Center for
Human-Compatible
Artificial
Intelligence

CHAI Progress Report | 2022

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

April 2022

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.

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* — 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].

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

[2] All numeric references point to items in the "Specific Outputs" section of the 2020 CHAI progress report.

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.

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.

CHAI Progress Report *July 2020 - December 2021*

This report describes the growth, outputs, and new program engagements of the Center for Human-Compatible AI in the 18 months since the 2020 CHAI Progress Report.

The Center's objective is to develop the conceptual and technical wherewithal to reorient the general thrust of AI research towards provably beneficial systems. Embedded within the Berkeley AI Research Lab at the University of California Berkeley, CHAI is focused on technical AI safety. It is supported by gifts from independent donors who share interest in achieving its objective.

In the 18 months from mid-2020 to the end of 2021, CHAI produced significant outputs. CHAI researchers published 54 papers, 13 podcast episodes, and 5 newsletters. 6 CHAI students completed their PhD's and accepted positions at DeepMind, MIT, the Digital Life Initiative at Cornell Tech, Stanford HAI / CISAC, Oregon State University and Cruise. CHAI's founder and faculty director Stuart Russell was named an Officer of the Most Excellent Order of the British Empire (OBE); he gave 29 invited talks, notably the 2021 Reith Lectures, broadcast on the BBC over four hours, and the 2020 annual Turing Lecture at the Alan Turing Institute. New translations of Russell's trade book *Human Compatible* were published in Chinese, Korean, Japanese, Turkish, and Ukrainian. CHAI co-PI Pieter Abbeel received the 2021 ACM Prize in Computing, the most prestigious mid-career award in computer science.

As an organization CHAI continued to grow, adding 1 postdoc, 8 new affiliate faculty, and 8 new graduate students.

During this time, public awareness of AI safety and impacts has increased as major governments introduced regulations and increased policy efforts to protect citizens, states and nations from unintended harms and malicious uses of AI, as well as to direct resources towards beneficial applications of AI. CHAI has advised extensively on draft regulations in the EU and US.

CHAI Talks/ Honors

Professor Anca Dragan was featured on the popular podcast Artificial Intelligence with Lex Fridman. They discussed human-robot interaction during semi-autonomous driving and reward engineering. Professor Dragan also spoke at the WIRED25 summit, explaining some of the challenges robots face when interacting with people.

Stuart Russell - talks/ honors

On May 26 2020, Professor Stuart Russell gave the annual Turing Lecture, sponsored by the Alan Turing Institute in London, on the topic of "Provably Beneficial Artificial Intelligence". Over 700 people attended, making it the most highly attended Turing Lecture so far. The lecture itself is here. Prof. Russell also provided written answers to the many questions from audience members here. His talk was also written about here.

On June 16, 2020, Professor Russell spoke at the World Peace Forum, held in Beijing, on the subject of lethal autonomous weapons.

On August 30, 2020, Professor Russell gave the keynote lecture for the annual European Conference on AI, held at Santiago de Compostela, Spain, on the subject of "How Not to Destroy the World With AI."

On September 10, 2020, his book *Human Compatible* was the subject of a meet-the-author session at the American Political Science Association's annual conference in San Francisco.

On September 22, 2020, he gave the keynote lecture for the annual United Nations Global Summit on AI for Good, held in Geneva. His talk was titled "On Beneficial AI" and asked how we are to define "Good" in the phrase "AI for Good".

On December 17, 2020, CHAI co-hosted the third Positive AI Economic Futures Workshop in collaboration with the World Economic Forum. More than 100 participants met virtually to plan scenarios of a future economy and society transformed by AI. The workshop brought together AI experts, science fiction authors, economists, policymakers, and social scientists.

Stuart Russell debated Melanie Mitchell, Professor at the Santa Fe Institute, in the latest episode of The Munk Debates, a debate series on major policy issues. In the episode, titled The Rise of Thinking Machines, Prof. Russell argued for the question, “Be it resolved, the quest for true AI is one of the great existential risks of our time.” Listen to the debate here or on any podcast service.

