FUTURE OF HUMANITY INSTITUTE
ANNUAL REVIEW 2016

In 2016, we continued our mission of helping the world think more systematically about how to craft a better future. We advanced our core research areas of macrostrategy, technical artificial intelligence (AI) safety, AI strategy, and biotechnology safety. The Future of Humanity Institute (FHI) has grown by one third, increasing the capacities of our research and operations teams, and strengthening our engagement at all levels with important actors and stakeholders in our focus areas. We are in the process of transforming FHI into a more highly focused, scalable and even more impactful organization.

AI safety

FHI researchers and collaborators have continued to advance the understanding of scalable AI control. We strengthened our collaboration with DeepMind, publishing a series of internationally recognized papers, most notably Laurent Orseau and Stuart Armstrong's Safely Interruptible Agents.

DeepMind collaboration

Researchers from FHI and DeepMind have weekly meetings and monthly seminars on technical AI safety. Alternating between Oxford and London, researchers from both organizations provide updates on advances in the field. During the most recent seminar, Laurent Orseau gave an overview presentation of his work on agent models, and Owain Evans led a discussion on Cooperative Inverse Reinforcement Learning, a recent AI Safety paper from Stuart Russell's group.

Technical research

Stuart Armstrong, FHI Research Fellow of the Alexander Tamas Initiative on Artificial Intelligence and Machine Learning, and Laurent Orseau presented Safely Interruptible Agents, a collaboration between DeepMind and FHI, at the Conference on Uncertainty in Artificial Intelligence (UAI). Orseau and Armstrong's research explores a method to ensure that certain reinforcement learning agents can be safely interrupted repeatedly by human or automatic overseers without the agent learning from these interruptions. The paper was mentioned in over one hundred media articles and was the subject of pinned tweets by Demis Hassabis and Shane Legg, co-founders of DeepMind.

Jan Leike, FHI Research Associate, and Machine Intelligence Research Institute (MIRI) Research Fellows Jessica Taylor and Benya Fallenstein presented new results at UAI that resolve a longstanding open problem in game theory in A formal solution to the grain of truth problem. During UAI, Jan was awarded the Best Student Paper for Thompson sampling is asymptotically optimal in

general environments, which was co-authored with Tor Lattimore of the University of Alberta, Laurent Orseau of Google DeepMind and Marcus Hutter of ANU. In November, Jan published Exploration Potential, proposing a new quantity to measure how much a reinforcement learning agent has explored its environment class.

Owain Evans, FHI Research Fellow of the Alexander Tamas Initiative, has expanded and managed our AI Safety and Machine Learning internship program and led the work on two papers at the intersection of Machine Learning (ML) and AI safety; Agent-Agnostic Human-in-the-Loop Reinforcement Learning with David Abel, Andreas Stuhlmüller, and John Salvatier, and Active Reinforcement Learning: Observing Rewards at a Cost with David Krueger, Jan Leike, and John Salvatier. Both papers were presented at workshops at the Conference on Neural Information Processing Systems (NIPS). Additionally, Owain continued to write the online textbook Modeling Agents with Probabilistic Programs, which is nearing completion.

Eric Drexler explored and extended the R&D-automation model of potentially safe recursive improvement of advanced AI, and presented results to several research groups in the California bay area. Anders Sandberg has, along with visiting researcher Devi Borg, partially finished a project looking at the potential use of neural interfaces as a tool to facilitate AI safety.

Broader engagement

As part of our strategy for deeper integration with the wider ML community, FHI researchers participated in NIPS, the International Conference on Machine Learning (ICML), UAI, the International Joint Conference on Artificial Intelligence (IJCAI), and several other ML and AI conferences.

In May, ten members of MIRI flew to Oxford from California for the week long workshop “The Control Problem in AI” hosted at FHI. In May and June, we teamed up with MIRI to co-host a twenty-two-day Colloquium Series on Robust and Beneficial AI (CSRBAI) in Berkeley. The colloquium brought together safety-conscious AI scientists from academia and industry to share their recent work. Over fifty people attended from twenty-five different institutions, with an average of fifteen people presenting on any given talk or workshop day.

