Center for Human-Compatible Artificial Intelligence

PROGRESS REPORT | 2023

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

May 31, 2023

Contact: chai-admin@berkeley.edu

CHAI Progress Report *May 2022 - May 2023*

This report describes the growth, outputs, and new program engagements of the Center for Human-Compatible AI in the year since the April 2022 CHAI progress report.

The Center for Human-Compatible Artificial Intelligence's objective is to develop the conceptual and technical wherewithal to reorient the general thrust of artificial intelligence (AI) research towards provably beneficial systems. Embedded within the Berkeley AI Research Lab at the University of California Berkeley, CHAI is a vibrant interdisciplinary research hub focused on technical AI safety. It is supported by gifts from independent donors who share an interest in achieving its objective.

This has been an extraordinary year for broad public awareness of both the potential and the perils of AI. The release of ChatGPT in November 2022 set off a competitive sprint among Big Tech, the venture and tech startup community, and the AI research world to make ever more capable large language models (LLM's). As with the various other deep learning models that put AI in the limelight over the last decade, CHAI has been engaged in the debates regarding their safe and beneficial development and deployment.

CHAI continues to produce essential research in collaboration with a broad range of collaborators and partner labs. Our publications this year include work in

- **Assistance games.** How AI systems can best learn from human instruction, especially imperfect human instructions.
- **Recommender Systems.** Recommenders are some of the largest deployed AI systems, and are already causing preventable harms.
- **Adversarial Robustness.** How can we break existing deep learning systems, and how can we build systems on well-founded principles that are harder to break?
- **Multi-Agent Cooperation.** How can agents (human or AI) work together to solve problems?
- **Social Impacts.** How the use of AI might transform society, and how we can direct it along beneficial paths.

A major thread running through our work is the attempt to create **provably safe and beneficial AI.** That is, it's not enough that an AI system seems to work – it must be possible to prove that it will work across a wide range of both expected and unexpected situations. This is ever more important in the era of large language models, which are not *designed* in the traditional sense – a trained deep neural network performs its task according to unknown principles. In practice, every kind of deep learning system, including LLMs, shows severe vulnerabilities and unpredictable failure modes.

This was vividly illustrated by CHAI research this year showing a simple way to beat superhuman Go playing AIs. Such systems are now built using deep learning, meaning that we

can't explain how they work, and in fact they do not really "understand" Go in the traditional sense. CHAI researchers were able to beat it using a silly strategy that no human would fall for.

Given the staggering size of the models (most estimates put GPT4 at over 600 billion parameters and trillions of training tokens), and their increased capability in numerous diverse tasks,it seems natural that governments, business ecosystems, educational institutions and (not least) media have focused so much attention on the future implications of these more concrete examples of apparently intelligent agents, accessible to millions.

The release of ChatGPT was a turning point, bringing the promise and perils of increasingly powerful multi-purpose AI beyond hypothetical debate and into mainstream discussion. It is an important opportunity for CHAI to increase its attention on emerging regulation, tracking efforts globally and providing guidance and direction where CHAI's particular focus on scientific and technical routes to safe and beneficial AI are most needed.

In the 12 months from May 2022 through April 2023, CHAI produced significant outputs. CHAI researchers published 26 papers. Four CHAI students completed their PhD's and accepted or began positions at DeepMind, Berkeley AI Research, Carnegie Mellon, and a new, AI-safety focused research company called FAR.ai. CHAI's founder and faculty director Stuart Russell gave hundreds of invited talks and media appearances, and engaged with numerous organizations creating real policy from erstwhile principles, including the EU and its constituent governments, members of US congress including senate majority leader Charles Schumer, and the OECD. CHAI Senior Scientist Jonathan Stray co-authored a Supreme Court brief on social media algorithms that was cited in oral arguments before the Court ultimately agreed that making platforms liable for the speech of their users was not the correct way to regulate increasingly powerful AI in media applications.
(Anything going on with the Co-PI's?)

As an organization CHAI continued to grow, adding **_** postdocs, **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.

NSF PSBAI Workshop

In October, CHAI held its first government-funded workshop, the NSF Convergence Accelerator Workshop on Provably Safe and Beneficial AI (PSBAI), with the aim of producing a research

agenda towards the design, creation, and practical deployment of verifiable, well-founded AI systems.