In the last quarter of 2020, Prof. Russell presented a keynote at Forum Humanum, a panel at the Paris Peace Forum, a keynote and roundtable at the Governance Of and By Digital Technology conference, a keynote at The Hague Conference on Responsible AI for Peace, Justice and Security, the annual lecture at the UC San Diego Institute for Practical Ethics, a keynote at the Fujitsu Labs Annual Technical Symposium, and a lecture at the Oxford Review of Economic Policy editorial seminar.

Stuart Russell accepted invitations to join the Scientific Advisory Board of the Sorbonne Center for AI in Paris and the Academic Committee of the Tsinghua University AI International Governance Institute.

He presented a plenary in the WEF 2021 Global Technology Governance Summit, a lecture at Applied Physics Lab Colloquium at Johns Hopkins University, a lecture at the MIT China Innovation and Entrepreneurship Forum, a panel at the Oxford China Forum, a keynote at the Center for the Future Mind at Florida Atlantic University, a lecture at the Microsoft Research New England Colloquium, the 18th Kim Okgill Memorial Lecture at Ewha Women's University, and a keynote at the inaugural Conference of the Stanford Existential Risk Institute.

Stuart Russell participated in Debates on the Future of Robotics Research, a virtual workshop at the International Conference on Robotics and Animation (ICRA) 2021. CHAI alum Jaime Fernández Fisac served as General Co-Chair on the organizing team. Watch recordings of the debates at roboticsdebates.org.

Stuart Russell gave a featured interview in the BBC Panorama documentary, "Are you scared yet, human?" He gave the keynote "How Not to Destroy the World With AI" to the IJCAI Workshop on Adverse Impacts of Artificial Intelligence and the French National AI conference Plate-Forme Intelligence Artificielle, Bordeaux. He also presented it as an invited lecture to the Ecole des Hautes Etudes Commerciales de Paris (HEC). He gave the keynote "Provably Beneficial AI and the Problem of Control," at the National Academies Conference on "How Artificial Intelligence and Machine Learning Transform the Human Condition." He also presented it as the inaugural lecture in the Prato Dialog series. He gave the keynote "Poorly Designed AI" to the EU Humane AI Network and the keynote "AI for the Comprehensive Nuclear-Test-Ban Treaty" to the CTBTO Science and Technology Summit, Vienna.

Stuart Russell delivered the 2021 Reith Lectures, perhaps the most widely broadcast series of talks on CHAI's subject matter.. The Reith Lecture series began in 1948 and is considered the most prestigious lecture series in the English-speaking world. Previous lecturers include Bertrand Russell, Arnold Toynbee, J. Robert Oppenheimer, George Kennan, J. K. Galbraith, Edward Said, Wole Soyinka, and Stephen Hawking. Russell is the first computer scientist selected for this honor. The lectures took place in London, Manchester, Edinburgh, and Newcastle in early November and were broadcast weekly in December on BBC Radio 4 and the BBC World Service, reaching an audience in the tens of millions.

Image credit: BBC (2021)

On November 9, Stuart Russell visited Windsor Castle and received the OBE from Prince William, Duke of Cambridge.

Student/ recent grad jobs, talks, honors

Tom Gilbert graduated in 2021 with his PhD in Machine Ethics and Epistemology. He is now working at the Digital Life Initiative at Cornell Tech as a postdoc.

Vael Gates graduated in 2021 with their PhD in Neuroscience with a focus on computational cognitive science. They are working at Stanford HAI / CISAC as a postdoc.

Karthika Mohan has finished her postdoc with CHAI. She joined Oregon State University as an assistant professor.

Micah Carroll graduated from U.C. Berkeley with a B.A. in statistics in 2019 and has been continuing his work at CHAI as a PhD student.

The *MIT Technology Review* highlighted Adam Gleave in a feature on his paper "Adversarial Policies: Attacking Deep Reinforcement Learning," coauthored with fellow CHAI members Michael Dennis, Cody Wild, Neel Kant, Sergey Levine, and Stuart Russell. Their research was accepted to be presented at the International Conference on Learning Representations (ICLR).

Rohin Shah spoke on the Machine Ethics podcast about alignment problems in AI, constraining AI behavior, current AI vs future AI, recommendation algorithms and extremism, and other topics.

In July 2021, CHAI alumni, Jaime Fernández Fisac, began working as an Assistant Professor of Electrical Engineering at Princeton University. Jaime's work combines safety analysis, machine learning techniques, and insights from cognitive science to enable robots to strategically plan their interaction with humans.