AI strategy

AI strategy is an area where FHI has a comparative advantage, and we have increased our efforts here. We released three academic papers on openness, policy, and strategy, hired Miles Brundage as our Policy Research Fellow for the Strategic Artificial Intelligence Research Centre (SAIRC), and laid the groundwork for several important collaborations on AI strategy.

Openness, policy, and strategy

In April, FHI’s Director Nick Bostrom finished Strategic Implications of Openness in AI Development which covers a breadth of areas including long-term AI development, singleton versus multipolar scenarios, race dynamics, responsible AI development, and identification of possible failure modes.

The paper has been accepted to Global Policy, to be published in 2017. In December, Nick Bostrom, Allan Dafoe (Assistant Professor of Political Science at Yale University), and Carrick Flynn (FHI Research Project Manager), released a new working paper, Policy Desiderata in the Development of Machine Superintelligence, outlining key considerations for coordination and governance in the development of advanced AI.

Miles Brundage developed a game theoretic model of openness in AI and is in the process of implementing it as a computer simulation. Miles is also co-organizing the joint Oxford-Cambridge February 2017 SAIRC workshop on bad actors and AI, with confirmed participants from Google, OpenAI, the Information Society Project, the Electronic Frontier Foundation, and Microsoft. Carrick Flynn and Professor Allan Dafoe established the Global Politics of AI Research Group, which has the mission of helping researchers and political actors to adopt the best possible strategy around the development of AI. The group currently consists of eleven research members more than thirty volunteers.

Toby Ord completed a working draft of "Lessons from the Development of the Atomic Bomb," which looks at the technical, political, and strategic aspects of the development of the atomic bomb comparing these to considerations surrounding the development of AI. It is forthcoming in the Bulletin of Atomic Scientists.

FHI hosted a four-day workshop called Policies for Responsible AI Development with the Cambridge Centre for the Study of Existential Risk (CSER). Anders Sandberg finished a paper on Energetics of the brain and AI. FHI Research Associates Katja Grace and Allan Dafoe, and John Salvatier completed an updated survey of AI expert's predictions which will likely be published in 2017. Additionally, Allan Dafoe, and Stuart Russell published Yes, We Are Worried About the Existential Risk of Artificial Intelligence in MIT Technology Review, in response to an article by Oren Etzioni.

Engagement with governments

Policymakers have become increasingly interested in strategic questions relating to AI and have consulted with FHI on a number of occasions. In April and May, FHI researcher Owen Cotton-Barratt gave written and oral evidence to the UK Parliament's Science and Technology Commons Select Committee. Following this, in October, the committee released Robotics and artificial intelligence. The report refers to the oral evidence provided by Dr Cotton-Barratt and references the written submission of the joint FHI/CEA Global Priorities Project (GPP) as part of its recommendation to the government that it establish a standing Commission on Artificial Intelligence. In January 2017, the UK Government announced its intention to create this standing commission. FHI continues to advocate at the national and regional levels; Miles Brundage met with policy makers and analysts at the European Commission in Brussels and Niel Bowerman gave evidence to the legal affairs committee of the European parliament on the ethics of AI. FHI also submitted a response to the White House's Request for Information on the Future of Artificial Intelligence.

Biotech safety

FHI is delighted to have hired our first international policy specialist on biotechnology, Piers Millett. Piers was formerly the Acting Head of the UN Biological Weapons Convention Implementation Support Unit and consults for the World Health Organization on research and development for public health emergencies. Since starting at FHI, Piers has participated in nine workshops, conferences and events, including a WHO informal consultation, the International Genetically Engineered Machines Competition (iGEM) and the 8th Review Conference of the Biological Weapons Convention. Piers is developing a short review that links together issues around biotechnology, an outline of a publication on the cost efficiency of dealing with different types of biorisks, and is liaising with biotech stakeholders in the UK on opportunities to strengthen arrangements for dealing with biorisks.