The inaugural PSBAI workshop was held in Berkeley, CA over the weekend of October 7-9th, 2022, where 51 attendees from a range of disciplines gathered for focussed work. The attendees, many of them leading experts in their fields, spanned a range of technical disciplines including artificial intelligence, programming languages, formal methods, control theory, game theory, statistics, and safety engineering, but also included key interdisciplinary researchers in philosophy, health care, defense, media, and law. There were several senior industry researchers in attendance, as well as government representatives. Attendees came from the UK, Canada, and Australia, as well as the US.

A sister workshop on the Ethical Design of AIs (EDAI) was held virtually with a plenary session on September 22 and three subsequent working group sessions. These two workshops covered complementary aspects of the challenge of ensuring that AI is safe and beneficial when integrated into individual lives and social systems. EDAI included sophisticated consideration of ethical principles, human-centered design, and AI governance, as well as the articulation of many domain-specific challenges. PSBAI attacked the corresponding challenge of making it possible to build systems that can implement the outputs of these design processes.

Publications

Prof. Stuart Russell

Journal Articles

- Joar Skalse, Matthew Farrugia-Roberts, Stuart Russell, Alessandro Abate, and Adam Gleave, Invariance in policy optimisation and partial identifiability in reward learning. In *Proc. ICML-23*, 2023.
- Tony Tong Wang, Adam Gleave, Tom Tseng, Nora Belrose, Kellin Pelrine, Joseph Miller, Michael D Dennis, Yawen Duan, Viktor Pogrebniak, Sergey Levine, and Stuart Russell, Adversarial policies beat superhuman Go AIs. In *Proc. ICML-23*, 2023.
- Niklas Lauffer, Ameesh Shah, Micah Carroll, Michael D Dennis, Stuart Russell, Who needs to know? Minimal knowledge for optimal coordination. In *Proc. ICML-23*, 2023.
- Tony T. Wang, Adam Gleave, Tom Tseng, Nora Belrose, Kellin Pelrine, Joseph Miller, Michael D Dennis, Yawen Duan, Viktor Pogrebniak, Sergey Levine, Stuart Russell, Adversarial policies beat superhuman Go AIs. arXiv:2211.00241, February 2023.
- Paria Rashidinejad, Hanlin Zhu, Kunhe Yang, Stuart Russell, and Jiantao Jiao, Optimal conservative offline RL with general function approximation via augmented Lagrangian. In *Proc. ICLR-23*, 2023.
- Alexander Lew, George Matheos, Matin Ghavamizadeh, Nishad Gothoskar, Stuart Russell, and Vikash Mansinghka, SMCP3: SMC with Probabilistic Program Proposals. In

Proc. Twenty-Sixth International Conference on Artificial Intelligence and Statistics, Valencia, Spain, 2023.

- Stuart Russell, AI weapons: Russia's war in Ukraine shows why the world must enact a ban. *Nature*, **614**, 620-623, 2023. (21 February, 2023)
- Adam Gleave, Mohammad Taufeeque, Juan Rocamonde, Erik Jenner, Steven H. Wang, Sam Toyer, Maximilian Ernestus, Nora Belrose, Scott Emmons, and Stuart Russell, imitation: Clean Imitation Learning Implementations. arxiv.org:2211.11972, November 2022.
- Samer B. Nashed, Justin Svegliato, Abhinav Bhatia, Stuart Russell, and Shlomo Zilberstein, Selecting the Partial State Abstractions of MDPs: A Metareasoning Approach with Deep Reinforcement Learning. In *Proc. IROS-22*
- Paria Rashidinejad, Banghua Zhu, Cong Ma, Jiantao Jiao, and Stuart Russell, Bridging Offline Reinforcement Learning and Imitation Learning: A Tale of Pessimism. *IEEE Transactions on Information Theory*, DOI: 10.1109/TIT.2022.3185139, 2022.
- Scott Emmons, Caspar Oesterheld, Andrew Critch, Vincent Conitzer, and Stuart Russell, For Learning in Symmetric Teams, Local Optima are Global Nash Equilibria. In *Proc. ICML-22*.
- Micah Carroll, Dylan Hadfield-Menell, Stuart Russell, and Anca Dragan, Estimating and Penalizing Induced Preference Shifts in Recommender Systems. In *Proc. ICML-22*.
- Kenji Doya, Arisa Ema, Hiroaki Kitano, Masamichi Sakagami, and Stuart Russell, Social impact and governance of AI and neurotechnologies. *Neural Networks*, 152, 542-554, 2022.

