Q1 2017 FHI Quarterly Briefing

In the first 3 months of 2017, FHI has continued its work exploring crucial considerations for the long-run flourishing of humanity in our four research focus areas:

- **Macrostrategy** - understanding which crucial considerations shape what is at stake for the future of humanity.
- **AI safety** - researching computer science techniques for building safer artificially intelligent systems.
- **AI strategy** - understanding how geopolitics, governance structures, and strategic trends will affect the development of advanced artificial intelligence.
- **Biorisk** - working with institutions around the world to reduce risk from especially dangerous pathogens.

Key outputs you can read

- Strategic Implications of Openness in AI Development. Global Policy 2017. Nick Bostrom.
  This paper reacts to a pressing question facing the AI community. Openness builds trust and speeds up some kinds of research relative to others. But it also creates risks. This paper is the first thorough investigation of the risks and benefits of openness in AI development with existential risk in mind.
- Agent-agnostic Human-in-the-loop Reinforcement Learning. ArXiv 2017. David Abel, John Salvatier, Andreas Stuhlmüller, Owain Evans.
  Safe exploration is one of the key parts of the Concrete Problems in AI Safety research agenda. This paper contributes to that project by building a general framework for thinking about human input into reinforcement learning and using it to explore possible guarantees of avoiding catastrophic accidents during training.
- Existential Risk: Diplomacy and Governance. Global Priorities Project 2017. Sebastian Farquhar, John Halstead, Owen Cotton-Barratt, Stefan Schubert, Haydn Belfield, Andrew Snyder Beattie.
  Working on behalf of the Finnish government, this project reviewed over a hundred possible interventions to select three which were easily accessible. It recommends more co-operative extreme pandemic scenario planning, better geo-engineering governance, and development of institutions for future generations. (This work began at the Global Priorities Project, whose policy work has now joined FHI.)

Key workshops and conferences

- Bad Actors and Artificial Intelligence
  FHI hosted a workshop on the misuse of artificial intelligence by state and non-state actors. The workshop was co-organised with our partners at the Centre for the Study of Existential Risk and the Centre for the Future of Intelligence and brought together over 30 experts from diverse disciplines to identify problems and potential resolutions. The workshop will lead to a report summarising its output.
- Beneficial Artificial Intelligence at Asilomar
  Several FHI staff spoke at and attended the BAI2017 conference in Asilomar. This gave us an opportunity to connect with colleagues from around the world and exchange beliefs and ideas.
- AAAI 2017
  Several FHI staff attended the AAAI conference this spring. This gave us an opportunity for our technical AI safety research team to share ideas with other researchers in artificial intelligence, and also to connect new research in AI to the work at the Strategic AI Research Centre on AI policy.

Future of Humanity Institute
Faculty of Philosophy, University of Oxford
Suite 1, Littlegate House, 16/17 St Ebbe's Street, Oxford OX1 1PT

Individual research updates

Macrostrategy and AI Strategy
Nick Bostrom started work on a project addressing the options that would be available to the AI safety community in worlds with short timelines to AGI among other strategic issues. He spoke at the Beneficial AI conference and presented is work at the International Studies Association, a widely respected scholarly association in the field of international studies.

Miles Brundage co-chaired the Bad Actors and Artificial Intelligence workshop. He has been modelling openness in AI, participated in red team/blue team exercises on adverse AI Outcomes at Arizona State University and attended AAAI 2017.

Owen Cotton-Barratt presented his work on classifying risks from extinction at the Beneficial AI conference. He has been working on risk management institutions in biosafety and exploring AI-safety benchmarks.

Sebastian Farquhar and his co-authors published a policy report on existential risk governance interventions as part of his project with the Finnish Government. He participated in the Beneficial AI and Bad Actors conferences.

Toby Ord gave a seminar on his work on long-run consequences, participated in the Beneficial AI and Bad Actors conferences and finished a paper on the Fermi paradox with his co-authors, Eric and Anders. He has been laying the groundwork for a book on existential risk.

Anders Sandberg submitted an entry to the Oxford Research Encyclopedia on Natural Hazard Science on human extinction. He is working on a number of papers from differential technological development to the biosecurity risk pipeline. He lectured at the Geneva Centre for Security Policy, presented to the Finnish Government, British Interplanetary Society, AAAS annual meeting, Wilton Park, UKSRN and many other venues.

Andrew Snyder-Beattie led a large grant proposal project which has now been shortlisted and has been working on a number of papers in biosecurity.

AI Safety
FHI and DeepMind continue to co-host monthly seminars aimed at deepening the ongoing fruitful collaboration between AI safety researchers in these organisations.

Stuart Armstrong has been developing ideas on a range of topics including oracle design, value learning, qualia, low-impact AI, and extending his work on interruptibility.

Eric Drexler has been drafting a limited-circulation set of documents on development-oriented approaches to AI safety. He helped to initiate a project on structured transparency and has been finishing a paper on the Fermi Paradox with his co-authors, Anders and Toby. He participated in the Beneficial AI and Bad Actors conferences.

Owain Evans pre-published a paper on human-in-the-loop reinforcement learning, completed an online interactive book on building agents at agentmodels.org and co-authored a long post with Jacob Steinhardt on Inverse Reinforcement Learning.

Biosecurity
Piers Millett has provided an internal briefing to the WHO's R&D Blueprint process on prioritisation of pathogens. He submitted a paper on genome editing and drafted two other papers with co-authors. He continued his work building networks across the community and assisted the planning and execution of CSER's conference on biorisk. He has been supporting the development of post-Brexit biotech governance, and oversight for the post-Ebola biobank.

Funding
This quarter FHI received a large grant from the Open Philanthropy Institute for roughly £1.7m. This puts FHI in a healthy financial position, although we continue to accept donations. We expect to spend approximately £1.3m over the course 2017. Including three new hires but no further growth, our current funds plus pledged income should last us until early 2020. Additional funding would likely be used to add to our research capacity in machine learning, technical AI safety and AI strategy. If you are interested in discussing ways to further support FHI, please contact Niel Bowerman.