragan. "Planning for cars that coordinate with people: leveraging effects on human actions for planning and active information gathering over human internal state." Autonomous Robots 2018.

Sophia Sanborn, Michael Chang, Sergey Levine, Thomas Griffiths. "Sparse Skill Coding: Learning Behavioral Hierarchies with Sparse Codes." ICLR 2020 submission.

Rohin Shah, Noah Gundotra, Pieter Abbeel, Anca D. Dragan. "On the Feasibility of Learning, Rather than Assuming, Human Biases for Reward Inference." ICML 2019.

Rohin Shah, Dmitrii Krasheninnikov, Jordan Alexander, Pieter Abbeel, Anca Dragan. "Preferences Implicit in the State of the World." ICLR 2019.

Megan Shearer, Gabriel Rauterberg, Michael P. Wellman. "An Agent-Based Model of Financial Benchmark Manipulation." ICML 2019.

Amitai Shenhav, Sebastian Musslick, Falk Lieder, Wouter Kool, Thomas L Griffiths, Jonathan D Cohen, Matthew M Botvinick. "Toward a Rational and Mechanistic Account of Mental Effort." Annual Review of Neuroscience, 40, 9f4b26db33-124 (2017).

Arunesh Sinha, Michael P. Wellman. "Incentivizing Collaboration in a Competition." AAMAS 2019.

Sarath Sreedharan, Siddharth Srivastava, David Smith, Subbarao Kambhampati. "Why Can't You Do That, HAL? Explaining Unsolvability of Planning Tasks." IJCAI 2019.

Elis Stefansson, Jaime F. Fisac, Dorsa Sadigh, S. Shankar Sastry, Karl H. Johansson. "Human-robot interaction for truck platooning using hierarchical dynamic games." European Control Conference 2019.

Liting Sun, Wei Zhan, Masayoshi Tomizuka, Anca D. Dragan. "Courteous Autonomous Cars." IROS 2018.

Gokul Swamy, Siddharth Reddy, Sergey Levine, Anca D. Dragan. "Scaled Autonomy: Enabling Human Operators to Control Robot Fleets." ICRA 2020.

Prasad Tadepall, Cameron Barrie, Stuart J. Russell. "Learning Causal Trees with Latent Variables via Controlled Experimentation." AAAI 2019.

Alexander Todorov, Stefan Uddenberg, Joshua Peterson, Thomas Griffiths, Jordan Suchow. "Data-Driven, Photorealistic Social Face-Trait Encoding, Prediction, and Manipulation Using Deep Neural Networks." Patent application 2020.

Sam Toyer, Felipe Trevizan, Sylvie Thiebaux, Lexing Xie. "ASNets: Deep Learning for Generalised Planning." JAIR 2020.

Ruibo Tu, Cheng Zhang, Paul Ackermann, Karthika Mohan, Hedvig Kjellström, Kun Zhang. "Causal Discovery in the Presence of Missing Data." AISTATS 2019.

Aaron Tucker, Adam Gleave, Stuart Russell. "Inverse reinforcement learning for video games." NeurIPS 2018 Deep RL Workshop.

Alexander Matt Turner, Dylan Hadfield-Menell, Prasad Tadepalli. "Conservative agency via attainable utility preservation.." AIES 2020.

Vivek Veeriah, Junhyuk Oh, Satinder Singh. "Many-Goals Reinforcement Learning." (Preprint) 2018.

Xintong Wang, Chris Hoang, Michael P. Wellman. "Learning-Based Trading Strategies in the Face of Market Manipulation." ICML 2019 Workshop on AI in Finance.

Michael Wellman, Eric Sodomka, Amy Greenwald. "Self-confirming price-prediction strategies for simultaneous one-shot auctions." Games and Economic Behavior, 102, 339–372 (2012).

Andrew Whalen, Thomas L. Griffiths, Daphna Buchsbaum. "Sensitivity to Shared Information in Social Learning." Cognitive Science 2018.

Bryce Wiedenbeck, Fengjun Yang, Michael P. Wellman. "A Regression Approach for Modeling Games with Many Symmetric Players." AAAI 2018.

Mason Wright and Michael P. Wellman. "Evaluating the Stability of Non-Adaptive Trading in Continuous Double Auctions." AAMAS 2018.

Yi Wu, Siddharth Srivastava, Nicholas Hay, Simon Du, Stuart Russell. "Discrete-Continuous Mixtures in Probabilistic Programming: Generalized Semantics and Inference Algorithms." ICML 2018.

IEEE Transactions on Robotics. "Bayesian Relational Memory for Semantic Visual Navigation." ICCV 2019.

Kelvin Xu, Ellis Ratner, Anca Dragan, Sergey Levine, Chelsea Finn. "Learning a Prior over Intent via Meta-Inverse Reinforcement Learning." ICML 2019.

Albert Zhan, Stas Tiomkin, Pieter Abbeel. "Preventing Imitation Learning with Adversarial Policy Ensembles." ICLR 2020.

Shun Zhang, Edmund H. Durfee, Satinder P. Singh. "Minimax-regret querying on side effects for safe optimality in factored Markov decision processes." IJCAI 2018.

Jason Y. Zhang, Anca D. Dragan. "Learning from Extrapolated Corrections." ICRA 2019.

Zeyu Zheng, Junhyuk Oh, Satinder Singh. "On Learning Intrinsic Rewards for Policy Gradient Methods." NeurIPS 2018.

Allan Zhou, Dylan Hadfield-Menell, Anusha Nagabaudi, Anca Dragan. "Expressive Robot Motion Timing." HRI 2017.

Allan Zhou, Anca D. Dragan. "Cost Functions for Robot Motion Style." IROS 2018.