Thanard Kurutach graduated with the thesis Learning, Planning, and Acting with Models and is working at Cruise as an AI/ML Research Scientist.

Rohin Shah and Dylan Hadfield-Menell accepted jobs after the completion of their PhD studies. Rohin is working at DeepMind as a Research Scientist and Dylan is working as an Assistant Professor at MIT in July 2021.

Thomas Krendl Gilbert, PhD candidate in Machine Ethics and Epistemology, has been awarded the Simons Institute Law and Society Fellowship for its upcoming research program on Theory of Reinforcement Learning.

Alyssa Li Dayan, a PhD student advised by Stuart Russell, was awarded a fellowship from The Hertz Foundation.

In December 2020 Daniel Filan launched the AI X-risk Research Podcast (AXRP). In each episode, Daniel interviews the author of a paper and discusses how it might reduce the existential risk of AI. In the first three episodes, Daniel features CHAI researchers Andrew Critch, Rohin Shah, and Adam Gleave.

The Science article Who needs a teacher? Artificial intelligence designs lesson plans for itself features three papers co-authored by members of CHAI. The papers, which explore the learning paradigm known as autocurricula, are *Emergent Complexity and Zero-shot Transfer via Unsupervised Environment Design* (Michael Dennis, Stuart Russell, Andrew Critch), *A Novel Automated Curriculum Strategy to*

Solve Hard Sokoban Planning Instances (Bart Selman), and *Automatic Curriculum Learning through Value Disagreement* (Pieter Abbeel).

Rachel Freedman spoke to the Berkeley High School STEMinist Club on the topic of human-compatible AI research.

The TalkRL podcast featured Michael Dennis and Tom Gilbert.

Daniel Filan's podcast AXRP released three new episodes featuring conversations with Beth Barnes, Vanessa Kosoy, and Evan Hubinger.

Research Engineer Steven Wang is starting a Master's of Computer Science at ETH Zurich. Research Engineer Cody Wild is joining Google Research. Operations Assistant Noor Brody is now a Data Engineer at SimpleLab Inc.

Affiliate talks/ honors

Caroline Jeanmarie, Director of Strategic Research and Partnerships at CHAI, gave a presentation at the 2020 Foresight AGI Strategy Meeting on the U.S. Guidance for Regulation of AI Applications.

Caroline Jeanmaire, then Director of Strategic Research and Partnerships at CHAI, was recognized as one of 100 Brilliant Women in AI Ethics in 2021. The list is published annually by Women in AI Ethics, which has the mission to make AI more diverse and accessible, in part by recognizing rising stars in the field.

On October 21 2020, CHAI celebrated the launch of The Alignment Problem by UC Berkeley Visiting Scholar and CHAI Affiliate Brian Christian. Journalist Nora Young hosted an interview with Christian and an audience Q&A. The Alignment Problem tells the story of the ethics and safety movement in AI and reports directly from those working on the field's frontier, including current CHAI researchers. Read more and watch a recording of the interview and Q&A here.

The LA Times Book Prizes named *The Alignment Problem* by Brian Christian a finalist for best Science & Technology book of 2020. It was also named by Microsoft CEO Satya Nadella as one of the five books that inspired him in 2021. "Clear and compelling," he writes, the book "moves us from the theoretical to the practical while attempting to answer one of our industry's most pressing questions: How do we teach machines, and what should we teach them?"

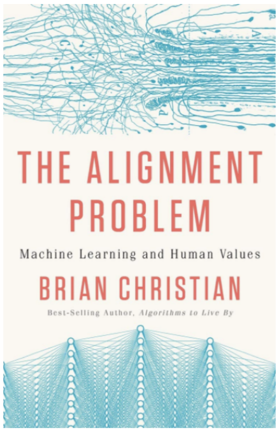
THE ALIGNMENT PROBLEM
Machine Learning and Human Values
BRIAN CHRISTIAN
Best-Selling Author, Algorithms to Live By

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

Brian also gave a keynote address at the annual Linux Foundation Member Summit about the alignment problem and the role the open-source community can play in addressing issues of AI safety. And Brian is featured in the New York Times podcast The Ezra Klein Show, in an episode titled "If 'All Models Are Wrong,' Why Do We Give Them So Much Power?"