In March, Owen Cotton-Barratt, Sebastian Farquhar, and Andrew Snyder-Beattie released the policy working paper Beyond risk-benefit analysis: pricing externalities for gain-of-function research of concern, outlining an approach for handling decisions about Gain-of-Function (GoF) research using liability insurance. At the start of November, Piers Millett and Eric Drexler participated in a biological engineering horizon scanning workshop co-hosted by CSER and FHI. This workshop, and its horizon scanning process, is intended to result in a peer-reviewed publication highlighting the 15-20 developments of greatest likely impact.

Macrostrategy

Our ongoing work on macrostrategy involves forays into deep issues in several fields, including detailed analysis of future technology capabilities and impacts, existential risk assessment, anthropics, ethics, reasoning under uncertainty, the Fermi paradox, and other indirect arguments. We are developing concepts and analytic tools that make it possible to think systematically about the long-term expected value of present actions. We have also engaged extensively with partners and stakeholders in these areas. FHI staff have worked out a number of “top funding ideas” for funders whose aim is to improve the long-term future, and presented that to a large philanthropic foundation, some major donors, and a government agency. Some of these ideas have already been adopted.

Existential and catastrophic risk

In February, twenty academics and policy-makers from the UK, USA, Germany, Finland, and Sweden gathered at the University of Oxford to discuss governance in existential risks. Following up on this workshop, researchers from FHI and CEA travelled to Helsinki to discuss existential risk policy with several Finnish government agencies and to attend meetings held at the Office of the President. Also, Toby Ord has started preparing to write a book on existential risk, and the Global Priorities Project released the Global Catastrophic Risks Report 2016.

Institute for Effective Altruism and DFID

FHI recruited Prof. William MacAskill and Prof. Hilary Greaves to set up the Oxford Institute for Effective Altruism, a proposed new research centre within Oxford University. They intend to create a

research group spanning economics and philosophy in which effective altruism can be studied more formally in an academic setting. In addition, Toby Ord has been meeting with senior decision makers at the UK Department for International Development (DFID), helping them incorporate ideas from effective altruism into their priority setting.

Policy research

In January, Nick Bostrom, Anders Sandberg and Tom Douglas had The Unilateralist's Curse: The Case for a Principle of Conformity published in Social Epistemology. In July, Owen Cotton-Barratt, along with Marc Lipsitch of Harvard University, and Nicholas G. Evans from the University of Pennsylvania published Underprotection of Unpredictable Statistical Lives Compared to Predictable Ones in Risk Analysis. The paper argues that society may under-invest in protecting lives from large but low-probability catastrophes relative to smaller but more frequent occurrences.

Technical trends and other research

FHI researchers did an internal multi-week exploratory and agenda-setting project on technology forecasting. In follow-up, we organized the Workshop on AI Safety and Blockchain in December, which featured blockchain and crypto-economic luminaries such as Vitalik Buterin and Wei Dai. FHI researchers and half a dozen volunteers have started a related project looking at the possibility space created by advances in technologies relevant to surveillance and structured transparency.

Eric Drexler helped to organize and lead an international research workshop at the University of Cambridge hosted by CSER on the implementation of atomically precise manufacturing, including its current status and potential development into the future. Anders Sandberg had a paper accepted to Bioethics, two in Cambridge Quarterly of Healthcare Ethics, and has submitted a paper on the aestivation hypothesis for publication. Toby Ord, Anders Sandberg and Eric Drexler have been finalising the paper "Dissolving the Fermi Paradox." Ben Levinstein worked with FHI until summer 2016, and his work resulted in, among other things, the papers Cheating Death in Damascus with Nate Soares and A Pragmatist's Guide to Epistemic Utility. Research Associate Robin Hanson published a new book titled The Age of Em: Work, Love, and Life when Robots Rule the Earth. The book explores a world where humanity hasn't created strong AI but are running emulations of humans on computers. On 19 October, FHI and Robin Hanson organised a workshop and public talk held at the Oxford Martin School on the The Age of Em.

