

Effective Altruism Global
2015

State of AI Safety 2015

In 2014, unprecedented attention was brought to the topic of safety in long-term AI development. Following an open letter by Stuart Russell, Max Tegmark, Frank Wilczek, and Stephen Hawking, Nick Bostrom's book *Superintelligence* spurred academic and public discussion about potential risks from the eventual development of superintelligent machines. In a 2014 survey, academic AI researchers gave a median estimate of 2075 for the development of AI systems able to perform most tasks at human-equivalent levels, a median estimate of a 75% chance that superintelligent machines would become possible within the following 30 years, and a mean estimate of an 18% chance that the outcome would be "extremely bad" for humanity. While the difficulty of such predictions makes a survey far from conclusive, it does reflect growing interest in AI safety, and indicates that it is not too early to begin work.

In the year since, there have been several notable developments. The Future of Life Institute's conference "The Future of AI: Opportunities and Challenges", held in January in Puerto Rico, brought together about 80 top AI experts to discuss AI safety, and resulted in an open letter and research priorities document on keeping AI development safe and beneficial in the short and long term. Elon Musk pledged $10M in grant funding for projects addressing parts of this agenda, and was joined by the Open Philanthropy Project in awarding about $7M to the winners of the first call for proposals, which attracted 300 applicants from academic and non-profit research groups. Eric Horvitz and Russ Altman launched the One Hundred Year Study on Artificial Intelligence at Stanford, and Cambridge University is seeking collaborators and funding for the Turing Project, a global collaboration on AI safety headed by Stuart Russell, Nick Bostrom, Huw Price, and Murray Shanahan. The Machine Intelligence Research Institute, the oldest research group in this area, is expanding its program and has become significantly more coordinated with academia, adding Bart Selman and Stuart Russell to its advisory board.

Given increasing activity and interest in long-term AI safety, it is a good time to take stock of the broader strategic situation. What are the most important considerations for initiating a new field of research, or for engaging in outreach to the public or to governments? How can we best move toward a better understanding of the potential risks of AI development and chart a course to maximize positive outcomes? At this meeting, we hope to make some progress on these questions.

Agenda

11:30 **Doors open,** registration –
Burbank Room, Google Bldg QD7, 369 N Whisman Rd, Mountain View

12:10 **Introduction**
Niel Bowerman

12:20 **Establishing safety as a core topic in the field of AI**
Stuart Russell

- *How can safety be established as a core topic in the field of AI?*
- *How should AI safety be fit into existing journals and conferences?*
- *What role should industry and professional associations play?*

12:30 **Technical research we can do today**
Nate Soares

- *What sort of practical research can usefully be done in advance?*
- *What sort of theoretical research can usefully be done in advance?*
- *Which research communities need to be involved?*
- *How can we build bridges to, and between, those communities?*

12:40 **AI policy: what and when?**
Sebastian Farquhar

- *What types of AI governance do we want?*
- *At which points in time will those governance options be possible/productive?*
- *What steps should we take in the next year to get on the right political trajectory?*

12:50 **Desirable development dynamics**
Nick Bostrom

- *Which values should AI subserve and how should that be determined?*
- *Which development trajectories offer the best odds of achieving such an outcome?*
- *How can we establish trust and cooperation while avoiding problematic tech races?*
- *Which aspects should be open and which should remain restricted or confidential?*
- *How can AI safety folk form strongly win-win relationships with AI developers?*

1:00 - 2:00 **Discussion session 1**

2:00 - 3:00 **Lunch**

3:00 **International coordination**
Mustafa Suleyman

- *What is the current situation of international coordination?*
- *What goals should we be aiming for, and why?*
- *What are the most relevant risks and possibilities in international coordination?*
- *What short-term activities would contribute positively to long-term coordination?*

3:10 **Differential progress toward AI safety**
Paul Christiano

- *What capabilities and challenges are differentially important for beneficial AI?*
- *What available projects will best advance those capabilities?*
- *How can we anticipate what capabilities will be important?*
- *What concrete challenges can guide work on beneficial AI?*

3:20 **What are the key technical uncertainties?**
Daniel Dewey

- *What are they key technical uncertainties in AI safety?*
- *What would it take to form a solid technical understanding of the risks?*
- *To what extent is a solid technical case understanding of risk useful or necessary?*