Pieter Abbeel launched a new podcast, The Robot Brains, where he hosts leading experts in AI robotics.

The Sunday Show podcast interviewed Jonathan Stray in the episode "Can social media help depolarize society?" They discuss his latest paper, Designing Recommender Systems to Depolarize, which appeared in First Monday.

Hires, new affiliates, CHAI workshops and other demographics and funding related info.

Jonathan Stray joined CHAI as a visiting scholar in April 2021. He was hired as a full-time Senior Scientist in August 2022 He holds an MSc in Computer Science from the University of Toronto and an MA in Journalism from the University of Hong Kong. At CHAI he works on the design of recommender systems for better personalized news and information. Jonathan joined CHAI full-time as Specialist, Recommender Systems in April 2022.

Justin Svegliato joined CHAI as a postdoc mentored by Stuart Russell. He received his PhD in Computer Science at UMass Amherst. The goal of his research is to build autonomous systems that operate in the open world for long periods of time in an efficient, reliable, and ethical way.

CHAI added eight new affiliate members: Jakob Foerster (Assistant Professor of Engineering Science, Oxford), Moritz Hardt (Associate Professor of Computer Science, Berkeley, and Director, Social Foundations of Computation, Max Planck Institute for Intelligent Systems, Tübingen), Rediet Abebe (Assistant Professor of Computer Science, Berkeley), Niko Kolodny (Professor and Chair of Philosophy, Berkeley), Nika Haghtalab (Assistant Professor of Computer Science, Berkeley), Brian Christian (author), Vincent Corruble (Associate Professor of Computer Science, Sorbonne Université), and Tom Lenaerts (Professor of Machine Learning, Université Libre de Bruxelles).

CHAI was featured in three recent episodes of the Future of Life Podcast. Professor Stuart Russell joined Harvard Psychology Professor Steven Pinker to discuss the foundations, benefits, and existential threat of AI. PhD student Rohin Shah and MIRI researcher Buck Shlegeris shared their thoughts on the current state of research efforts for beneficial AI. Research scientist Andrew Critch discussed the paper "AI Research Considerations for Human Existential Safety," coauthored with David Krueger.

On December 17, 2020, CHAI hosted the third Positive AI Economic Futures Workshop in collaboration with the World Economic Forum. More than 100 participants met virtually to plan scenarios of a future economy and society transformed by AI. The workshop brought together AI experts, science fiction authors, economists, policymakers, and social scientists.

Professor Tom Lenaerts visited CHAI from the Computer Science Department of the Université Libre de Bruxelles. Prof. Lenaerts and Michael Dennis investigated the relationship between translucent players and commitment devices, considering an evolutionary game theory perspective. Tom Lenaerts has become a new affiliate since his time with CHAI.

Noyuri Mima is a visiting scholar who has been with CHAI since August 2021. is professor of Computer Science at the Future University Hakodate (FUN) in Hokkaido, Japan as a key member of the founding team. Prof. Mima works in the intersection of CS, cognitive psychology, and education.

Anni Hellman is a visiting scholar who has been with CHAI since August 2021. She is visiting CHAI as part of her sabbatical from the European Commission.

New Collaborators Joar Skalse (Oxford) and Oliver Richardson (Cornell) will work alongside Adam Gleave. Antoni Lorente (King's College London, Universitat Autònoma de Barcelona) will work through the fall with Thomas Krendl Gilbert.

On June 7-8, 2021, CHAI held its fifth annual workshop. Over 150 attendees, including professors, students, industry researchers, and policymakers met virtually to discuss AI safety and beneficial systems.

- Participants came from over 100 institutions and attended 26 talks, six dialogues, one panel, and plenaries given by Stuart Russell, Peter Eckersley, Glen Weyl, Adam Kalai, and Anca Dragan. Attendees participated in organized meet-ups and conversation in the virtual space Gather.town. View the workshop program here.

Two Kavli Centers for Ethics, Science, and the Public – at the University of California, Berkeley, and the University of Cambridge – are launching to engage the public in identifying and exploring ethical considerations and impacts born from scientific discovery. Stuart Russell will be the inaugural Director of the new center at Berkeley, and AI will be one of the three initial foci along with neuroscience and gene editing.

