





**The Ethics of Value Alignment 2017**  
**Long Beach, CA**

	<p><b>Susan Anderson</b> is Professor Emerita of Philosophy at the University of Connecticut. Her specialty is applied ethics, most recently focusing on biomedical ethics and machine ethics. She has received funding from the National Endowment for the Humanities and, with Michael Anderson, from NASA and the NSF. She is the author of three books in the Wadsworth Philosophers Series, as well as numerous articles. Together with Michael Anderson, her research in machine ethics was selected for Innovative Applications of Artificial Intelligence as an emerging application in 2006. Scientific American (Oct. 2010) features an invited article on their research in which the first robot whose behavior is guided by an ethical principle is debuted.</p>
	<p><b>Michael Anderson</b> is a professor of computer science at the University of Hartford. His interest in further enabling machine autonomy brought him first to diagrammatic reasoning where he co-chaired Diagrams 2000, the first conference on the topic, held at the University of Edinburgh, co-edited the book Diagrammatic Representation and Reasoning, and published an article on the topic in Artificial Intelligence, the premier journal in the field of AI. This interest has currently led him, in conjunction with Susan Leigh Anderson, to establish machine ethics as a bona fide field of study. They co-chaired the AAAI Fall 2005 Symposium on Machine Ethics, co-edited an IEEE Intelligent Systems special issue on machine ethics, and co-authored an invited article on the topic for Artificial Intelligence Magazine. They have published "Machine Ethics" with Cambridge University Press (2011). His research has been funded by the NSF and NASA.</p>
	<p><b>Brent Barron</b> is Director of Public Policy at CIFAR (the Canadian Institute for Advanced Research) where he is responsible for building relationships with CIFAR's government partners and engaging the policy community around cutting edge science. He played an important role in the development of the Pan-Canadian Artificial Intelligence Strategy, and now oversees CIFAR's AI &amp; Society program, examining the social, ethical, legal, and economic effects of AI. Prior to this role, Brent held a variety of positions in the Ontario Public Service, most recently in the Ministry of Research, Innovation and Science. Brent holds a Master's in Public Policy from the University of Toronto, as well as a Bachelor's in Media Studies from Western University</p>
	<p><b>Nick Bostrom</b> is Professor at Oxford University, where he is the founding Director of the Future of Humanity Institute. He also directs the Strategic Artificial Intelligence Research Center. He is the author of some 200 publications, including <i>Anthropic Bias</i> (Routledge, 2002), <i>Global Catastrophic Risks</i> (ed., OUP, 2008), <i>Human Enhancement</i> (ed., OUP, 2009), and <i>Superintelligence: Paths, Dangers, Strategies</i> (OUP, 2014), a New York Times bestseller. He is best known for his pioneering work in five areas: (i) existential risk; (ii) the simulation argument; (iii) anthropics; (iv) impact of future technology, especially AI; and (v) macrostrategy (links between long-term outcomes and present actions).</p>
	<p><b>Meia Chita-Tegmark</b> is a Ph.D. candidate in Developmental Sciences at Boston University and an alumna of the Harvard Graduate School of Education. She conducts research in the Social Development and Learning Lab at Boston University. Meia is interested in a variety of topics in developmental psychology, such as atypical social development, attention mechanisms and</p>

	<p>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.</p>
	<p><b>Joshua Cohen</b> is a political theorist, trained in philosophy, with a special interest in issues that lie at the intersection of democratic norms and institutions. He has written extensively on issues of democratic theory, particularly deliberative democracy and its implications for personal liberty, freedom of expression, religious freedom, and political equality. He has also written on issues of global justice, including the foundations of human rights, distributive fairness, supranational democratic governance, and labor standards in supply chains. Cohen serves as co-editor of <i>Boston Review</i>, a bimonthly magazine of political, cultural, and literary ideas. He has published <i>Philosophy, Politics, Democracy</i> (Harvard University Press, 2009); <i>Rousseau: A Free Community of Equals</i> (Oxford University Press, 2010); <i>The Arc of the Moral Universe and Other Essays</i> (Harvard University Press, 2011); and edited (with Alex Byrne, Gideon Rosen, and Seana Shiffrin) <i>The Norton Introduction to Philosophy</i> (forthcoming 2014).</p>
	<p><b>Andrew Critch</b> is a Research Fellow at the Machine Intelligence Research Institute (MIRI), and a visiting postdoc at the UC Berkeley Center for Human Compatible AI. He earned his Ph.D. 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.</p>
	<p><b>Stephanie Dinkins</b> is an artist interested in creating platforms for ongoing dialogue about artificial intelligence as it intersects race, gender, aging, and our future histories. She is particularly driven to work with communities of color to develop deep-rooted AI literacy and co-create more culturally inclusive equitable artificial intelligence. Her work often employs lens-based practices, the manipulation of space, and technology to grapple with notions of consciousness, agency, perception, and social equity. Stephanie's art is exhibited internationally at a broad spectrum of community, private, and institutional venues – by design. Stephanie is a 2017 A Blade of Grass Fellow and a 2018 Truth Resident at EYEBEAM, NY. Artist residencies include NEW INC, Blue Mountain Center; Aim Program, Bronx Museum; The Laundromat Project; Santa Fe Art Institute, Art/Omi and Center for Contemporary Art, Czech Republic. Her work has been cited in media outlets such as Art In America, The New York Times, Washington Post, and Baltimore Sun, and SLEEK Magazine.</p>
	<p><b>Anca Dragan</b> 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.</p>