Professor Russell has been interviewed by a wide-array of global news outlets over the last year. A selection appears below:

Media

- If we succeed. *Daedalus*, Spring 2022, 43-57.
- The best of Radio Davos over the last year, *World Economic Forum*, August 4, 2022. Features this interview.
- The Foundations of Artificial Intelligence, interview by Daniel bashir, *The Gradient* podcast, October 6, 2022.
- Politicians must prepare for AI or face the consequences. *The House* (magazine of the UK Houses of Parliament), Oct 14, 2022.
- AI experts are increasingly afraid of what they're creating, by Kelsey Pieper, *Vox*, November 28, 2022.
- Rethinking the purpose of AI, interview by Mark Leonard, *The World in 30 Minutes*, European Council on Foreign Relations, December 2, 2022.
- Are we living in an AGI World?, interview by Kay Firth-Butterfield, *In AI We Trust?* podcast, January 18, 2023.
- Hot Topic at the WEF: Artificial Intelligence, interview by Harry Stitzel, SRF (Swiss Radio and Television), January 20, 2023.
- WAICF '23: Renowned AI Professor: Don't be fooled by ChatGPT, by Ben Wodecki, *AI Business*, February 10, 2023.

- Man beats machine at Go in human victory over AI, by Richard Waters, *Financial Times*, February 17, 2023.
- Interview Of The Week: Stuart Russell, Director Of The Center For Human-Compatible AI, interview by Jennifer Schenker, *The Innovator*, March 3, 2023.
- The Trouble With AI, interview by Sam Harris, *Making Sense* podcast, March 7, 2023.
- How to stop AI waging war on humans, interview by Danny Fortson, *The Times*, March 11, 2023.
- Interview on AI, by Jan Christoph Wiechmann, *Die Stern*, March 19, 2023.
- AI Expert Wants Developers to Take Responsibilities Seriously, interview by Ed Ludlow, *Bloomberg TV*, March 30, 2023.
- Beyond ChatGPT: Stuart Russell on the Risks and Rewards of AI, interview by Jerry Kaplan, Commonwealth Club, San Francisco, and National Public Radio, April 3, 2023.
- We must slow down the race to God-like AI, by Ian Hogarth, *Financial Times*, April 12, 2023.
- Stuart Russell on why A.I. experiments must be paused, interview by Michael Smerconish, *CNN*, April 1, 2023.
- AI has much to offer humanity. It could also wreak terrible harm. It must be controlled.. *The Guardian*, April 2, 2023.
- An Open Letter Asks AI Researchers To Reconsider Responsibilities, interview by Ira Flatow, *NPR Science Friday*, April 7, 2023.
- Stuart Russell calls for new approach for AI, a `civilization-ending' technology, by Rachel Leven, *Berkeley News*, April 7, 2023.
- AI: Blessing Or Curse?, interview by Aayush Ailawadi, *India Today*, April 10, 2023.
- AI guru Prof Stuart Russell explains why he signed a letter with Elon Musk and others to pause AI development, by Rachna Manojkumar Dhanrajani, *Business Today* magazine, April 11, 2023.
- Interview with Stuart Russell: Why I called for a moratorium on giant AI experiments, interview by Chen Yikai, *Beijing News*, April 13, 2023.
- What Keeps a Leading AI Scientist Up At Night, interview by Benjamin Hart, *New York Magazine*, April 17, 2023.
- interview by Sophy Ridge, *Sophy Ridge on Sunday*, Sky, April 30, 2023.
- Artificial intelligence: Powerful AI systems 'can't be controlled' and 'are causing harm', says UK expert, by Adam Arnold, *Sky News*, April 30, 2023.
- 'Genuinely scary': Sophy Ridge panics when expert explains how easily he can frame her, by Ciaran McGrath, *Daily Express*, April 30, 2023.
- How to stop runaway AI, *Freethink* podcast, May 8, 2023.
- Ministers not doing enough to control AI, says UK professor, interview by Harry Taylor, *The Guardian*, May 13, 2023.
- AI `could be like an alien invasion' says British professor, interview by Rhys Blakely, *The Times*, London, May 13, 2023.
- AI pioneer warns UK is failing to protect against `existential threat' of machines, by Rob Freeman, *The Independent*, May 13, 2023.
- How can humans maintain control over AI, forever?, *Boston Globe*, May 15, 2023.

