Beneficial AGI

Participants

2019, Puerto Rico

future of life INSTITUTE

Anthony Aguirre is a Professor of Physics at the University of California, Santa Cruz. He has worked on a wide variety of topics in theoretical cosmology and fundamental physics, including inflation, black holes, quantum theory, and information theory. He also has strong interest in science outreach, and has appeared in numerous science documentaries and is author of the formcoming book *Cosmological Koans*. He is a co-founder of the Future of Life Institute, the Foundational Questions Institute, and Metaculus (http://www.metaculus.com/).

Daniela Amodei manages the natural language processing team and is head of people operations at OpenAI. She previously spent 5 years at Stripe, where she most recently managed several business and operations teams in the Risk organization, and led early recruiting efforts for the company. Prior to Stripe, Daniela was a congressional aide for a Congressman in Washington, DC, worked in field operations on two political campaigns and spent time working in international development.

Nima Anari is a researcher in the Computer Science Theory Group at Stanford University. He will be at the Simons Institute for the Theory of Computing during spring 2019 as a Microsoft Research Fellow for the Geometry of Polynomials program. Previously, Nima was a research fellow at the Simons Institute for the Theory of Computing, and a postdoctoral researcher at Stanford MS&E. He obtained his Ph.D. in Computer Science at UC Berkeley. Nima is broadly interested in the design and analysis of algorithms and more generally theoretical computer science.

Amanda Askell recently completed her PhD in philosophy at New York University with a thesis in infinite ethics, prior to which she received a BPhil in philosophy from the University of Oxford. Her research interests include formal ethics, formal epistemology, and decision theory. She currently works on AI policy at OpenAI, with a focus on the dynamics of AI development. She has also been working on how to identify and train good human judges of debate for the debate approach to AI alignment.

Shahar Avin is a postdoctoral researcher at the Centre for the Study of Existential Risk (CSER). He works with CSER researchers and others in the global catastrophic risk community to identify and design risk prevention strategies, through organizing workshops, building agent-based models, and by frequently asking naive questions. Prior to CSER, Shahar worked at Google for a year as a mobile/web software engineer. His PhD was in philosophy of science, on the allocation of public funds to research projects. His undergrad was in physics and philosophy of science, which followed his mandatory service in the IDF.

Jimmy Ba is an Assistant Professor in the Department of Computer Science at the University of Toronto. His research focuses on developing novel learning algorithms for neural networks. He is broadly interested in questions related to reinforcement learning, computational cognitive science, artificial intelligence, computational biology and statistical learning theory. His long-term research goal is to address a computational question: How can we build general problem-solving machines with human-like efficiency and adaptability? Jimmy completed PhD under the supervision of Geoffrey Hinton. Both his master (2014) and undergrad degrees (2011) are from the University of Toronto under Brendan Frey and Ruslan Salakhutdinov. He was a recipient of Facebook Graduate Fellowship 2016 in machine learning.

Elizabeth Barnes is a research assistant to the chief scientist at DeepMind. Prior to this she studied computer science at Cambridge and interned at the Centre for Human-Compatible AI.

Seth Baum is Co-Founder and Executive Director of the Global Catastrophic Risk Institute, a nonprofit and nonpartisan think tank. He leads an interdisciplinary research agenda of risk and policy analysis of catastrophic risks, focusing primarily on artificial intelligence and nuclear war. His work on AI includes social science for improving the human institutions that develop and influence AI, risk analysis for evaluating AI decisions, and ethics for what AI should be designed to do. His AI work includes basic research as well as outreach to important decision-makers in government, industry, and other sectors. He is based in New York City.

Haydn Belfield is Academic Program Manager at the Centre for the Study of Existential Risk (CSER), where he works across all of CSER's research projects. He has a background in policy and politics, including as a Policy Associate to the University of Oxford's Global Priorities Project, as a Senior Parliamentary Researcher to a British Shadow Cabinet Minister, and a degree in Philosophy, Politics and Economics from Oriel College, University of Oxford.

Yoshua Bengio is a professor in the Department of Computer Science and Operations Research at the University of Montréal, and a pioneer of deep learning. He is also scientific director of Mila, member of the NeurIPS advisory board, program co-director of the Canadian Institute for Advanced Research (CIFAR) Neural Computation and Adaptive Perception program, and Canada Research Chair in Statistical Learning Algorithms. Yoshua is currently the action editor for the Journal of Machine Learning Research and an associate editor for Neural Computation.

Liv Boeree is a science communicator, TV presenter and games specialist. An astrophysics graduate turned professional poker player, she has won multiple championship titles on the international poker circuit. Today, her primary focus is science and rationality outreach via talks, videos and written articles. A strong supporter of the EA movement, Liv co-founded Raising for Effective Giving in 2014 - a nonprofit that fundraises for EA-approved causes from within the poker community. Liv also works with EffectiveGiving.org, a newly launched network of major donors who seek to maximise the impact of their philanthropy.

Blake Borgeson is the technical co-founder of Recursion Pharmaceuticals, where they use big experiments and data analysis to look for new treatments for rare diseases. A while ago, he co-founded and led the initial technical development of buildasign.com. Blake worked for several years with the Marcotte lab at UT-Austin on bioinformatics research questions. He also supports several existential risk organizations working to help humanity survive and thrive through the oncoming wave of incredible technology we're all building right now, particularly MIRI (Machine Intelligence Research Institute).

Nick Bostrom is a Swedish-born philosopher with a background in theoretical physics, logic, computational neuroscience, and artificial intelligence. He is Professor at Oxford University, where he leads the Future of Humanity Institute as its founding director. He is the author of some 200 publications, including Anthropic Bias (2002), Global Catastrophic Risks (2008), Human Enhancement (2009), and Superintelligence: Paths, Dangers, Strategies (2014), a New York Times bestseller. Bostrom's work, which traverses science, philosophy, ethics, and technology, has illuminated the links between our present actions and long-term global outcomes, casting a new light on the human condition.

Malo Bourgon is COO of the Machine Intelligence Research Institute (MIRI), where he oversees MIRI's day-to-day operations and program activities. Before becoming COO, Malo worked for MIRI as a program management analyst, helping implement many of MIRI's current systems, processes, and program activities. He also co-chairs the committee on the Safety and Beneficence of Artificial General Intelligence and Artificial Superintelligence of the IEEE Global Initiative for Ethical Considerations in Artificial Intelligence and Autonomous Systems, and is an advisor to the Berkeley Existential Risk Initiative (BERI). Malo joined MIRI in 2012 shortly after completing a master's degree in engineering at the University of Guelph.

Jared Brown is an analyst in homeland security and emergency management policy at the U.S. Congressional Research Service, where he advises and supports Members of Congress and their staffs in their legislative and oversight responsibilities. Jared also works with the Future of Life Institute and the Global Catastrophic Risk Institute to affect incremental reductions in global catastrophic and existential risk by influencing the U.S. policymaking process, including as it relates to emerging technologies. Jared has a Master's in Public Policy, with a specialty in Humanitarian Crisis Policy, from Georgetown University's McCourt School of Public Policy.

Sasha Brown has been at DeepMind for over a year, working with the engineering teams to help put ethics into practice. Her previous jobs include Google UK Charities team, Obama's 2012 campaign, and a brief but career defining stint as the yellow M&M on Twitter. She has degrees in neuroscience, and behavioural economics and social policy, as well as an MBA from INSEAD.

Miles Brundage recently joined OpenAI, where he works as a Research Scientist on the policy team. Previously, he was a Research Fellow at the University of Oxford's Future of Humanity Institute (where he remain a Research Associate). He is also a PhD candidate in Human and Social Dimensions of Science and Technology at Arizona State University and a member of Axon's AI and Policing Technology Ethics Board.

Erik Brynjolfsson is Director of the MIT Initiative on the Digital Economy, Schussel Family Professor at the MIT Sloan School, and Research Associate at The National Bureau of Economic Research (NBER). His research examines the effects of information technologies and AI on the economy. At MIT, his courses include the Economics of Information, and The Analytics Lab. He is one of the most cited scholars in information systems and economics, and co-author of *The Second Machine Age: Work, Progress and Prosperity in a Time of Brilliant Technologies*, as well as *Machine | Platform | Crowd*. Erik is currently on sabbatical at Stanford University where he is assessing the types of tasks that will be most affected by AI and the implications for income, work and the economy.