Publications

CHAI co-published a report with the World Economic Forum. The "Positive AI Economic Futures" report was written by CHAI's Stuart Russell and Caroline Jeanmaire along with Daniel Susskind of Oxford University. It attempts to articulate positive and realistic visions of a future where Artificial Intelligence can do most of what we currently call work. You can read a summary of the report here.

Prof. Stuart Russell

Professor Stuart Russell's book *Human Compatible* was published in Chinese, as well as Turkish, Korean, Japanese, and Ukranian:

- Stuart Russell, İnsanlık için Yapay Zekâ: Yapay Zekâ ve Kontrol Problemi. Translated to Turkish by Barış Satılmış. Ankara: Buzdağı Yayınevi, 2021.
- Stuart Russell, 어떻게 인간과 공존하는 인공지능을 만들 것인가. Translated to Korean by Lee Han-eum. Seoul: Gimm-Young, 2021.
- Stuart Russell, AI新生. Translated to Japanese by Nobuhiko Matsui. Tokyo: Misuzu Shobo, 2021.
- Stuart Russell, Сумісний з людиною. Штучний інтелект і проблема контролю, Translated to Ukrainian by Viktoria Zenhyea. Kiev: Bookchef, 2020

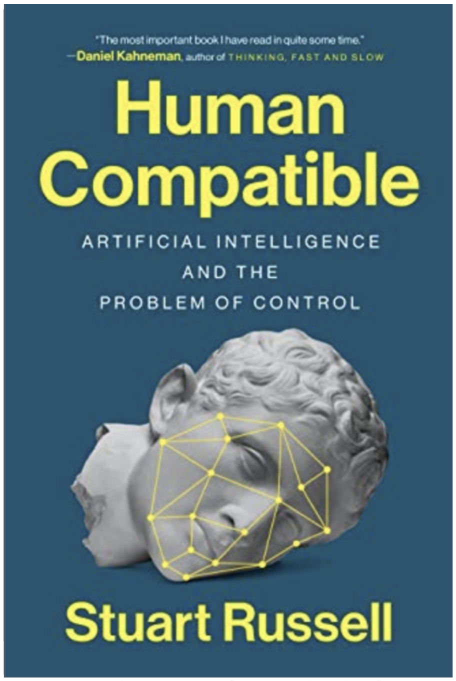
"The most important book I have read in quite some time."
—Daniel Kahneman, author of THINKING, FAST AND SLOW
Human Compatible
ARTIFICIAL INTELLIGENCE AND THE PROBLEM OF CONTROL
Stuart Russell

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

On the topic of lethal autonomous weapons:

- Stuart Russell, It's time to ban autonomous killer robots before they become a threat, Financial Times. August 5, 2021.
- Stuart Russell, Anthony Aguirre, Emilia Javorsky and Max Tegmark, Lethal Autonomous Weapons Exist; They Must Be Banned. IEEE Spectrum, June 16, 2021.

Two chapters in 'Reflections on Artificial Intelligence for Humanity', based on the Global Forum on AI for Humanity held in Paris in 2019:

- Raja Chatila, Virginia Dignum, Michael Fisher, Fosca Gianotti, Katharina Morik, Stuart Russell, David Sadek, and Karen Yeung, "Trustworthy AI." In Bertrand Braunschweig and Malik Ghallab (eds.), Reflections on Artificial Intelligence for Humanity, Springer, 2021.

Jocelyn Maclure and Stuart Russell, "AI for Humanity: The Global Challenges." In Bertrand Braunschweig and Malik Ghallab (eds.), Reflections on Artificial Intelligence for Humanity, Springer, 2021.

Recent Graduates

Rohin Shah published his PhD dissertation, "Extracting and Using Preference Information from the State of the World." He is now working as a Research Scientist on the technical AGI safety team at DeepMind.

Affiliates

Dylan Hadfield-Menell, CHAI alumnus and Assistant Professor at MIT, published his thesis The Principal-Agent Alignment Problem in Artificial Intelligence in August 2021.

IEEE ISTAS20

- Thomas Krendl Gilbert co-authored "AI Development for the Public Interest: From Abstraction Traps to Sociotechnical Risks,"

IJCAI 2020 AI Safety Workshop

- "Choice Set Misspecification in Reward Inference" by Rachel Freedman, Rohin Shah, and Anca Dragan received the Best Paper Award at the IJCAI 2020 AISafety workshop.