- Stuart Russell and Toby Walsh, Examining regulation for ChatGPT, interview by Kelly Forbes, *AI Asia Pacific Institute Podcast*, May 15, 2023.
- Living with AI: 2021 Reith lecturer Stuart Russell on recent developments in artificial intelligence, interview by Anita Anand, *BBC Radio 4*, May 23, 2023.
- 'We do not understand how these systems work': Expert Stuart Russell on why AI systems need urgent regulations, interview by Aayush Ailawadi, *Business Today* magazine, May 28, 2023.
- We've reached a turning point with AI, expert says, interview by Jessica Chia and Bethany Cianciolo, *CNN*, May 31, 2023.

Recent Graduates

- Adam Gleave (FAR.ai)
- Michael Dennis (DeepMind)
- Paria Rashidinejad (Postdoc at BAIR)
- Andrea Bajcsy (Prof at CMU)

Affiliates

Journal Articles

- Tom Lenaerts and Azel Abels. Dealing with Expert Bias in Collective Decision-Making. *Artificial Intelligence*.
- Tom Lenaerts, Francisco Santos, and Elias Fernandez. EGTtools: Evolutionary Game Dynamics in Python. *iScience*.
- Tom Lenaerts. Inferring Strategies From Observations in Long Iterated Prisoner's Dilemma Experiments. *Scientific Reports*.
- Tom Lenaerts. Delegation to Artificial Agents Fosters Prosocial Behaviors in the Collective Risk Dilemma. *Scientific Reports*.
- Tom Lenaerts et al. Fast Deliberation is Related to Unconditional Behavior in Iterated Prisoners' Dilemma Experiments. *Scientific Reports*.
- Tom Lenaerts et al.
- Gillian Hadfield. Legal Markets. *Journal of Economic Literature*.
- Vincent Conitzer, Gillian Hadfield, and Shannon Vallor. The Impact of Auditing for Algorithmic Bias. *Communications of the ACM*.

Conference Papers

- Pulkit Verma, Shashank Rao Marpally, and Siddharth Srivastava. Discovering User-Interpretable Capabilities of Black-Box Planning Agents. *KR 2022*.
- David Krueger. Defining and Characterizing Reward Hacking. *NeurIPS 2022*.

Media

- Brian Christian. How a Google Employee Fell for the Eliza Effect. *The Atlantic.*

- Thomas Krendl Gilbert and Nathaniel Lubin. Social media is polluting society. Content moderation alone won't fix the problem. *MIT Technology Review*.

Working Papers

- Jonathan Stray, Dylan Hadfield-Menell, et al. Building Human Values into Recommender Systems: An Interdisciplinary Synthesis.
- Siddharth Srivastava and Rushang Karia. Relational Abstractions for Generalized Reinforcement Learning on Symbolic Problems.

Research Fellows

- Justin Svegliato, Samer Nashed, and Su Lin Blodgett. Fairness and Sequential Decision Making: Limits, Lessons, and Opportunities.
- George Obaido et al. An Interpretable Machine Learning Approach for Hepatitis B Diagnosis. *Applied Sciences.*
- Justin Svegliato et al. Competence-Aware Systems. *Artificial Intelligence Journal*.
- Erdem Biyik and Megha Srivastava. Assistive Teaching of Motor Control Task to Humans. *NeurIPS 2022*.

Graduate Students

Conference Papers

- Caspar Oesterhald*, Johannes Treutlein*, Emergy Cooper, and Rubi Hudson. Incentivizing Honest Performance Predictions with Proper Scoring Rules. *UAI 2023*.
- Micah Carroll and Smitha Milli. Emotional and Political Effects of Twitter's Ranking Algorithm. *Knight Institute Symposium on "Algorithmic Amplification.*"
- Micah Carroll et al. Harms from Increasingly Agentic Algorithmic Systems. *FAccT 2023*.
- Rachel Freedman, Peter Barnett, Justin Svegliato, and Stuart Russell. Active Reward Learning from Multiple Teachers. *AAAI 2023 Workshop on AI Safety*.
- Rachel Freedman and Oliver Daniels-Koch. The Expertise Problem: Learning from Specialized Feedback. *NeurIPS 2022*.
- Micah Carroll, Dylan Hadfield-Menell, Stuart Russell, and Anca Dragan. Estimating and Penalizing Induced Preference Shifts in Recommender Systems. *ICML 2022*.
- Micah Carroll, Anca Dragan, et al. Uni[MASK]: Unified Inference in Sequential Decision Problems. *NeurIPS 2022*.
- Micah Carroll, Anca Dragan, et al. Optimal Behavior Prior: Improving Human-AI Collaboration Through Generalizable Human Models. *NeurIPS 2022*.
- Micah Carroll, Anca Dragan, et al. Time-Efficient Reward Learning via visually Assisted Cluster Ranking. *NeurIPS 2022*.