Xiongshan "Sam" Cai is the Deputy Director and the Chief Researcher of the Legal Research Center at Tencent Research Institute. He is also a supervisor for LL.M. students at the Communication University of China. He was a visiting scholar at Yale University. Mr. Cai received a master's degree from the Transnational Law and Business University in Korea and another master's degree from the University of Paris I: Panthéon-Sorbonne's EU and International Economic Law Programme.

Jeff Cao is a senior research fellow at Tencent Research Institute, his research areas include AI/robot law & ethics, platform law, data governance, regulation of self-driving technologies, legal tech, and digital intellectual property. He is also an expert at Internet Society of China, a part-time researcher at East China University of Political Science and Law, and a member of National AI Standardization Group. He has published dozens of articles on the Internet, and in newspapers and academic journals.

Natalie Cargill is the Founder & Executive Director of Effective Giving, a nonprofit which helps high-net-worth donors maximize the impact of their philanthropy. She coordinates a group of donors who collectively give $70m annually to one of the world's most pressing problems, and has lectured on the principles of effective giving at the University of Cambridge and King's College London. Natalie is also an associate barrister at Serjeants' Inn Chambers, a tier-1 set, known for cases of ethical and political importance. She graduated with a first-class degree from the University of Oxford and has worked with the UN Human Rights Council, 80,000 Hours, and Sentience Politics.

Brad Carson is a professor at the University of Virginia. He is also a senior advisor at Boston Consulting Group. He was appointed by President Barack Obama in 2015 as the acting Under Secretary of Defense for Personnel & Readiness at the Department of Defense. Mr. Carson earlier served as the Under Secretary of the U.S. Army and as General Counsel of the U.S. Army. From 2001-2005, he served two terms as a United States Congressman from Oklahoma. Later, he was appointed to the faculty of the business and law schools at the University of Tulsa, where he directed the National Energy Policy Institute and taught academic courses on energy policy, property law, negotiation and game theory, globalization, and law and literature.

Meia Chita-Tegmark is a postdoctoral researcher working on human-robot interactions at the Tufts School of Engineering. She did her M.Ed. at the Harvard Graduate School of Education and her Ph.D. in the Department of Psychological and Brain Sciences at Boston University, focusing on a variety of topics in developmental psychology, such as atypical social development, attention mechanisms and learning strategies. Meia has strong interests in the future of humanity and big picture questions, and she is a co-founder of the Future of Life Institute.

Brian Christian is the author, with Tom Griffiths, of *Algorithms to Live By*, a #1 Audible bestseller, Amazon best science book of the year and MIT Technology Review best book of the year. He is the author of *The Most Human Human*, a Wall Street Journal bestseller, New York Times Editors' Choice, and New Yorker favorite book of the year. Brian holds degrees in computer science, philosophy, and poetry from Brown University and the University of Washington, and his books have been translated into nineteen languages. He is the Director of Technology at McSweeney's Publishing, an active open-source contributor to projects such as Ruby on Rails, and a Visiting Scholar at the University of California, Berkeley.

Teddy Collins is a strategy researcher at DeepMind, and coauthor of "Team of Teams: New Rules of Engagement for a Complex World".

Ariel Conn is the Director of Media and Outreach for the Future of Life Institute. She oversees online outreach and communication efforts, as well as leading collaboration efforts with other organisations. Her work covers a range of fields, including artificial intelligence (AI) safety, AI policy, lethal autonomous weapons, nuclear weapons, biotechnology, and climate change. Ariel has degrees in English, physics, and geophysics and she's worked with NASA, the Idaho National Laboratory, the National Energy Technology Laboratory, MIT, and Virginia Tech.

Andrew Critch is with the UC Berkeley Center for Human Compatible AI. He earned his PhD in mathematics at UC Berkeley, and during that time, he cofounded the Center for Applied Rationality and the Summer Program on Applied Rationality and Cognition (SPARC). Andrew has been with MIRI since 2015, and his current research interests include logical uncertainty, open source game theory, and avoiding arms race dynamics between nations and companies in AI development.

Jessica Cussins is an AI policy specialist for FLI and a Research Fellow at the UC Berkeley Center for Long-Term Cybersecurity. She is passionate about the ethics and governance of emerging technologies and has worked in technology policy with organizations including The Belfer Center for Science and International Affairs and The Future Society. She received her master's degree in public policy from the Harvard Kennedy School and her bachelor's in anthropology from the University of California, Berkeley with highest distinction honors.

Allan Dafoe is Associate Professor and Senior Research Fellow in the International Politics of Artificial Intelligence at the University of Oxford. He is also Director of the Center for the Governance of AI (GovAI), at University of Oxford's Future of Humanity Institute. Allan's research focuses on characterizing and building the field of AI governance, and on global cooperation around transformative AI. Allan's prior research examined the liberal peace, the role of reputation and honor as motives for war, the role of technology in shaping history, and statistical methods for credible causal inference. For more information see: www.governance.ai and www.allandafoe.com

Pranab Das is the Principal Advisor for the Diverse Intelligences initiative at Templeton World Charity Foundation, Inc. He is currently Professor of Physics at Elon University. Pranab's current research collaborations include a project focusing on the "active matter" found inside living cells. Earlier in his career, he specialized in the dynamics of neural network architectures. He previously served as the Executive Editor of the International Society for Science and Religion's Library Project. Previously, he headed the Global Perspectives on Science and Spirituality program, a multiyear, multimillion dollar project to identify and support non-Western contributions to the science and spirituality dialogue. His background and interests include neural networks, computational dynamics, soft condensed matter, and the history and philosophy of science.

Tucker Davey is a writer, editor, and researcher for the Future of Life Institute, where he focuses on AI and climate change. He graduated from Boston College in 2016 with a major in Political Science and minors in Philosophy and Hispanic Studies. After spending his postgraduate summer working at an orphanage in Honduras, he turned his focus to effective altruism and existential risk mitigation with FLI. Tucker likes to think about the relationship between technology and society, and he's in the constant midst of figuring out how he can best use his writing and imagination to make the world a better place.

Gaia Dempsey is the Founder and CEO of 7th Future, a consultancy that partners and co-invests with technology leaders and communities to build and launch global-impact innovation models, with a commitment to openness, integrity, resilience, and long-term thinking. Prior to 7th Future, in 2010, Gaia co-founded DAQRI, an augmented reality hardware company that delivers a complete professional AR platform to the industrial and enterprise market.

Anca Dragan is an Assistant Professor in the EECS Department at UC Berkeley. Her goal is to enable robots to work with, around, and in support of people. She runs the InterACT lab, which focuses on algorithms that move beyond the robot's function in isolation, and generate robot behavior that also accounts for interaction and coordination with end-users. She works across different applications, from assistive robots, to manufacturing, to autonomous cars, and draws from optimal control, planning, estimation, learning, and cognitive science. She serves on the steering committee for the Berkeley AI Research Lab and is a co-PI for the Center for Human-Compatible AI.

Eric Drexler is currently a Senior Fellow at the Oxford Martin School, and Senior Research Fellow at the Future of Humanity Institute, University of Oxford. His work centers on development-oriented models of the emergence of high-level AI technologies, and is currently supported by a European Research Council grant in AI macrostrategy.

Allison Duettmann is a researcher and program coordinator at Foresight Institute, a non-profit institute for technologies of fundamental importance for the future of life. Her research focuses on the reduction of existential risks, especially from artificial general intelligence. At Existentialhope.com, she keeps a curated map of resources, organizations and people working toward positive long-term futures for life. The project is collaborative and for everyone who wants to work toward grand futures but doesn't know where to start. Allison speaks and moderates panels on existential risks and existential hope, AI safety, cybersecurity, longevity, blockchains, ethics in technology, and more. Allison holds an MS in Philosophy & Public Policy from the London School of Economics, where she developed an ethical framework for AGI that relies on Rawl's Reflective Equilibrium and NLP.

David Duvenaud is an assistant professor at the University of Toronto. His research focuses on constructing deep probabilistic models to help predict, explain and design things. Previously, he was a postdoc in the Harvard Intelligent Probabilistic Systems group with Ryan Adams. He did his Ph.D. at the University of Cambridge and his M.Sc. at the University of British Columbia, where he worked mostly on machine vision. David spent a summer working on probabilistic numerics at the Max Planck Institute for Intelligent Systems, and the two summers before that at Google Research, doing machine vision. He co-founded Invenia, an energy forecasting and trading firm where he still consults. David is also a founding member of the Vector Institute.