**Owain Evans** leads a collaborative project on “Inferring Human Preferences” with Andreas Stuhlmuehler, Jessica Taylor and Noah Goodman. He is working on techniques for learning about human beliefs, preferences and values from observing human behavior or interacting with humans. This draws on machine learning (e.g. inverse reinforcement learning and probabilistic programming), cognitive science, and analytic philosophy.



**Nathanael Fast** is an Associate Professor of Management and Organization at the University of Southern California. His research focuses on the psychological, technological, and social tools people use to lead and influence others. His work examines the determinants and consequences of power and status hierarchies in groups and organizations as well as the social psychological mechanisms that lead people, ideas, and practices to become and stay prominent. He also studies the psychology of social networks and how new and emerging technologies are shaping the future of humanity.



**Nathan Gardels** is the editor-in-chief of The WorldPost and a senior adviser to the Berggruen Institute. He has been editor of New Perspectives Quarterly since it began publishing in 1985. He has served as editor of Global Viewpoint and Nobel Laureates Plus (services of Los Angeles Times (Syndicate/Tribune Media) since 1989. Gardels has written widely for The Wall Street Journal, Los Angeles Times, New York Times, Washington Post, Harper's, U.S. News & World Report, and the New York Review of Books. His books include, *At Century's End: Great Minds Reflect on Our Times* and *The Changing Global Order*. He is coauthor with Hollywood producer Mike Medvay of *American Idol After Iraq: Competing for Hearts and Minds in the Global Media Age*. Since 1986, Nathan has been a Media Fellow of the World Economic Forum (Davos). He has lectured at the Islamic Educational, Scientific, and Cultural Organization (ISESCO) in Rabat, Morocco, and the Chinese Academy of Social Sciences in Beijing, China. He has been a member of the Council of Foreign Relations, as well as the Pacific Council, for many years. Nathan is co-author with Nicolas Berggruen of *Intelligent Governance for the 21st Century*, a Financial Times Book of the Year.



**Nils Gilman** is the Vice President of Programs at the Berggruen Institute where he joined in August 2017. From 2013 to 2017, he served as Associate Chancellor and Chief of Staff to the Chancellor at UC Berkeley, and as the Founding Executive Director of Social Science Matrix, Berkeley's flagship interdisciplinary social science research center. Earlier in this career, he worked as a research director and scenario planning consultant at the Monitor Group and Global Business Network, and in software companies such as Salesforce.com and BEA Systems. He is the author of *Mandarins of the Future: Modernization Theory in Cold War America* (2004), *Deviant Globalization: Black Market Economy in the 21st Century* (2011), as well as numerous articles on intellectual history and political economy. He holds a B.A., M.A., and Ph.D. in History from UC Berkeley.



**Joshua D. Greene** is Professor of Psychology, a member of the Center for Brain Science faculty, and the Director of the Moral Cognition Lab at Harvard University. His research has focused on the psychology and neuroscience of moral judgment and decision-making. His broader interests cluster around the intersection of philosophy, psychology, and neuroscience. He is the author of *Moral Tribes: Emotion, Reason, and the Gap Between Us and Them*.



**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.



**Sam Harris** is a neuroscientist and the author of five New York Times best sellers. His books include *The End of Faith*, *Letter to a Christian Nation*, *The Moral Landscape*, *Free Will*, *Lying*, *Waking Up*, and *Islam and the Future of Tolerance* (with Maajid Nawaz). *The End of Faith* won the 2005 PEN Award for nonfiction. His writing and public lectures cover a wide range of topics—neuroscience, moral philosophy, religion, meditation practice, human violence, rationality—but generally focus on how a growing understanding of ourselves and the world is changing our sense of how we should live. He also regularly hosts a popular podcast, and in September 2016 he released a TED talk titled: Can we build AI without losing control over it?