3:30 **Identifying neglected intervention points**
Owen Cotton-Barratt

- *What are the major factors that plausibly determine AI's impact?*
- *What are the major points of leverage for affecting AI's impact?*
- *When can or should we decide to push off a problem to be solved later?*
- *What methods for positively influencing AI's impact are receiving less attention than they deserve today?*

3:40 - 4:40 **Discussion session 2**

4:40 - 5:00 **Break**

5:00 **Introduction to next steps session**
Niel Bowerman

5:10 **Technical research**
Stuart Russell

- *What are the most important open technical questions?*
- *What existing areas of technical research should be directed at AI safety?*
- *What areas of research should be expanded?*
- *What can be done now, and what later?*

5:20 **Field building**
Bart Selman

- *How can more researchers in this area be created?*
- *How should AI safety be placed in conferences and journals?*
- *How can AI safety best be supported by professional societies?*
- *What funding sources are most suitable for this research?*

5:30 **Cooperation**
Victoria Krakovna

- *What are key areas for coordination?*
- *How do they interact with one another?*
- *What steps can improve communication within / between academia, industry, effective altruism, and government?*
- *What is our message to each of these groups?*

5:40 - 6:40 **Next steps discussion session**

6:50 **Summary of the discussion and conclusion**
Niel Bowerman

7:00 **Dinner** – Google Bldg QD1, 464 Ellis St

Attendees



Alexander Tamas is a partner of Vy Capital, which he co-founded in March 2013. Prior to Vy, Alexander Tamas was partner at DST from 2008 to 2013. Through a series of transactions, he helped to consolidate a number of leading Russian Internet brands under Mail.ru Group and subsequently led the $7bn IPO of the company. Alexander was a board member and managing director of Mail.ru.



Bart Selman is a Professor of Computer Science at Cornell University. He previously was at AT&T Bell Laboratories. His research interests include computational sustainability, efficient reasoning procedures, planning, knowledge representation, and connections between computer science and statistical physics. He has (co-)authored over 100 publications, and has received an NSF Career Award and an Alfred P. Sloan Research Fellowship. He is an AAAI Fellow and a Fellow of the American Association for the Advancement of Science.



Bryan Johnson is an entrepreneur and investor. Bryan launched OS Fund in 2014 with $100 million of his personal capital; his investments include endeavors to cure age-related diseases and radically extend healthy human life to 100+ (Human Longevity), make biology a predictable programming language (Gingko Bioworks & Synthetic Genomics), replicate the human visual cortex using artificial intelligence (Vicarious), and mine an asteroid (Planetary Resources).



Daniel Dewey is the Alexander Tamas Research Fellow on Machine Superintelligence and the Future of AI at the Future of Humanity Institute. He was previously at Google, Intel Labs Pittsburgh, and Carnegie Mellon University.



Dario Amodei is a research scientist at Baidu, where he works with Andrew Ng and a small team of AI scientists and systems engineers to solve hard problems in deep learning and AI, including speech recognition and natural language processing. Dario earned his PhD in physics at Princeton University on Hertz and NDSEG fellowships. His PhD work, which involved statistical mechanics models of neural circuits as well as developing novel devices for intracellular and extracellular recording, was awarded the 2012 Hertz doctoral Thesis Prize.



Elon Musk is the founder, CEO and CTO of SpaceX and co-founder and CEO of Tesla Motors. In recent years, Musk has focused on developing competitive renewable energy and technologies (Tesla, SolarCity), and on taking steps towards making affordable space flight and colonization a future reality (SpaceX). He has spoken about the responsibility of technology leaders to solve global problems and tackle global risks, and has also highlighted the potential risks from advanced AI.



Francesca Rossi is a professor of computer science at the University of Padova, Italy. Currently she is on sabbatical at Harvard University with a Radcliffe fellowship. Her research interests are within artificial intelligence, and include constraint reasoning, preferences, multi-agent systems, and computational social choice. She has been president of the international association for constraint programming (ACP) and she is now the president of International Joint Conference on Artificial Intelligence (IJCAI), as well as the associate editor in chief for JAIR (Journal of AI Reseach).