Peter Eckersley is Director of Research at the Partnership on AI, a collaboration between the major technology companies, civil society and academia to ensure that AI is designed and used to benefit humanity. He leads PAI's research on machine learning policy and ethics, including projects within PAI itself and projects in collaboration with the Partnership's extensive membership. Peter's AI research interests are broad, including measuring progress in the field, figuring out how to translate ethical and safety concerns into mathematical constraints, and setting sound policies around high-stakes applications such as self-driving vehicles, recidivism prediction, cybersecurity, and military applications of AI. Prior to PAI, Peter was Chief Computer Scientist at EFF, and led a team that worked on numerous computer security, privacy and Internet policy topics.

El Mahdi El Mhamdi pioneered Byzantine-Resilient Machine Learning, devising a series of provably safe algorithms he recently presented at NeurIPS and ICML. Interested in theoretical biology, his work also includes the analysis of error propagation in networks, applied to both neural and bio-molecular networks. He also works on the safety of reinforcement learning, introducing a trade-off between Byzantine-resilient perception and safe-interruptibility. After a Bachelor in Morocco, he moved to France, Germany then Switzerland to graduate in Physics. Before his PhD at EPFL, he worked as a researcher in condensed-matter Physics and co-founded Wandida, now an EPFL science video library, and Mamfakinch, a Moroccan citizen-media winning the 2012 Google & Global Voices Breaking Borders Award.

Owain Evans is a research scientist at the University of Oxford, working on AI Safety and Reinforcement Learning. He is interested in inferring preferences, safe exploration in RL, and in using ML to improve human reasoning and deliberation. He led a collaboration on "Inferring Human Preferences", along with Andreas Stuhlmueller and Noah Goodman. He is a collaborator and board member at new non-profit Ought, working on techniques for breaking down reasoning ("Factored Cognition"). Owain received his Ph.D at MIT.

Daniel Filan completed his undergraduate education at the Australian National University, where he worked on an honours thesis on algorithmic information theory. He is currently on break from an AI alignment PhD program with CHAI at UC Berkeley, where he thinks about transparency for machine learning, doing an internship programming Haskell at MIRI.

Jaime Fisac is a final-year Ph.D. student at UC Berkeley, advised by Profs. Shankar Sastry, Claire Tomlin, and Anca Dragan. He is interested in developing analytical and computational tools to safely deploy robotic and AI systems in the physical world. His goal is to ensure that autonomous systems such as self-driving cars, delivery drones, or home robots can operate and learn in the open while satisfying safety constraints at all times. His research focuses on the following areas: Safety analysis for learning robotic systems, scalable safety for multi-agent systems, and safe human-centered robotic and AI systems.

Iason Gabriel is a Senior Research Scientist at DeepMind Ethics and Society, where his work focuses on value alignment, responsible innovation, and globally beneficial AI. Before joining DeepMind, he was a Supernumerary Teaching Fellow in Politics at St John's College, Oxford and a member of the Centre for the Study of Social Justice (CSSJ). Iason previously earned a doctorate in Politics and an M.Phil in International Relations from the University of Oxford. He also spent time as a visiting scholar Harvard University and at Princeton University where he worked on a number of projects that address the problem of global justice. Outside of this environment, he spent a number of years working for the United Nations in Lebanon and Sudan on post-conflict recovery initiatives.

Danit Gal is a Project Assistant Professor at the Cyber Civilizations Research Center at the Keio University Global Research Institute in Tokyo, Japan. She is interested in technology geopolitics, safety and security, diversity and inclusion, and maximizing shared social benefit. Danit chairs the IEEE P7009 standard on the Fail-Safe Design of Autonomous and Semi-Autonomous Systems and serves on various committees of The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. She is an Associate Fellow at the Leverhulme Centre for the Future of Intelligence at the University of Cambridge, and an Affiliate at the Center for Information Technology Policy at Princeton University. Danit also serves as a board trustee at the Seed Token Project. Prior to joining Keio University, she was a Yenching Scholar at Peking University and International Strategic Advisor to the iCenter at Tsinghua University in Beijing, China.

Yarin Gal is an Associate Professor of Machine Learning at the University of Oxford Computer Science department, and head of the Oxford Applied and Theoretical Machine Learning Group (OATML). He is also the Tutorial Fellow in Computer Science at Christ Church, Oxford, and Fellow at the Alan Turing Institute. Previously, Yarin was a Research Fellow in Computer Science at St Catharine's College at the University of Cambridge. He obtained his PhD from the Cambridge machine learning group, working with Prof Zoubin Ghahramani and funded by the Google Europe Doctoral Fellowship. Prior to that he studied at Oxford Computer Science department for a Master's degree under the supervision of Prof Phil Blunsom.

Scott Garrabrant leads the agent foundations team at the Machine Intelligence Research Institute. He earned his PhD in mathematics from UCLA studying applications of theoretical computer science to enumerative combinatorics. His main research area is in logical uncertainty, and he is the primary author of "Logical Induction" (2016), a highly general method for assigning probabilities to logical sentences. He is also interested in other aspects of building a theory of embedded agency.

Carla Gomes is a Professor of Computer Science and the director of the Institute for Computational Sustainability at Cornell University. Her research area is Artificial Intelligence with a focus on large-scale constraint-based reasoning, optimization, and machine learning. Recently, she has become deeply immersed in the establishment of new field of Computational Sustainability, a new interdisciplinary research field, with the overarching goal of studying and providing solutions to computational problems for balancing environmental, economic, and societal needs for a sustainable future.

Joseph Gordon-Levitt is an American actor and filmmaker. He has starred in many films, including 500 Days of Summer, Inception, 50/50, The Dark Knight Rises, and Snowden (2016). He also founded the online production company hitRECord in 2004 and has hosted his own TV series, HitRecord on TV, since January 2014. As one of his recent projects with hitRECord, he is developing a short animated series called USAI, based on the question: Could a machine beat a human at the game of electoral politics?

Katja Grace is a researcher at the Machine Intelligence Research Institute (MIRI) in Berkeley. She contributes to AI Impacts, an independent research project focused on social and historical questions related to artificial intelligence outcomes. Her analyses include Algorithmic Progress in Six Domains (2013). She writes the blog Meteuphoric, and is a part-time PhD student in Logic, Computation, and Methodology at Carnegie Mellon University. Katja previously studied game theory—especially in signaling and anthropic reasoning.

Joshua Greene is Professor of Psychology and member of the Center for Brain Science faculty at Harvard University. For over a decade his lab has used behavioral and neuroscientific methods to study moral judgment, focusing on the interplay between emotion and reason in moral dilemmas. More recent work examines how the brain combines concepts to form thoughts and how thoughts are manipulated in reasoning and imagination. Other interests include conflict resolution and the social implications of advancing artificial intelligence. He is the author of *Moral Tribes: Emotion, Reason, and the Gap Between Us and Them*.

Roger Grosse is an Assistant Professor of Computer Science at the University of Toronto, and a founding member of the Vector Institute. His group's research focuses on machine learning, especially deep learning and Bayesian modeling. They aim to develop architectures and algorithms that train faster, generalize better, give calibrated uncertainty, and uncover the structure underlying a problem. They are especially interested in scalable and flexible uncertainty models, so that intelligent agents can explore effectively and make robust decisions at test time.

Gillian Hadfield is Professor of Law and Professor of Strategic Management at the University of Toronto, Faculty Affiliate at the Vector Institute and Center for Human-Compatible AI, and Senior Policy Advisor at OpenAI. Her research is focused on innovative design for legal and dispute resolution systems in advanced and developing market economies; governance for artificial intelligence (AI); the markets for law, lawyers, and dispute resolution; and contract law and theory. She teaches Contracts; Problems in Legal Design; Legal Design Lab, and Responsible AI at the University of Toronto. Previously, she was Professor of Law and Professor of Economics at the University of Southern California from 2001 to 2018. Her book is titled *Rules for a Flat World: Why Humans Invented Law and How to Reinvent It for a Complex Global Economy*.