Working Papers

- Lawrence Chan, Buck Shlegeris, Fabien Roger, and Euan McLean. Language Models are Better than Humans at Next-Token Prediction.

- Lawrence Chan, Richard Ngo, Sören Mindermann. The Alignment Problem from a Deep Learning Perspective.

CHAI Talks/ Honors

Stuart Russell - talks/ honors

In May 2022, Professor Russell participated in three panels on AI at the 2022 World Economic Forum meeting in Davos, Switzerland.

On July 29th, Profesor Russell was awarded the IJCAI Research Excellence Award for 2022. From the website: "The Research Excellence award is given to a scientist who has carried out a program of research of consistently high quality throughout an entire career yielding several substantial results. Past recipients of this honor are the most illustrious group of scientists from the field of Artificial Intelligence."

On October 18th at the invitation of Lord McFall, Speaker of the House of Lords in the UK Parliament, Professor Russell delivered the Lord Speaker's Lecture on the topic of "Artificial Intelligence: Promise and Peril." On October 20th, Professor Russell delivered the keynote address at the launch of Israel's AI safety research community in Haifa followed by a "fireside chat" with Professor Emeritus Yoav Shoham.

On December 9th, Professor Russell participated in a symposium preceding the Nobel Prize award ceremonies in Stockholm.

In January, Professor Russell spoke at two events at the 2023 World Economic Forum meeting in Davos on "Why Human Perspective is Essential to Unlock the Transformative Value of AI" and "The Long View for Technology." At the same meeting, he was appointed co-chair of the World Economic Forum Council on the Future of AI.

Student/recent grad jobs, talks, honors

Rachel Freedman and Oliver Daniels-Koch won the AI Risk Analysis award at the NeurIPS ML Safety Workshop for their paper "The Expertise Problem: Learning from Specialized Feedback." Rachel was also named a Rising Star in AI Ethics by the Women in AI Ethics global initiative.

Erdem Bıyık gave an invited talk on his research on learning preferences for interactive autonomy at Sonoma State University.

Micah Carroll was accepted to the first iteration of the AI Policy Hub at UC Berkeley, aiming to explore how the latest recommender systems research could be translated into policy-relevant

information. Micah's research on incentives for manipulation in recommender systems was also featured on the Berkeley Engineering website.

Nisan Stiennon gave an invited talk on "Metagames and Imprecise Probability" at the Topos Institute's weekly seminar. Nissan also gave an invited talk at an AI safety research organized by Effective Altruism Philadelphia.

Jonathan Stray joined with Brandie Nonnecke from Berkeley's CITRIS Policy lab, the Center for Democracy and Technology, and other scholars to file a brief for the Supreme Court in an upcoming case that asks whether "targeted recommendations" should be protected under the same law that protects the "publishing" of third party content. This law, called CDA Section 230, protects virtually all major apps and platforms that provide access to other people's information -- from search engines to music streaming to social media. CHAI is of course very concerned with the personal and societal effects of recommender systems; however, we believe that creating a legal distinction around the type of algorithm used will only disincentivize innovative new algorithms for selecting content -- exactly the type of safe and beneficial algorithms that we wish to create. Instead, we argued that platforms should be liable based on whether they contributed to harmful effects, not on what type of algorithm they use.

Affiliate talks/ honors

Brian Christian was interviewed by a number of media outlets including NPR, the New York Times, The New Yorker, Radio New Zealand, and Last Week Tonight with John Oliver. Brian also delivered the opening keynote address at "The Value Connection Workshop" in Amsterdam, an interdisciplinary workshop on human values, hosted by TU Delft and sponsored by the Royal Netherlands Academy of Arts and Sciences and the Delft Design for Values Institute.