Gaverick Matheny is IARPA's Director of the Office for Anticipating Surprise and also manages IARPA's OSI, FUSE and ForeST Programs. He previously worked for the Future of Humanity Institute at Oxford University, the World Bank, the Center for Biosecurity, the Center for Global Development, the Applied Physics Laboratory, and on national security projects for the US government. He holds a PhD in Applied Economics from Johns Hopkins University, an MPH from Johns Hopkins, an MBA from Duke University.



Geoff Anders is the founder of Leverage Research, a research institute that studies psychology, cognitive enhancement, scientific methodology, and the impact of technology on society. Geoff is deeply interested in the potential impact of new technologies, especially artificial intelligence.



Howie Lempel leads the Open Philanthropy Project's investigations of potential focus areas related to reducing global catastrophic risks. He has also contributed to its investigations of U.S. policy causes. Howie has a B.A. in Social Studies and Mathematical Economics from Wesleyan University.



John Hering is CEO and co-founder of mobile security company Lookout. BusinessWeek named John a Best Young Tech Entrepreneur in 2010, Fortune named him a Smartest Person in Tech 2011, more recently, the editors of MIT Tech Review dubbed John a "35 under 35" entrepreneur. John is a frequent presenter at mobile and tech industry at events including: RSA, CTIA, Black Hat Technical Security Conference, DEFCON, Fortune Brainstorm. Additionally, John is an angel investor in multiple Silicon Valley startups.



Kerry Vaughan is finishing up a PhD in philosophy specializing in applied ethics and value theory, and is simultaneously working at the Centre of Effective Altruism on expansions of effective altruist outreach and funding. Prior to joining CEA, Kerry worked at the Laura and John Arnold Foundation, where he ascended from a summer volunteer to running the Technology and Innovation Department in less than one year.



Mustafa Suleyman is a British entrepreneur, co-founder of artificial intelligence/machine learning company DeepMind Technologies, acquired by Google. He is also co-founder of Reos Partners, a global conflict resolution firm specialising in addressing complex social challenges.



Nate Soares is the executive director of the Machine Intelligence Research Institute. He joined the MIRI research team in April 2014, quickly earning a strong reputation for his strategic insight and high productivity. Nate is the primary author of most of MIRI's technical agenda, including the overview document "Aligning Superintelligence with Human Interests: a Technical Research Agenda" (2014) and the AAAI conference paper "Corrigibility" (2015). Prior to MIRI, Nate worked as a software engineer at Google.



Nick Bostrom is a Professor in the Faculty of Philosophy at Oxford University and founding Director of the Future of Humanity Institute within the Oxford Martin School. He is the author of some 200 publications, including Anthropic Bias, Global Catastrophic Risks, Human Enhancement, and, most recently, the book *Superintelligence: Paths, Dangers, Strategies* (OUP, 2014). He has received the Eugene R. Gannon Award for the Continued Pursuit of Human Advancement and been named One of the Top 100 Global Thinkers by Foreign Policy Magazine.



Niel Bowerman recently joined the management team at the Future of Humanity Institute and is a co-founder of the Centre for Effective Altruism. Niel has a PhD (DPhil) in Physics from Oxford University. Niel was a member of President Obama's Energy and Environment Policy Team and was Climate Science Advisor to the President of the Maldives. Niel has attended two G8 Summits, and has delivered speeches at the European Parliament, World Bank, UK Treasury, and at the World Economic Forum in Davos.



Owen Cotton-Barratt is the Director of Research at the Global Priorities Project, a non-partisan Oxford think tank which develops policy proposals targeted at the best opportunities to make people better off. He is also a Lecturer in mathematics at St. Hugh's College, Oxford. He previously worked as a Research Fellow at the University of Southampton, and holds a DPhil in mathematics from the University of Oxford.



Paul Christiano is a PhD student in theoretical computer science at UC Berkeley. His work in algorithms and learning theory has won best paper and best student paper awards at the Symposium on Theory of Computation. Paul is also a research associate at the Future of Humanity Institute, and a co-founder of AI Impacts, an exploratory project to help understand the likely impacts of progress in AI.