Dylan Hadfield-Menell is a fifth year Ph.D. student at UC Berkeley, advised by Anca Dragan, Pieter Abbeel, and Stuart Russell. His research focuses on the value alignment problem in artificial intelligence. His goal is to design algorithms that learn about and pursue the intended goal of their users, designers, and society in general. Dylan's recent work has focused on algorithms for human-robot interaction with unknown preferences and reliability engineering for learning systems. He's also interested in work that bridges the gap between AI theory and practical robotics and work on the problem of integrated task and motion planning.

John Havens is Executive Director of The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. He is also Executive Director of The Council on Extended Intelligence and author of, *Heartificial Intelligence: Embracing our Humanity to Maximize Machines* (Penguin, 2016). @johnchavens

John Hering is Executive Chairman and Founder of Lookout. He has grown Lookout's footprint to tens of millions of users globally across consumer, enterprise, and government sectors. John is a frequent presenter at mobile and technology industry events including: RSA, Mobile World Congress, Black Hat Technical Security Conference, DEFCON, and Fortune Brainstorm. Additionally, John is an investor in dozens of technology startups focused on the areas of cybersecurity, artificial intelligence, enterprise software, and transportation technologies.

Danny Hernandez measures and forecasts AI progress at OpenAI.

José Hernández-Orallo is Professor at the Universitat Politecnica de Valencia, Spain, and Associate Fellow at the Leverhulme Centre for the Future of Intelligence. He received a B.Sc. and a M.Sc. in Computer Science from UPV, partly completed at the École Nationale Supérieure de l'Électronique et de ses Applications (France), and a Ph.D. in Logic with a doctoral extraordinary prize from the University of Valencia. His academic and research activities have spanned several areas of artificial intelligence, machine learning, data science and inductive programming. His most recent book "The Measure of All Minds" addresses an integrated view of the evaluation of natural and artificial intelligence (Cambridge University Press, 2017, PROSE Award 2018).

Irina Higgins is a Research Scientist at DeepMind, where she works in the Neuroscience team. Her work aims to bring together insights from the fields of machine learning and neuroscience to understand the computational principles underlying sensory representation learning. Before joining DeepMind, Irina was a British Psychological Society Undergraduate Award winner for her achievements as an undergraduate student in Experimental Psychology at Westminster University, followed by a DPhil at the Oxford Centre for Computational Neuroscience and Artificial Intelligence, where she focused on understanding the computational principles underlying speech processing in the auditory brain, and also worked on developing poker AI, applying machine learning in the finance sector, and speech recognition at Google Research.

Cyrus Hodes is advising the UAE Minister of State for AI. He oversees the Global Governance of AI Roundtable at the World Government Summit and the use of AI towards the SDGs (AI4SDG Center, Global Data Commons). He is a Partner at FoundersX Ventures, a silicon-valley based VC firm focusing on early stage AI startups. Driven by AI safety, he co-founded the AI Initiative within The Future Society, a 501(c)3 incubated at Harvard Kennedy School. He has hands-on experience in crisis countries (Afghanistan, RDC, South Sudan) leading humanitarian aid programs. Member: OECD Expert Group on AI; Council on Extended Intelligence; IEEE Ethically Aligned Design. Education: Sciences Po Paris, MA in Geostrategy (Paris II), MPA (Harvard).

Reid Hoffman is co-founder and executive chairman of LinkedIn and a partner at venture capital firm Greylock Partners. Prior to LinkedIn, Reid served as executive vice president at PayPal, where he was also a founding board member. He currently serves on the boards of Airbnb, Convoy, Edmodo, Xapo, Linkedin, and a number of not-for-profit boards, including Kiva, Mozilla Corporation, Endeavor, and Do Something. He is also the co-author of two New York Times best-selling books: The Start-up of You and The Alliance. Reid earned a master's degree in philosophy from Oxford University, where he was a Marshall Scholar, and a bachelor's degree with distinction in symbolic systems from Stanford University.

Tiejun Huang is a professor, and the Chair of the Department of Computer Science, School of EE&CS, Peking University. His research areas include visual information processing and neuromorphic computing. Professor Huang is the secretary general of the Artificial Intelligence Industry Technology Innovation Alliance, the advisory board of Computing Now of the IEEE Computer Society.

Tim Hwang is Director of the Harvard-MIT Ethics and Governance of AI Initiative, a philanthropic project working to ensure that machine learning and autonomous technologies are researched, developed, and deployed in the public interest. Previously, he was at Google, where he was the company's global public policy lead on artificial intelligence, leading outreach to government and civil society on issues surrounding the social impact of the technology. Dubbed "The Busiest Man on the Internet" by Forbes Magazine, his current research focuses on the geopolitical aspects of computational power and machine learning hardware, and the parallels between historical common law development and platform content policies.

Joichi "Joi" Ito is an activist, entrepreneur, venture capitalist and scholar focusing on the ethics and governance of technology, tackling complex problems such as climate change and redesigning the systems that support scholarship and science. As director of the MIT Media Lab and a Professor of the Practice in Media Arts and Sciences, he supports researchers at the Media Lab to deploy design, science and technology such AI, blockchain and synthetic biology to transform society in substantial and positive ways. Ito is a Visiting Professor of Law from Practice at the Harvard Law School, where he and Professor Jonathan Zittrain teach The Ethics and Governance of Artificial Intelligence. Ito is co-author with Jeff Howe of Whiplash: How to Survive Our Faster Future, and he writes a monthly column for WIRED magazine.

De Kai is a Distinguished Research Scholar at the International Computer Science Institute (ICSI) at Berkeley and a Professor of Computer Science and Engineering at the Hong Kong University of Science and Technology (HKUST), which he joined in its founding year in 1992 after a PhD in Computer Science from UC Berkeley and a postdoctoral fellowship at the University of Toronto; he also holds an Executive MBA from Kellogg (Northwestern University) and HKUST, and a BS in Computer Engineering from UCSD. He was selected by the Association for Computational Linguistics as one of only 17 scientists worldwide to be awarded the honor of founding ACL Fellow in 2011, with a citation for "significant contributions to machine translation and the development of inversion transduction grammar" which pioneered the integration of syntactic and semantic models into statistical machine translation paradigms.

Holden Karnofsky is the Executive Director of the Open Philanthropy Project, a philanthropic organization funded by Cari Tuna and Dustin Moskovitz. The Open Philanthropy Project funds a variety of work on AI safety, including its AI Fellows Program as well as work at OpenAI, the Center for Human-Compatible AI, Mila, the Future of Humanity Institute and the Machine Intelligence Research Institute. Open Philanthropy also works in a variety of other causes including criminal justice reform, farm animal welfare, biosecurity and pandemic preparedness, and life sciences. Previously, Holden co-founded GiveWell.

Zac Kenton is a Postdoc at the University of Oxford, in the Oxford Applied and Theoretical Machine Learning (OATML) group, working under Yarin Gal. His primary research interest is technical AI safety, currently working on robustness and safe exploration in reinforcement learning. Zac was previously a Research Assistant under Owain Evans at the Future of Humanity Institute, University of Oxford and a Visiting Researcher at the Montreal Institute for Learning Algorithms (MILA), under Yoshua Bengio. He also worked as a Data Scientist at ASI Data Science. He completed his PhD in 2017 at Queen Mary University of London in Theoretical Physics, where he worked on string theory and cosmology. Prior to his PhD, Zac studied Mathematics at the University of Cambridge.

Max Kleiman-Weiner is a Harvard Data Science and CRCS Postdoctoral Fellow at Harvard & MIT. His goal is to understand how the human mind works with enough precision that it can be implemented on a computer. He draws on insights from how people learn and think to engineer smarter and more human-like algorithms for artificial intelligence. Max received his PhD in Computational Cognitive Science from MIT, advised by Josh Tenenbaum, where he received fellowships from the Hertz Foundation and NSF. Previously, he was a Marshall Scholar in Statistics at Oxford, a Fulbright Fellow in Beijing, and before that was an undergraduate at Stanford. Max is also Chief Scientist of Diffeo, a start-up that he co-founded.

Victoria Krakovna is a research scientist in AI safety at DeepMind, focusing on objective specification problems, and a co-founder of the Future of Life Institute. Her PhD thesis in statistics and machine learning at Harvard University focused on building interpretable models. Viktoriya gained numerous distinctions for her accomplishments in math competitions, including a silver medal at the International Mathematical Olympiad and the Elizabeth Lowell Putnam prize.