Brian Christian was named one of the inaugural recipients of the National Academies Eric and Wendy Schmidt Awards for Excellence in Science Communication, given by The National Academies of Sciences, Engineering, and Medicine in partnership with Schmidt Futures for his book "The Alignment Problem." He will be honored with an event hosted at the National Academy of Sciences in Washington, DC in November.

Regulation & Governance

Over the past year, CHAI has been deeply involved in shaping the emerging regulatory response to AI. While CHAI has been discussing policy as a hypothetical for many years, it is now becoming a reality.

Professor Russell's involvement was instrumental in elevating recommender systems to "high-risk" in the European Union's AI Act and has met with a variety of global policymakers. CHAI's work has already set the stage for future generations of safer AI systems.

Over the last 12 months, Professor Russell has met with a wide-array of global policymakers:

- Canada:
  - 1/5/23: Presentation to Policy Horizons (Federal Government)
- UK:
  - 1/17/23: Foreign Office consultation on AI existential risk
  - 3/20/23: 10 Downing Street meeting on AI risk/opportunity
  - 5/10/23: Govt Office for Science meeting
  - 5/12/23: Meeting with former head of UK Office for AI
- US
  - 1/13/23: meeting with Sen. Heinrich (NM) staff
  - 2/1/23: meeting with State Department on AI organizations and engagement
  - 4/13/23: meeting with Representative Bill Foster (IL)
  - 4/14/23: meeting with Senator Schumer (NY) staff
  - 4/14/23: meeting with Senator Blumenthal (CT) staff
  - 4/14/23: meeting with Senator Heinrich staff
  - 5/2/23: Keynote at Brookings Institute Forum for Cooperation on AI
- EU
  - Multiple calls with MEPs and EU Commission negotiators on EU AI Act
- France
  - 2/13/23: Institut Montaigne
  - 2/13/23: Finance ministry
- Singapore
  - 3/31/23: Meeting with Minister of Communication and Information and staff
- Argentina
  - 3/17/23: Ministry of Science keynote
- Netherlands
  - 4/12/23: Parliament/ERO keynote
- China
  - 6/8/23-6/10/23: multiple meetings in Beijing (senior academics and government officials)
- World Economic Forum
  - Co-chair of Global Futures Council on AI
  - 4/27/23: Keynote at WEF meeting on Responsible AI Leadership
- Lethal Autonomous Weapons Systems (LAWS):
  - 2/15/23: Keynote at Summit on Responsible AI in Military Affairs, the Hague
  - 2/23/23: Keynote at Costa Rica summit on LAWS (LatAm/Caribbean nations)
  - 4/24/23: Testimony to House of Lords Select Committee
  - 5/3/23: US/India Track II discussions
  - 5/18/23: CCW Geneva EU/Philippines event keynote
  - 5/18/23: CCW Geneva Dinner with ambassadors of Norway, Netherlands
  - 6/15/23: Testimony to All-Party Parliamentary Group on AI Weapons
- GPAI
  - Member, Responsible AI Working Group

- OECD
    - Co-chair of Expert Group on AI Futures
    - 1/26/23 onwards: several meetings with OECD officials to revise AI definitions
    - 3/27/23: Keynote at OECD Conference on AI and Work
    - 4/19/23: Keynote at OECD Conference on AI Governance
- UNESCO
    - Member, AI Principles implementation group
    - 5/25/23: Keynote for UNESCO Ministerial meeting on AI in Education
- UNECE
    - 3/24/23: Discussion on AI compliance

Hires, new affiliates, CHAI workshops and funding related activity.

Jonathan Stray joined CHAI as a visiting scholar in April 2021. He was hired as a full-time Senior Scientist in April 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.

Brian Judge was hired as a full-time Policy Fellow in May 2022. He holds a PhD in Political Science from the University of California, Berkeley.

CHAI held its 7th annual workshop on June 16-18, 2023 at the Asilomar Conference Grounds in Pacific Grove, California. In each of 8 subject areas, the over 200 attendees were brought up to speed in a plenary tutorial, and then attended deep dives with no more than 2 tracks. The subjects included: LLM alignment, Human Value Learning, Well-founded AI, Cooperative AI, Robust & Trustworthy AI, AI Governance, Explainability & Interpretability, and Human Cognition. There were sidebar meetings on topics such as AI ethics and AI safety, democratic and deliberative processes for AI alignment, and China AI safety and governance, This was the largest ever CHAI workshop. Feedback from participants was enthusiastic.