**John Hepburn** is the Vice President, Research and International at the University of British Columbia (UBC). He is also a Professor in the departments of Chemistry and Physics & Astronomy. John was influential in building the capacity of UBC's research portfolio and creating new institutional partnerships in China and around the world. His previous positions at UBC include Dean of the Faculty of Science and Head of Chemistry. He is also a former Canada Council Killam Research Fellow and a Fellow of the Royal Society of Canada. John has received a number of honours and awards, including the Rutherford Medal in Physics from the Royal Society of Canada, Elected Fellowships in the American Physical Society and the Chemical Institute of Canada, and an Alfred P. Sloan Foundation Fellowship.



**Pamela Hieronymi** is a Professor of Philosophy at UCLA. She conducts research at the intersection of many different subfields: ethics, philosophy of mind, philosophy of action, and the lively discussion of moral responsibility and free will. Her recent work has focused on the agency we exercise over our own attitudes, in particular, over our beliefs and intentions. This interest grew out of her interest in the source and nature of the motives worth having, their justification, and our responsibility for them. Pamela is currently working on a manuscript bringing her recent work to bear on the problems of free will and moral responsibility.



**Viktoriya Krakovna** is a research scientist in AI safety at DeepMind and a co-founder of the Future of Life Institute. Her Ph.D. 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.



**Simon Lacoste-Julien** is an Assistant Professor at the Montreal Institute of Learning Algorithms, at the Université de Montréal. Until August 2016, he was a researcher at INRIA in the SIERRA project team which is part of the Computer Science Department of École Normale Supérieure in Paris. Simon did his PhD in Computer Science at the University of California, Berkeley, and (basically) a B.Sc. Triple Honours in Mathematics, Physics and Computer Science at McGill University. He then worked with Zoubin Ghahramani as a postdoc in the Machine Learning Group of the University of Cambridge.



**Shane Legg** is a machine learning researcher and founder of DeepMind. He is interested in measures of intelligence for machines, neural networks, artificial evolution, reinforcement learning and the theory of learning. He obtained his Ph.D. from IDSIA in Switzerland, and his thesis proposed a formal definition of machine intelligence, for which he was awarded the \$10,000 Canadian Singularity Institute research prize. He spent a postdoctoral year at the Swiss Finance Institute building models of human decision making, followed by two years at the Gatsby Computational Neuroscience Unit at UCL, where he is now an honorary fellow.



**Patrick Lin** is the Director of the Ethics + Emerging Sciences Group, based at California Polytechnic State University, San Luis Obispo, where he is an Associate Philosophy Professor. Other current and past affiliations include: Stanford Engineering, Stanford Law, U.S. Naval Academy, Dartmouth College, Notre Dame, World Economic Forum, and UNIDIR. He is well published in technology ethics, especially on robotics and AI—including the books *Robot Ethics* (MIT Press, 2012) and *Robot Ethics 2.0* (Oxford University Press, forthcoming in 2017)—as well as cyberwar/security, nanotechnology, human enhancement, space exploration, and other areas. He regularly gives invited briefings to industry, media, and government; and he teaches courses in ethics, political philosophy, philosophy of technology, and philosophy of law.



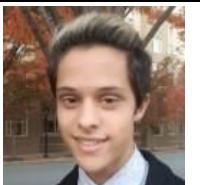
**Richard Mallah** is Director of AI Projects at the Future of Life Institute, where he works to support the robust, safe, beneficent development of advanced AI via meta-research, analysis, research organization, and advocacy. Mallah serves on the Executive Committee of the Institute of Electrical and Electronics Engineers’ initiative on autonomous systems ethics, and chairs associated working groups. Richard serves as a Senior Advisor to both Cambridge Semantics, Inc., where he led creation of the industry’s highest-rated enterprise text analytics system, and to the AI Initiative of the Future Society at the Harvard Kennedy School.



**Thomas K. Metzinger** is full Professor and Director of the theoretical philosophy group and the research group on neuroethics/neurophilosophy at the department of philosophy, Johannes Gutenberg University of Mainz, Germany. He is a 2014-2019 Fellow at the Gutenberg Research College. He is the Founder and Director of the MIND group and Adjunct Fellow at the Frankfurt Institute of Advanced Studies, Germany. His research centers on analytic philosophy of mind, applied ethics, philosophy of cognitive science, and philosophy of mind.