János Kramár is a research engineer at DeepMind working on ways of analyzing multi-agent systems. In 2016 he interned at the Montreal Institute for Learning Algorithms, co-developing Zoneout, a state-of-the-art regularization method for recurrent neural nets. He also co-authored a paper surveying and calling for prospective research on secure environments for testing advanced AI systems, as well as a research priorities survey for FLI's Future of AI conference back in 2015. He holds a Masters in Statistics from Harvard University, and has worked in algorithmic trading. He was a top competitor in math and programming competitions at the national level in Canada, earning a bronze medal at the International Math Olympiad.

David Krueger: In the words of Yoshua Bengio, David is the "resident fear monger" at MILA, where he's published seminal works on normalizing flows and RNN regularization. He joined the field of deep learning in 2013, aiming to steer the development of AI towards socially beneficial outcomes, and hoping to be learn why Xrisk fears were misguided (it turns out they're not). Over the last 5 years, he's maintained that AI-Xrisk is primarily an issue of global coordination, mandating a completely unprecedented level of international cooperation. Recently, he's been interning at FHI (2016) and DeepMind (2018), shifting his research to focus more on technical safety problems (value learning, calibrated confidence, safety via myopia) and plotting his next move after graduation; will he... pivot into policy? start an AI safety bootcamp? go to China? All options are on the table!

Ramana Kumar is a research scientist in AGI safety at DeepMind, and a research associate at the Machine Intelligence Research Institute. Prior to DeepMind, he was a researcher in the Trustworthy Systems group at Data61. Kumar holds an MPhil in Advanced Computer Science from the University of Cambridge and completed his PhD at Cambridge in 2016 as a Gates Scholar. His dissertation, Self-compilation and Self-verification, earned the John C. Reynolds Award for an outstanding dissertation in the area of Programming Languages. Kumar is a lead developer of CakeML, a functional programming language with end-to-end machine-checked proofs of correctness. His research vision involves the use of formal methods to understand and construct well-behaved computer systems.

Martina Kunz is a research associate at the Leverhulme Centre for the Future of Intelligence (CFI) and a research affiliate at the Centre for the Study of Existential Risk (CSER), both at Cambridge University. For her PhD (soon to be completed) at the Cambridge Centre for Environment, Energy and Natural Resource Governance, she used natural language processing, data science and artificial intelligence to map and evaluate the regulatory techniques and effects of international environmental treaty systems. Prior to her PhD at Cambridge, Martina studied and worked at the University of Geneva, Tsinghua University, and the Graduate Institute of International and Development Studies. Her work at CFI focuses on international strategies, policies and governance mechanisms that could help mitigate the risks and harness the opportunities presented by AI.

Sean Legassick is co-lead of the Ethics and Society group at DeepMind, where he has spent the last three years researching the social, economic, ethical, legal and philosophical implications of AI. He is a software engineer and technology policy specialist with over 20 years of industry experience, together with a BA in Artificial Intelligence from Sussex University, and an MA in Sociology from Goldsmiths College, University of London. He was a lead developer at Demon Internet, the UK's first Internet provider, and has made notable contributions to the Apache Software Foundation and Django Foundation open source software projects.

Jan Leike is a Senior Research Scientist at DeepMind and a Research Associate at the Future of Humanity Institute, University of Oxford. Previously he was a PhD student at the Australian National University and wrote his dissertation on the theory of reinforcement learning. He is working on long-term technical problems of safety and alignment of reinforcement learning agents: How do we learn a good reward function? How can we design agents such that they are incentivized to act in accordance with our intentions? How can we avoid degenerate solutions to the objective function?

Jade Leung is the Head of Research & Partnerships with the Center for the Governance of Artificial Intelligence (GovAI) at the Future of Humanity Institute. She is also a researcher with the Center and a Rhodes scholar, undertaking a PhD in the geopolitics of AI at Oxford. Her research focuses on modelling relationships between firms, the government, and the research community with the view to understanding how these dynamics seed cooperation and conflict. As Head of Research & Partnerships, Jade plays a central role in steering GovAI's research priorities and projects. She also leads on our partnerships strategy, focusing on building high value relationships with key stakeholders.

Yang Liu is a Leverhulme Research Fellow in the Faculty of Philosophy and Associate Fellow in the Leverhulme Centre for the Future of Intelligence at the University of Cambridge. His research interests include Logic, Foundations of Probability, Mathematical and Philosophical Decision Theory, and Philosophy of Artificial Intelligence (AI). Before moving to Cambridge, he received his PhD in Philosophy from Columbia University. He is co-founder and co-organiser of a seminar series on Logic, Probability, and Games, as well as a conference series on Decision Theory and AI. More information about his research can be found at www.yliu.net.

Jelena Luketina is pursuing a PhD at Oxford University focusing on deep reinforcement learning. She used to work on transfer learning and optimization, as an intern at DeepMind and a research assistant at Aalto University. In her past life, she was a physics and applied mathematics student dabbling in science communication.

Tegan Maharaj is a senior PhD student at the Montreal Institute for Learning Algorithms (MILA). Her academic research has focused on understanding multimodal data with deep models, particularly for time-dependent data. At the practical end, Tegan has developed datasets and models for video and natural language understanding, and worked on using deep models for predicting extreme weather events. On the more theoretical side, her work examines how data influence learning dynamics in deep and recurrent models. Her most recent research efforts focus on ecosystem modeling. Tegan is concerned and passionate about AI ethics, safety, and the application of ML to environmental management, health, and social welfare.

Vishal Maini is Strategic Communications Manager at DeepMind, where he works on steering AI progress towards safe development and beneficial deployment. His focus areas include: technical AI safety, long-term strategic planning, education, recruiting, and coordination across key stakeholders. Vishal has his Bachelor's degree in Economics from Yale University, is the co-author of *Machine Learning for Humans*, and previously led growth at Upstart, a financial technology startup applying machine learning to credit risk assessment.

Jacques Mallah is a physicist and philosopher. He has written on computationalist interpretations of consciousness, computationalist interpretations of quantum mechanics, theoretical requirements for characterizing computation implementations, anthropic probability, and philosophy of mind. By day he is a medical physicist in radiation therapy. He holds a Ph.D. in physics from NYU.

Richard Mallah is Director of AI Projects at the Future of Life Institute, where he works to support the robust, safe, beneficent development of advanced artificial intelligence via meta-research, analysis, research organization, and advocacy. Richard also serves on the Executive Committee of IEEE's initiative on autonomous systems ethics, co-chairs that IEEE initiative's AGI committee, serves on the safety-critical AI and the labor & economy working groups at Partnership on AI, and serves as a senior advisor to the AI initiative of The Future Society. Richard also heads AI R&D at talent acquisition automation firm Avrio AI, where he leads innovation in machine learning, knowledge representation, and conversational agents. He is also an advisor to other startups and NGOs where he advises on AI, metaknowledge, and sustainability.

James Manyika is a senior partner at McKinsey & Company and chairman of the McKinsey Global Institute (MGI). Based in Silicon Valley, James has worked with the chief executives and founders of many of the world's leading technology companies. James was appointed by President Obama as vice chair of the Global Development Council at the White House (2012-16) and by the Commerce Secretary to the Digital Economy Board (2016-17). He serves on the boards of the Council on Foreign Relations, MacArthur Foundation and Hewlett Foundation, Oxford Internet Institute, MIT's Initiative on the Digital Economy, is a member of the standing committee for the Stanford-based 100 Year Study on AI, and a fellow at DeepMind and the Royal Society of Arts. A Rhodes Scholar, James received his DPhil. MSc. MA. from Oxford in AI and Robotics, Computation, BSc from University of Zimbabwe.

David Marble is a trustee and board member at The Casey and Family Foundation. He focuses on funding catastrophic and existential risk mitigation, and sees in AI both a potential existential risk and a panacea for other such risks. David is significantly influenced by the Effective Altruism movement and seeks a future for humanity characterized by meaningful, positive, intentional experience.

Jason Matheny is a member of the National Security Commission on Artificial Intelligence, established by Congress in 2018. Previously he was Assistant Director of National Intelligence, and Director of IARPA, a government organization that develops technologies for national intelligence. Before IARPA, he worked at Oxford University, the World Bank, the Applied Physics Laboratory, the Center for Biosecurity, and Princeton University, and was the co-founder of two biotechnology companies. He serves on the National Academies' Intelligence Community Studies Board and the AAAS Committee on Science, Engineering and Public Policy. He co-led the National AI R&D Strategic Plan released by the White House in 2016.