Scott Emmons, Andrew Critch, and Stuart Russell published Symmetry, Equilibria, and Robustness in Common-Payoff Games in the Games and Incentives Workshop 2021.

Daniel Filan, Stephen Casper, Shlomi Hod, Cody Wild, Andrew Critch, and Stuart Russell posted Clusterability in Neural Networks on arXiv.

Paria Rashidinejad, Banghua Zhu, Cong Ma, Jiantao Jiao, and Stuart Russell posted Bridging Offline Reinforcement Learning and Imitation Learning: A Tale of Pessimism on arXiv.

Paul Knott, Micah Carroll, Anca Dragan, and Rohin Shah co-authored the paper Evaluating the Robustness of Collaborative Agents alongside Sam Devlin, Kamil Ciosek, and Katja Hofmann.

Thomas Krendl Gilbert published the white paper Mapping the Political Economy of Reinforcement Learning Systems: The Case of Autonomous Vehicles on the Simons Institute website.

Brian Christian published the article Watch and Learn: Offline Reinforcement Learning on the Simons Institute website.

NeurIPS 2020 Papers

NeurIPS 2020 acceptances co-authored by CHAI researchers include these main conference papers (asterisk denotes equal contribution):

- Scott Emmons*, Ajay Jain*, Michael Laskin*, Thanard Kurutach, Pieter Abbeel, Deepak Pathak. "Sparse Graphical Memory for Robust Planning"

- Sam Toyer, Rohin Shah, Andrew Critch, Stuart Russell. "The MAGICAL Benchmark for Robust Imitation"
- Michael Dennis, Natasha Jaques, Eugene Vinitsky, Alexandre Bayen, Stuart Russell, Andrew Critch, Sergey Levine. "Emergent Complexity and Zero-shot Transfer via Unsupervised Environment Design" (Oral presentation, Reinforcement Learning)
- Paria Rashidinejad, Jiantao Jiao, Stuart Russell. "SLIP: Learning to Predict in Unknown Dynamical Systems with Long-Term Memory" (Oral presentation, Dynamical Sys/Density/Sparsity)
- Alexander Matt Turner, Neale Ratzlaff, Prasad Tadepalli. "Avoiding Side Effects in Complex Environments" (Spotlight presentation, Reinforcement Learning)
- Yuqing Du, Stas Tiomkin, Emre Kiciman, Daniel Polani, Pieter Abbeel, Anca Dragan. "AvE: Assistance via Empowerment"
- Hong Jun Jeon, Smitha Milli, Anca D. Dragan. "Reward-Rational (Implicit) Choice: A Unifying Formalism for Reward Learning"
- Kush Bhatia, Ashwin Pananjady, Peter Bartlett, Anca Dragan, and Martin Wainwright. "Preference Learning Along Multiple Criteria: A Game-theoretic Perspective"

In addition, the following papers appeared in workshops:

- Pedro Freire, Adam Gleave, Sam Toyer, Stuart Russell. "DERAIL: Diagnostic Environments for Reward and Imitation Learning" (Deep RL Workshop)
- Rohin Shah, Pedro Freire, Neel Alex, Rachel Freedman, Dmitrii Krasheninnikov, Lawrence Chan, Michael Dennis, Pieter Abbeel, Anca Dragan, Stuart Russell. "Benefits of Assistance over Reward Learning" (Cooperative AI Workshop; won a Best Paper Award)
- Eric J. Michaud, Adam Gleave, Stuart Russell. "Understanding Learned Reward Functions" (Deep RL Workshop)

AAAI 2021:

- Probabilistic Dependency Graphs by Oliver Richardson and Joseph Y. Halpern at Cornell University
- Asking the Right Questions: Learning Interpretable Action Models Through Query Answering by Pulkit Verma, Shashank Rao Marpally, and Siddharth Srivastava at Arizona State University
- Unifying Principles and Metrics for Safe and Assistive AI by Siddharth Srivastava

ICLR 2021 Papers

- Adam Gleave, Michael D. Dennis, Shane Legg, Stuart Russell, Jan Leike. "Quantifying Differences in Reward Functions" (Spotlight paper)
- Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, Jacob Steinhardt. "Measuring Massive Multitask Language Understanding"
- Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, Jacob Steinhardt. "Aligning AI With Shared Human Values"