**David Pearce** is a British utilitarian philosopher and transhumanist, who promotes the idea that there exists a strong ethical imperative for humans to work towards the abolition of suffering in all sentient life. His internet manifesto *The Hedonistic Imperative* details how he believes the abolition of suffering can be accomplished through “paradise engineering.” He co-founded the World Transhumanist Association in 1998, and the Abolitionist Society in 2002.



**Lucas Perry** is Project Coordinator for the Future of Life Institute. Lucas is passionate about the role that science and technology play in the evolution of all sentient life. He has a background in philosophy, and has studied at a Buddhist monastery in Nepal and engaged in a range of meditative retreats and practices, which work to inform his study of effective altruism and existential risks.



**Tenzin Priyadarshi** is the Founding Director of the Dalai Lama Center for Ethics and Transformative Values at MIT. He received his undergraduate degree (summa cum laude) and has a graduate degree in Comparative Philosophy of Religion from Harvard University where he was an Integral Honors Scholar (studying Philosophy and Physics). At the age of ten, he entered a Buddhist monastery in Rajgir near ancient Nalanda University and was subsequently ordained by His Holiness the Dalai Lama who is his spiritual mentor. He has been interviewed by NPR and articles on him and his work have appeared in the New York Times and the Boston Globe. He has given a number of talks at the American Academy of Arts and Sciences and various institutes of learning. Tenzin lectures internationally and is also President of The Prajnopaya Foundation, a worldwide humanitarian organization.



**Tobias Rees** is Assistant Professor of Anthropology with a dual appointment in the Departments of Social Studies of Medicine and Anthropology. Prior to joining the McGill Faculty, he held positions at the Universities of Freiburg (Germany) and Zurich (Switzerland). Tobias' expertise lies at the intersection of anthropology, art history, history of science, and the philosophy of modernity and concerns the critical study of knowledge/thinking. More specifically, he is interested in how categories that order knowledge mutate over time (or differ across space) – and in what effects these mutations have on conceptions of the human. The main areas of Tobias' research have been emergent phenomena in the life sciences and in medicine, with a particular focus on neurobiology/neuropharmacology.



**Francesca Rossi** is a research scientist at the IBM T.J. Watson Research Centre, and a Professor of computer science at the University of Padova, Italy, where she is currently on leave. Her research interests focus on AI, specifically on constraint reasoning, preferences, multi-agent systems, computational social choice, and collective decision-making. She is also interested in ethical issues in the development and behaviour of AI systems. She has published over 170 scientific articles in journals and conference proceedings, and as book chapters. She is an AAAI and a EurAI fellow, and a 2015 Radcliffe fellow. She has been President of IJCAI, an Executive Councillor of AAAI, and is Editor-in-Chief of JAIR. She co-chairs the AAAI committee on AI and ethics and is a member of the scientific advisory board of the Future of Life Institute. She is on the executive committee of the IEEE global initiative on ethical considerations on the development of autonomous and intelligent systems and she belongs to the World Economic Forum Global Council on AI and robotics.



**Bing Song** is the Director of the Berggruen Institute China Center.. The China Center focuses on bringing out Chinese voices on certain key issues which impact our world and the future of humanity, such as AI and gene-editing and the ethics, as well as new concepts and framework of the world order and new world institutions China has brought and will continue to bring to international geopolitics. Prior to joining the Berggruen Institute, Bing was a Senior Executive with Goldman Sachs China for over a decade. Prior to Goldman, Bing was a practicing lawyer in capital markets transactions. Earlier in her career, she worked for the Ford Foundation on judicial reforms in China and published extensively on comparative law.



**Max Tegmark** is a Professor of Physics at MIT, President of the Future of Life Institute, and Scientific Director of the Foundational Questions Institute. His research has ranged from cosmology to the physics of cognitive systems, and is currently focused at the interface between physics, AI and neuroscience. He is the author of over 200 publications and the book *Our Mathematical Universe: My Quest for the Ultimate Nature of Reality*. His work with the Sloan Digital Sky Survey on galaxy clustering shared the first prize in Science Magazine's "Breakthrough of the Year: 2003".



**Denis Therien** leads CIFAR's Partnerships Group with the goal of developing partnerships between CIFAR and other leading research organizations around the world. Prior to joining the Institute in 2011, he served as Vice President, Research and International Relations at McGill University from 2005 to 2010. A member of the McGill Computer Science faculty since 1978 and a tenured professor since 1991, he served as Director of the McGill School of Computer Science from 1997 to 2005. Among other distinctions, he received, in 2000, a Forschungspreise (Research Award) from Alexander van Humboldt Foundation, and in 2002 was named James McGill Professor at McGill University. He has co-authored over 100 publications on computational complexity theory.



**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.