Tasha McCauley is a technology entrepreneur. Her current work with GeoSim Systems centers around a new technology that produces high-resolution, fully interactive virtual models of cities. Prior to her involvement with GeoSim, she co-founded Fellow Robots, a robotics company based at NASA Research Park in Silicon Valley. She was formerly on the faculty of Singularity University, where she taught students about robotics and was Director of the Autodesk Innovation Lab. She sits on the Board of Directors of the Ten to the Ninth Plus Foundation, an organization focused on empowering exponential technological change worldwide.

Nico Miailhe co-founded The Future Society in 2014, and its AI Initiative in 2015. He is the co-Convener of the Global Governance of AI Roundtable (World Government Summit, Dubai), as well as a member of the AI Group of experts at OECD, of the World Bank's Digital Economy for All Initiative, of the Global Council on Extended Intelligence, and of the IEEE Global Initiative on Ethically Aligned Design. Nicolas teaches at the Paris School of International Affairs (Sciences Po), at the IE School of Global and Public Affairs in Madrid, and at the Mohamed bin Rashid School of Government in Dubai. He is a Senior Visiting Research Fellow with the Program on Science, Technology and Society at Harvard.

James Miller is a professor of economics at Smith College. He has a PhD from the University of Chicago and a J.D. from Stanford. He is the author of *Singularity Rising*, *Game Theory at Work*, and a microeconomics textbook. He has a podcast called Future Strategist and has written for Quillette. His interests include learning what the Fermi paradox teaches us about existential risks and what utility function analysis has to offer AGI theorists. He is a member and adviser to the board of cryonics provider Alcor.

Smitha Milli is a PhD student at UC Berkeley advised by Moritz Hardt and Anca Dragan who works on machine learning and cognitive science. Her recent work has focused on understanding dynamics that occur when humans strategically interact with machine learning systems.

Luis Moniz Pereira is Emeritus Professor of Computer Science, Universidade Nova de Lisboa, Portugal. He is affiliated with its NOVA Laboratory for Computer Science and Informatics (NOVA-LINCS). A fellow of the European Association for Artificial Intelligence (EurAI) and Doctor honoris causa T.U. Dresden, he was the founding president of the Portuguese Artificial Intelligence Association (APPIA). His research focuses on knowledge representation and reasoning, logic programming, cognitive sciences, and evolutionary game theory. More information at http://userweb.fct.unl.pt/~lmp/.

Richard Ngo is a research engineer on the AGI safety team at DeepMind, primarily working on reward modeling from human preferences. He is particularly interested in building reward models which remain predictable when scaled up to harder problems. He has also analysed approaches to AI safety at the Future of Humanity Institute. He graduated with distinctions from Oxford and Cambridge.

Jeremy Nixon is a Research Software Engineer at Google Brain, dedicated to understanding the principles underlying information and intelligence.

Seán Ó Héigeartaigh is the Executive Director of the Centre for the Study of Existential Risk (CSER) at the University of Cambridge. He is also Director of the AI: Futures and Responsibility Programme at the Centre for the Future of Intelligence, where he leads a team of researchers focused on foresight, governance and international community-building relating to the near- and long-term impacts of artificial intelligence.

Catherine Olsson is a Research Software Engineer at Google Brain working on ML robustness. Catherine graduated in Computer Science and Brain & Cognitive Science from MIT, and completed a Master's degree in Neuroscience at NYU.

David Orban is the Founder and Managing Partner of Network Society Ventures, a global investment firm focused on innovative ventures at the intersection of exponential technologies and decentralized networks. David is also Founder and a Trustee of Network Society Research, a London-based global think tank, whose aim is to allow individuals, enterprises and the society at large to deal positively with the unstoppable transformation brought by exponential technologies and decentralization. He is also a mentor for the Thiel Fellowship, a Scientific Advisory Board Member for the Lifeboat Foundation, an Advisor to Humanity+, a Founder of the Open Government Data working group, and an Advisor to the Institute of Ethics and Emergent Technologies and numerous startups in Europe and North America. David is the author of Something New, about the role of artificial intelligence in society.

Lucas Perry is Project Coordinator for the Future of Life Institute, where he focuses on existential risk mitigation efforts. He organized the Beneficial AI 2017 conference, has worked on AI safety grant making, hosts the AI Alignment Podcast, developed a conceptual landscape of the AI alignment problem, and participated in the UN negotiations for a nuclear weapons ban. Lucas has a background in philosophy and is passionate about the role technology will play in the evolution of all sentient life.

Marie-Therese Png is a doctoral candidate at the Oxford Internet Institute, and PhD research intern at DeepMind Ethics and Society. Her research focuses are Globally Beneficial AI, Intercultural AI Ethics, and Global Justice. Previously, Marie-Therese was Research Affiliate at the MIT Media Lab, where she founded Implikit.org, a neurotechnology project addressing implicit bias, and coordinated the community focused global biohacking movement at the MIT BioSummit. She was AI Policy Research Associate at the Harvard Artificial Intelligence Initiative, building the Global Civic Debate on AI, and the World Government Summit AI Roundtable. Marie-Therese holds a Master's from Harvard in Developmental Cognition and Intergroup conflict.

Tomaso Poggio is the Eugene McDermott Professor in the Dept. of Brain & Cognitive Sciences at MIT and the director of the NSF Center for Brains, Minds and Machines at MIT. He is a member of the Computer Science and Artificial Intelligence Laboratory and of the McGovern Brain Institute. He received the Laurea Honoris Causa from the University of Pavia for the Volta Bicentennial, the 2003 Gabor Award, the Okawa Prize 2009, the American Association for the Advancement of Science (AAAS) Fellowship, and the 2014 Swartz Prize for Theoretical and Computational Neuroscience. A former Corporate Fellow of Thinking Machines Corporation and a former director of PHZ Capital Partners, Inc., and Mobileye, he was involved in starting, or investing in, several other high tech companies including Arris Pharmaceutical, nFX, Imagen, Digital Persona and DeepMind.

Neil Rabinowitz is leads a research group at DeepMind focused on teasing apart how trained AI systems are solving their problems. His research interests primarily centre around asking what kinds of intelligence AI systems embody, building descriptions of how they behave, what their biases are, how these arise, and how we can interface with their internal computations from our human perspective. Neil has had diverse training in mathematics, physics, the philosophy of science, neuroscience, and machine learning. He received his DPhil from the University of Oxford, and pursued postdoctoral research at New York University, where he focused on how mammalian brains solve some of their natural computational challenges.

Peter Railton is the Kavka Distinguished University Professor of Philosophy at the University of Michigan. His main research has been in ethics, meta-ethics, moral psychology, and the philosophy of science. He has a special interest in the bearing of empirical research on the nature of action and value. Among his writings are *Facts, Values, and Norms* (Cambridge, 2003) and Homo Prospectus (co-authored, Oxford, 2016). He has been a visiting faculty member at Berkeley and Princeton, President of the American Philosophical Association (Central Division), a fellow of the Guggenheim Foundation, the ACLS, and the NEH. In 2003 he was elected to the American Academy of Arts and Sciences.

Tobias Rees is Reid Hoffman Professor of Humanities at the New School for Social Research; Director of the Transformations of the Human program at the Berggruen Institute; and a Fellow at the Canadian Institute for Advanced Research. He holds degrees in philosophy, anthropology, and neurobiology. As he sees it, AI is not only an engineering field but also a most far-reaching philosophical event –– one that cannot be framed (tamed) with our established ways of thinking + doing.

Francesca Rossi is the IBM AI Ethics Global Leader and a Distinguished Research Staff Member at IBM Research. Her research interests focus on AI, including constraint reasoning, preferences, multi-agent systems, collective decision making, bias, and value alignment. She is an AAAI and a EurAI fellow, she has been president of IJCAI and an executive councillor of AAAI. She is Editor in Chief of the Journal of AI Research. She is a member of the scientific advisory board of the Future of Life Institute and a deputy academic director of the Leverhulme Centre for the Future of Intelligence. She is a member of the board of directors of the Partnership on AI and of the European Commission high level expert group on AI.