Peter Norvig is a Director of Research at Google Inc; previously he directed Google's core search algorithms group. He is co-author of Artificial Intelligence: A Modern Approach, the leading textbook in the field, and co-teacher of an Artifical Intelligence class that signed up 160,000 students, helping to kick off the current round of massive open online classes. He is a fellow of the AAAI, ACM, California Academy of Science and American Academy of Arts & Sciences.



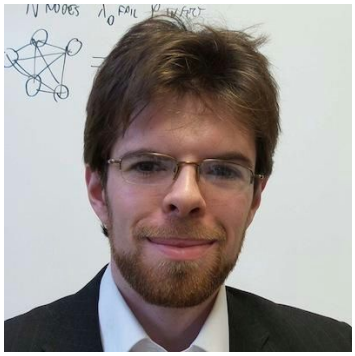
Riva-Melissa Tez has been a Cofounder at Kardashev Technologies, Founder at Berlin Singularity, and Cofounder at StickStar. Riva was previously Lecturer at DAB University Berlin and Cofounder at R.S. Currie & Co. Riva earned her Master of Philosophy (M.Phil.) at University College London, University of London in 2011. Riva is currently a regular columnist at The European.



Sam Altman studied computer science at Stanford University, and dropped out in 2005. Sam Altman was cofounder and CEO of Loopt, which was funded by Y Combinator in 2005 and acquired by Green Dot in 2012. At Green Dot, he was the CTO and is now on the Board of Directors. Sam also founded Hydrazine Capital. In 2014, Altman was named president of Y Combinator, regarded as the best seed-stage accelerator in the world by its founders. He is Chairman of the Board for Helion and uPower, two nuclear energy companies.



Scott Nolan is a Partner at Founders Fund, where he focuses on investments in technology-driven companies across sectors including energy, biotechnology, aerospace, healthcare and advanced manufacturing. Prior to Founders Fund, Scott was an early employee at SpaceX. There, he helped develop the propulsion systems used on the Falcon 1, Falcon 9 and Dragon vehicles and was responsible for the Dragon capsule's thermal and environmental control subsystems.



Seán Ó hÉigeartaigh is the Executive Director of the Centre for the Study of Existential Risk. Previously, he was Senior Academic Manager at the Future of Humanity Institute, where he managed the Oxford Martin Programme on the Impacts of Future Technology, and helped establish the FHI-Amlin Collaboration on Systemic Risk. Sean has a PhD in genome evolution, and a degree in human genetics.



Sebastian Farquhar is the director of the Global Priorities Project, a non-partisan Oxford think tank which develops policy proposals targeted at the best opportunities to make people better off. Seb has developed strategy for leading European public sector health and social care providers, and helped to implement large-scale change programmes in local areas serving more than 2 million people.



Steve Weber is a professor at the School of Information and the Department of Political Science at the University of California, Berkeley. After studying history and international development at Washington University, he received an M.D. and a Ph.D in political science from Stanford University. Steve is the author of several books about international politics and economics. He is also the editor of Globalization and the European Political Economy (Columbia University Press, 2000).



Stuart Russell is a Professor (and former Chair) of Electrical Engineering and Computer Sciences at the University of California at Berkeley and holds the Smith--Zadeh Chair in Engineering. He is a fellow of AAAI, ACM, and AAAS. His book *Artificial Intelligence: A Modern Approach* (with Peter Norvig) is the standard text in AI and has been translated into 13 languages. He also works for the United Nations, developing a new global seismic monitoring system for nuclear treaty verification.



Tom Dietterich is one of the founders of the field of Machine Learning. Among his research contributions was the application of error-correcting output coding to multiclass classification, the formalization of the multiple-instance problem, the MAXQ framework for hierarchical reinforcement learning, and the development of methods for integrating non-parametric regression trees into probabilistic graphical models. He is the President of AAAI.



Viktoriya Krakovna is a doctoral student in statistics at Harvard University. Her work focuses on Bayesian models and Markov chain Monte Carlo methods. She has gained numerous distinctions for her accomplishments in math competitions, including a silver medal at the International Mathematical Olympiad and the Elizabeth Lowell Putnam prize. She is a co-founder of the Future of Life Institute.