- David Lindner, Rohin Shah, Pieter Abbeel, Anca Dragan. "Learning What To Do by Simulating the Past"
- Cassidy Laidlaw, Sahil Singla, Soheil Feizi, Perceptual Adversarial Robustness: Defense Against Unseen Threat Models. ICLR 2021

ICML 2021

- Kimin Lee, Laura Smith, and Pieter Abbeel, PEBBLE: Feedback-Efficient Interactive Reinforcement Learning via Relabeling Experience and Unsupervised Pre-training. ICML 2021
- Johannes Treutlein, Michael Dennis, Caspar Oesterheld, and Jakob Foerster, A New Formalism, Method and Open Issues for Zero-Shot Coordination. ICML 2021
- Smitha Milli, Luca Belli, Moritz Hardt, Causal Inference Struggles with Agency on Online Platforms. ICML 2021
- Kimin Lee, Michael Laskin, Aravind Srinivas, Pieter Abbeel, SUNRISE: A Simple Unified Framework for Ensemble Learning in Deep Reinforcement Learning. ICML 2021 Spotlight paper

NeurIPS 2021

- Kimin Lee, Laura Smith, Anca Dragan, Pieter Abbeel, B-Pref: Benchmarking Preference-Based Reinforcement Learning.
- Michael Dennis, Minqi Jiang, Jack Parker-Holder, Jakob Foerster, Edward Grefenstette, Tim Rocktäschel Replay-Guided Adversarial Environment Design
- Arnaud Fickinger, Stuart Russell, Brandon Amos, Noam Brown Scalable Online Planning via Reinforcement Learning Fine-Tuning
- Cassidy Laidlaw, Stuart Russell Uncertain Decisions Facilitate Better Preference Learning
- Xin Chen, Sam Toyer, Cody Wild, Scott Emmons, Ian Fischer, Kuang-Huei Lee, Neel Alex, Steven H Wang, Ping Luo, Stuart Russell, Pieter Abbeel, Rohin Shah An Empirical Investigation of Representation Learning for Imitation
- Alexander Matt Turner, Logan Smith, Rohin Shah, Andrew Critch, Prasad Tadepalli Optimal Policies Tend To Seek Power, NeurIPS 2021 spotlight
- Tianjun Zhang*, Paria Rashidinejad*, Jiantao Jiao, Yuandong Tian, Joseph E. Gonzalez, Stuart Russell, MADE: Exploration via Maximizing Deviation from Explored Regions
- Paria Rashidinejad, Banghua Zhu, Cong Ma, Jiantao Jiao, Stuart Russell, Bridging Offline Reinforcement Learning and Imitation Learning: A Tale of Pessimism

Roel Dobbe, Thomas Krendl Gilbert, Yonatan Mintz, Hard Choices in Artificial Intelligence. Artificial Intelligence Volume 300, November 2021

FAccT 2021

- Smitha Milli, Luca Belli, Moritz Hardt, From Optimizing Engagement to Measuring Value.

AAMAS-21

- Charlotte Roman, Michael Dennis, Andrew Critch, and Stuart Russell, Accumulating Risk Capital Through Investing in Cooperation.

CVPR 2021

- Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, Dawn Song, Natural Adversarial Examples.

ICCV 2021

- Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, Dawn Song, Jacob Steinhardt, Justin Gilmer, The Many Faces of Robustness: A Critical Analysis of Out-of-Distribution Generalization.
- Avik Jain, Lawrence Chan, Daniel S. Brown, Anca D. Dragan, Optimal Cost Design for Model Predictive Control. Proceedings of the 3rd Conference on Learning for Dynamics and Control, PMLR 144:1205-1217, 2021

AABI 2021

- George Matheos, Alexander K. Lew, Matin Ghavamizadeh, Stuart Russell, Marco Cusumano-Towner, Vikash K. Mansinghka, Transforming Worlds: Automated Involutive MCMC for Open-Universe Probabilistic Models.
- Stuart Russell, Human-Compatible Artificial Intelligence. In Stephen Muggleton and Nick Chater (eds.), Human-Like Machine Intelligence, Oxford University Press, 2021