Jonathan Rothberg is best known for inventing high-speed, "Next-Gen" DNA sequencing and was awarded the National Medal of Technology by president Obama for this innovation. Jonathan brought to market the first new method for sequencing genomes since Sanger and Gilbert won the Nobel Prize in 1980. He sequenced the first individual human genome (Watson Genome), initiated the Neanderthal Genome Project with Svante Paabo, and with the sequencing of Gordon Moore paved the way to the sub $1,000 genome. Under his leadership his team helped understand the mystery behind the disappearance of the honey bee, uncovered a new virus killing transplant patients, and elucidated the extent of human variation—work recognized by Science magazine as the breakthroughs of the year for 2006 & 2007. He currently runs 4catalyzer, a medical device incubator, and is an adjunct professor at Yale.

Stuart Russell is Professor of Computer Science at Berkeley and director of the Center for Human- Compatible Artificial Intelligence. He is also an Honorary Fellow of Wadham College, Oxford. He is a co-author (with Peter Norvig) of the standard textbook, *Artificial Intelligence: a Modern Approach*. He is a recipient of the NSF Presidential Young Investigator Award, the IJCAI Computers and Thought Award, the AAAI Feigenbaum Prize, the Mitchell Prize of the American Statistical Association, and Outstanding Educator Awards from both ACM and AAAI. From 2012 to 2014 he held the Chaire Blaise Pascal in Paris. He is a Fellow of AAAI, ACM, and AAAS.

Dorsa Sadigh is an Assistant Professor in the Computer Science Department and Electrical Engineering Department at Stanford University. Her work is focused on the design of algorithms for autonomous systems that safely and reliably interact with people.

Anna Salamon is the co-founder and president of the Center for Applied Rationality (CFAR). She has previously done machine learning research for NASA and applied mathematics research on the statistics of phage metagenomics. She holds a degree in mathematics from UC Santa Barbara.

David Sanford is Chief of Staff, Office of Reid Hoffman at Linkedin Inc, and he is also an Advisory Council Member at New America California. David received his BA in Entrepreneurial Management from Stanford University.

Will Saunders is a PhD Student in Machine Learning and AI Safety at the University of Toronto, working on projects related to Iterated Distillation and Amplification. Prior to that, he interned at the Future of Humanity Institute working on safe reinforcement learning, and worked as a software engineer at Google.

John Schulman is a research scientist at OpenAI, where he works on reinforcement learning (RL), especially related to transfer learning and meta-learning. He leads the Games Team, where they (mostly) use games as a testbed for reinforcement learning. He co-developed some of the most widely used algorithms in reinforcement learning: Trust Region Policy Optimization and Proximal Policy Optimization; and some of the most widely used software: OpenAI Gym and Baselines. Previously, John received his PhD in Computer Science from UC Berkeley, where he was advised by Pieter Abbeel, working on robotics and reinforcement learning.

Bart Selman is the Professor of Engineering and Computer Science at Cornell University. He is the President-Elect of the Association for the Advancement of Artificial Intelligence (AAAI) and the co-Chair of a national study to determine the Roadmap for AI research to guide US government research investments in AI. Bart was previously at AT&T Bell Laboratories. His research interests include artificial intelligence, computational sustainability, efficient reasoning procedures, machine learning, deep learning, reinforcement learning, planning, knowledge representation, and connections between computer science and statistical physics. He has (co-)authored over 150 publications, including six best paper awards and two classic paper awards.

Andrew Serazin is President of Templeton World Charity foundation, where he is building interdisciplinary teams of scientists, technologists, and humanities scholars to pursue big questions of human purpose, the natural world, and ultimate reality. As a malaria researcher at Oxford and Notre Dame, as well as an executive at the Bill & Melinda Gates foundation in Seattle, he worked to harness insights from unconventional thinkers to advance solutions for nutrition, maternal and child health, and infectious diseases. He also founded Matatu, a venture-backed biotechnology company, to demonstrate the commercial and societal value of the vast community of beneficial bacteria.

Rohin Shah is a 5th year PhD student at UC Berkeley with the Center for Human-Compatible AI, working with Anca Dragan, Pieter Abbeel and Stuart Russell. His research interests have been changing a lot, but he is overall aimed at the problem of creating beneficial AI. Every week, he collects and summarizes recent progress relevant to AI alignment in the Alignment Newsletter. He has also been involved with effective altruism for several years.

Carl Shulman is a Research Associate at the Future of Humanity Institute, Oxford Martin School, Oxford University, where his work focuses on the long-run impacts of artificial intelligence and biotechnology. He is also an Advisor to the Open Philanthropy Project. Previously, he was a Research Fellow at the Machine Intelligence Research Institute and held positions at Clarium Capital Management and Reed Smith LLP. He attended New York University School of Law and holds a degree in philosophy from Harvard University.

Xiao (Jason) Si is currently dean of Tencent Research Institute, and deputy general counsel of Tencent. He is a member of Information Society 50 Forum, and a senior research fellow of International Copyright Research Center in NCAC. He is also a visiting scholar of Stanford Law School, president of the Shenzhen Copyright Association, and an advisor to postgraduates at Peking University Law School. Dr. Jason Si received his Ph.D. from Zhongnan University of Economics and Law. He has dedicated much of his academic works to studying legal and public issues associated with the Chinese internet industry and published dozens of articles accordingly. Before joining Tencent, he held key positions in Netease, Xunlei, and other renowned Chinese internet companies.

Tanya Singh has been the Executive Assistant to Prof. Nick Bostrom for over a year, and has recently taken on the role of Head of Operations at the Future of Humanity Institute. Tanya has worked across the domains of HR consulting, business development and data analytics (running a monthly P&L of up to $100 million) for an Indian e-commerce company, market research and user growth for Khan Academy, and as the product head of a FinTech company. Tanya has been profoundly influenced by the research done at the Future of Humanity Institute and is convinced that her work here is her most effective contribution to society.

Nate Soares is the executive director of the Machine Intelligence Research Institute (MIRI). He first joined MIRI in 2014 as a Research Fellow, quickly earning a strong reputation for his insight and his work ethic. Soares is the primary author of most of MIRI's technical agenda, including the overview document Agent Foundations for Aligning Superintelligence with Human Interests (2014) and the Association for the Advancement of Artificial Intelligence (AAAI) paper Corrigibility (2015). Prior to MIRI, Soares worked as a software engineer at Google.

Bing Song is Vice President of the Berggruen Institute and Director of the Institute's Beijing based China Center. She overseas research programs, institutional development of the China Center as well as China-based fellowship programs. Prior to joining the Berggruen Institute, Bing was a senior executive of Goldman Sachs' China business. Earlier in her career, Bing was a practicing lawyer for many years. She also undertook academic and policy research and published in the areas of administrative law, competition law and comparative procedural laws.

Charlotte Stix is a Research Associate at the Leverhulme Centre for the Future of Intelligence, University of Cambridge. She further advises Element AI on European AI Policy and is a fellow of the World Economic Forum's Global Future Council on Neurotechnologies and Brain Sciences. She was formerly at the European Commission's Robotics and Artificial Intelligence Unit, where she oversaw a budget of €18 million in projects and contributed to the formulation of EU-wide AI strategy. Previously, she was a Policy Officer at the World Future Council, and before that founded an award winning culture magazine while managing a team of 15 freelancers. She holds a MSc in Philosophy with a focus on the ethics of enhancement, as well as additional degrees in philosophy, fashion design, and the performing arts.

Hang Su is an assistant professor in the Department of Computer Science and Technology at Tsinghua University. Currently, he is serving as the Director Assistant of the Tsinghua Institute for Artificial Intelligence. His research interests lie in the fundamental theory of artificial intelligence with special focus on the interpretable and robust AI. Prof. Su has published more than 50 papers including the top tier conferences of CVPR, ECCV, TMI, etc. He has served as senior PC or PC members in the dominant international conferences including IJCAI, AAAI, CVPR, and worked as reviewers for top journals such as TPAMI, TIP, TMI, etc. Current, he also served as TCs in academic societies, including CV in CCF, ML in CAAI. He received "Young Investigator Award" from MICCAI2012, the "Best Paper Award" in AVSS2012, and "Platinum Best Paper Award" in ICME2018. More information is available at http://www.suhangss.me.