Engagement

To further our policy and strategy work, FHI has engaged with numerous governmental, non-profit, and commercial organizations.

Researchers have attended or spoken at dozens of events and conferences including Seoul World Strategy Forum, Effective Altruism Global, Zurich Economic Forum, C2 Montreal, CeBit, RSA, IP EXPO Europe, and the Windsor AI Debates, and have given talks for the Bank of England,

the US National Academies, NASDAQ, the Geneva Centre for Security Policy, Audi, SAP, Lynx Asset Management, Bloomberg London, and many other organisations.

FHI has hosted guests from the Prime Minister of Singapore's office, two delegations from Japan's Ministry of Trade, Economy, and Industry, the Swedish Ministry of Foreign Affairs, the French parliament, and the United Nations Interregional Crime and Justice Research Institute. FHI has engaged in consultations with the UK parliament, the UK Prime Minister's office, the Finnish Foreign Ministry, RAND, the US Government's Intelligence Advanced Research Project Activity, Defense Advanced Research Project Agency, and Department of Defense, among others. We also had more than 25 external speakers visit FHI.

FHI has engaged with Open Philanthropy, Barclays, Ethereum, FiveAI, the Breakthrough Initiatives, the Global Catastrophic Risk Institute, the Information Society Project at Yale Law School, the Global Challenges Foundation, the UN Office for the Coordination of Humanitarian Affairs, the Institute for Futures Studies, and The Future Society at the Harvard Kennedy School of Government, among many others. FHI is also working with the IEEE on developing safety and ethical recommendations and standards for the AI industry.

According to Google News, FHI was mentioned in the media around 2500 times in 2016, including President Obama's discussion of Nick Bostrom's work in a Wired interview.

Organization

Fundraising

We have continued to apply for academic grants in order to maintain our diverse set of funding sources. We won part of a grant from the Leverhulme Trust for the new Centre for the Future of Intelligence at the University of Cambridge. FHI also had a generous offer from Luke Ding to fund William MacAskill's full salary for five years, and to fund Prof. Hilary Greaves for four years from mid-2017 if alternative funding cannot be raised. Overall, FHI raised £1.4m in new pledged funding in 2016.

Staff

As part of our expansion in research and operations, we grew the organization by one third, to around 24 in total including research staff, research associates, and support staff. We have put a greater focus on our operations side, resulting in a smoother-running organisation.

On the research side, we hired Piers Millett, Miles Brundage, and Jan Leike (who later moved to DeepMind while continuing as an FHI research associate). Carrick Flynn was hired as Research Project Manager. On the operations side, we hired Kathryn Mecrow to work as our Administrative Officer. To bootstrap the Institute for Effective Altruism, we hired Will MacAskill and Hilary Greaves. We have hosted Daniel Filan, David Abel, and Fiona Furnari as research interns, and Devi Borg as a visiting researcher. We had Peter McIntyre working for us for a while, David Krueger collaborating with

us as an external consultant, and we have taken on David Kristoffersson on a temporary basis. Feng Zhou, Cecilia Tilli and Ben Levinstein left FHI in 2016.

The new year

We intend to make AI strategy an even stronger focus in 2017. For this reason, we are recruiting a leading international relations specialist with a strong interest in AI from a top US university to grow our capacity in this area. In particular, we plan to do more research into scenarios in which AI arrives unexpectedly soon.

We are likely to be provided with another suite by Oxford University within Littlegate House in order to expand our offices. We plan to strengthen our communications capacity and our “impact capacity” - our ability to go out into the world and have our ideas applied.

We have many papers and collaborations underway, and we are truly excited to continue developing these partnerships and collaborations, and to continue our transformation into a bigger, more focused, and even more high-performing organization in 2017.