Jaan Tallinn is a founding engineer of Skype and Kazaa. He is a co-founder of the Cambridge Centre for the Study of Existential Risk (cser.org), Future of Life Institute (futureoflife.org), and philanthropically supports other existential risk research organisations. Jaan is on the Board of Sponsors of the Bulletin of the Atomic Scientists (thebulletin.org), member of the High-Level Expert Group on AI at the European Commission, and has served on the Estonian President's Academic Advisory Board. He is also an active angel investor, a partner at Ambient Sound Investments (asi.ee), and a former investor in and director of the AI company DeepMind (deepmind.com).

Alexander Tamas is a founding partner of Vy Capital where he invests in frontier technology companies ranging from Quantum Computing, AI Hardware, Space Exploration, Blockchain Technology to Longevity and others. Prior to Vy, Alexander was a partner at Digital Sky Technologies (DST) and a Board Member and Managing Director of Mail.Ru, one of Russia's largest Internet platforms. At DST, Alexander led primary investment rounds into companies such as Facebook, Spotify, and Alibaba. Prior, he worked in the Tech Investment Banking Division of Goldman Sachs. Alexander is co-founder of many companies and philanthropic initiatives and a long time supporter of the AI safety program at the Future of Humanity Institute.

Max Tegmark is a professor doing AI and physics research at MIT, and advocates for positive use of technology as president of the Future of Life Institute. He is the author of over 200 publications as well as the books "Life 3.0: Being Human in the Age of Artificial Intelligence" and "Our Mathematical Universe: My Quest for the Ultimate Nature of Reality".

Josh Tenenbaum is a professor of Computational Cognitive Science at MIT. His research focuses on one of the most basic and distinctively human aspects of cognition: the ability to learn so much about the world, rapidly and flexibly. His current passion is to understand the nature and origins of common sense: What makes any human toddler more intelligent than any machine ever built? Tenenbaum also works actively in artificial intelligence, believing that if we can build machines that learn, see, think and act in more human-like ways, this will lead to more useful and beneficial AI systems as well as more powerful tools for understanding the human mind.

Helen Toner spent most of 2018 in Beijing, combining intensive Mandarin Chinese language study with research into the Chinese AI ecosystem as a Research Affiliate of Oxford University's Center for the Governance of AI, as well as spending two months over summer as an AI Policy Associate at OpenAI. Previously, Helen was a Senior Research Analyst at the Open Philanthropy Project, where she focused on policy and strategy issues related to progress in machine learning. Helen was a lead co-author on the seminal report The Malicious Use of AI, and has also been published in Foreign Affairs and the People's Daily.

Brian Tse is a Policy Affiliate at the Center for the Governance of AI at the University of Oxford. He wrote the Chinese version of the OpenAI Charter and presented his research on international cooperation at DeepMind. He has consulted the Open Philanthropy Project on their multi-million-dollar grants and is leading the effective altruism community in the Chinese-speaking world. Brian has worked at a leading Chinese AI hardware startup and as an investment banker at J.P. Morgan. A fluent speaker of three Chinese languages, Brian has studied at Harvard University, Tsinghua University, and the University of Hong Kong.

Andrey Ustyuzhanin is Head of Laboratory at the National Research University, Higher School of Economics, in Moscow. He is a researcher, data scientist, and software developer, and is particularly interested in finding applications of scientific approaches from distributed and heterogeneous system design, machine learning, statistics to various areas including particle physics, bioinformatics, robotics. At the National Research University, he organizes seminars & summer schools (http://www.hse.ru/mlhep2015/) focused on bridging the gap between data science and particular fields of fundamental science.

Jonathan Uesato is a research engineer at DeepMind. His primary interest is in ensuring deployed machine learning systems perform well on average, without ever performing too poorly even in the worst case. This has led him to research topics such as adversarial examples, scalable verification techniques for neural networks, and adversarial testing in reinforcement learning. Prior to joining DeepMind, he worked at Microsoft Research on combining statistical and symbolic approaches for neural program synthesis, and at Cruise Automation, on control algorithms for self-driving cars.

Carroll Wainwright is a Visiting Research Fellow at the Partnership on AI, and cofounder of Metaculus.

Wendell Wallach is a scholar, consultant, and author at Yale University's Interdisciplinary Center for Bioethics, where he chairs Technology and Ethics studies. He is also a senior adviser to The Hastings Center. His books include, A Dangerous Master: How to keep technology from slipping beyond our control, a primer on emerging technologies, and Moral Machines: Teaching Robots Right From Wrong (co-authored with Colin Allen), which mapped the then new field of enquiry called machine ethics or machine morality. He received the World Technology Award for Ethics in 2014 and for Journalism and Media in 2015, as well as a Fulbright Research Chair at the University of Ottawa for 2015-2016. The World Economic Forum appointed Mr. Wallach co-chair of its Global Future Council on Technology, Values, and Policy for the 2016-2018 term.

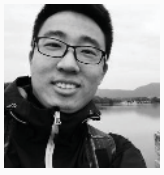
Adrian Weller is Programme Director for AI at The Alan Turing Institute, the UK national institute for data science and AI, where he is also a Turing Fellow leading a group on Fairness, Transparency and Privacy. He is a Senior Research Fellow in Machine Learning at the University of Cambridge, and at the Leverhulme Centre for the Future of Intelligence where he leads the project on Trust and Transparency. He is very interested in all aspects of AI, its commercial applications and helping to ensure beneficial outcomes for society. He serves on several boards including the Centre for Data Ethics and Innovation. Previously, Adrian held senior roles in finance.

Hiroshi Yamakawa is the Director of Dwango AI Laboratory, a Director and Vice Chief Editor of the Japanese Society for Artificial Intelligence, and a Fellow Researcher at the Brain Science Institute at Tamagawa University. He is specialized in AI, in particular, cognitive architecture, concept acquisition, neuro-computing, opinion collection. He is one of the founders of the Whole Brain Architecture Seminar and the SIG AGI in Japan.

Roman Yampolskiy is a tenured faculty member in the department of Computer Engineering and Computer Science at the Speed School of Engineering, University of Louisville. He is the founding and current director of the Cyber Security Lab and an author/editor of many books including Artificial Intelligence Safety and Security. Yampolskiy's research focuses on AI safety and cybersecurity. He tweets as @romanyam

Yi Zeng is Professor and Deputy Director at Research Center for Brain-inspired Intelligence, Institute of Automation, Chinese Academy of Sciences. He is a Professor at University of Chinese Academy of Sciences, and is a Fellow of the Berggruen Institute. His research interests focus on Brain-inspired Artificial General Intelligence, including Brain-inspired Cognitive Modeling, Brain-inspired Neural Networks, Brain-inspired Robotics, Philosophy and Ethics of Artificial Intelligence. He is devoted to realize Beneficial Brain-inspired Conscious Living Becomings. His recent work include Brain-inspired Cognitive Engine, Linking Artificial Intelligence Principles, and Harmonious Artificial Intelligence Principles.

Alex Zhu does technical AI safety research on Paul Christiano's research agenda and MIRI's research agenda. He placed first in the USA Math Olympiad in 2012 and 15th on the 2012 Putnam Mathematical Competition. He dropped out of MIT in 2015 to co-found AlphaSheets, a venture-backed startup he co-ran for 2 years, which he left in 2017 to do direct work on AI safety.



Xiaohu (Neil) Zhu is the Founder and Chief Scientist of University AI, an organization providing AI education and training for individuals and big companies in China. He is a consultant of National Engineering Lab for Deep Learning in China. He got a master degree on AI in Nanjing University with a background on algorithmic game theory, mathematical logic, complex networks, deep learning, and reinforcement learning. He started his general research on AGI/AI safety in 2016 and now focuses on mechanism design/value modeling. He has translated several articles on AGI/AI safety (from DeepMind, FHI, and OpenAI) and many articles on Deep Learning/Reinforcement Learning/GANs, and books like Neural Networks and Deep Learning and Learning TensorFlow.

Daniel Ziegler works on the OpenAI Safety Team as a research engineer, prototyping techniques for building powerful ML systems aligned with human values. In a past life, he worked on formal verification to make provably bug-free software.

Shivon Zilis is a project director at Neuralink and Tesla and an advisor to OpenAI. Prior to that she was a partner and founding member of Bloomberg Beta where she spent 6 years investing in and working with machine intelligence startups, non-profits, and incubators.