Future of Humanity Institute 2005-2024: Final Report

Anders Sandberg

Report #2024-1. Version 1.0

Preamble

Normally manifestos are written first, and then hopefully stimulate actors to implement their vision. This document is the reverse: an epitaph summarizing what the Future of Humanity Institute was, what we did and why, what we learned, and what we think comes next.

It can be seen as an oral history of FHI from some of its members. It will not be unbiased, nor complete, but hopefully a useful historical source. I have received input from other people who worked at FHI, but it is my perspective and others would no doubt place somewhat different emphasis on the various strands of FHI work.

Contents

Presentation of the Future of Humanity Institute

The James Martin seminar room with the whiteboards that acted as the shared brains of FHI. Toby Ord, Andrew Snyder-Beattie, Nick Bostrom, Cecilia Tilli, Anders Sandberg and Stuart Armstrong.

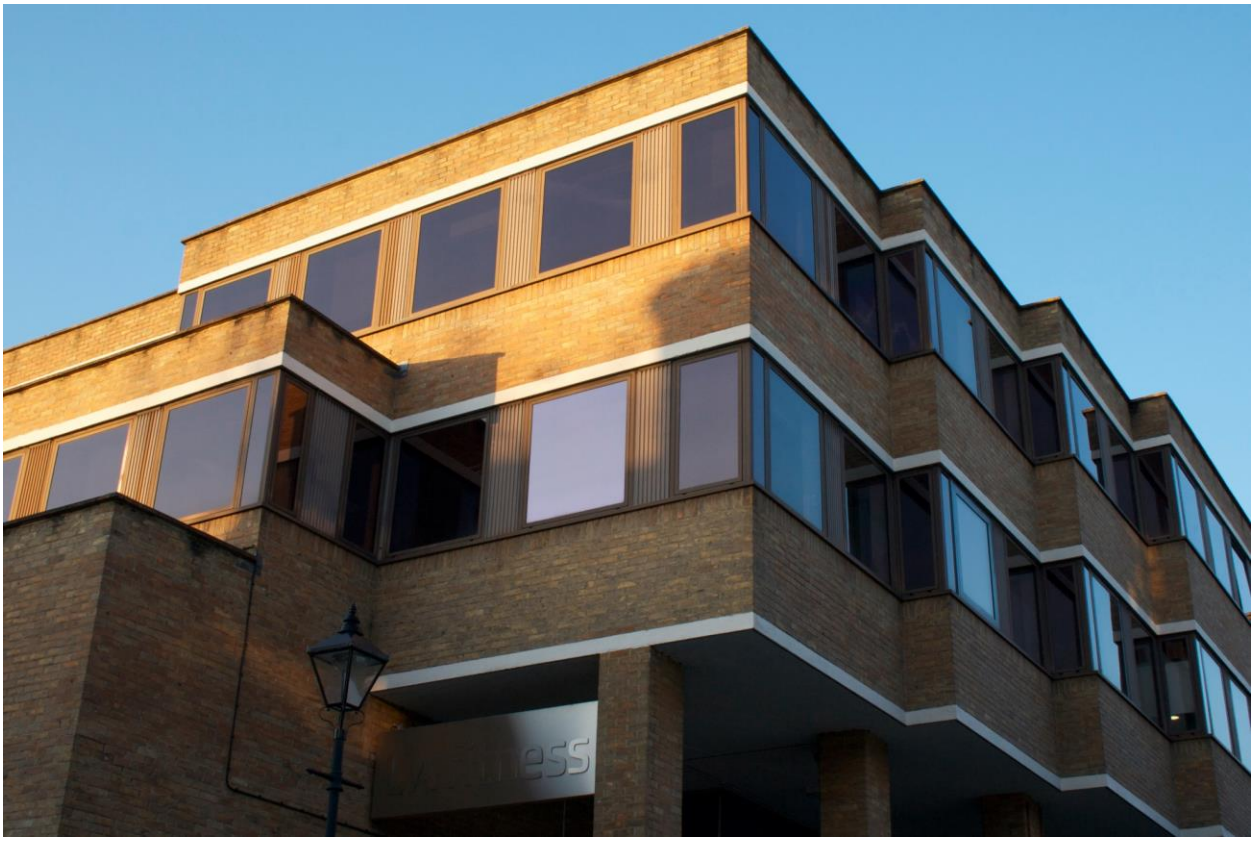
Established in 2005, initially for a 3-year period, the Future of Humanity Institute was a multidisciplinary research group at Oxford University. It was founded by Prof Nick Bostrom and brought together a select set of researchers from disciplines such as philosophy, computer science, mathematics, and economics to study big-picture questions for human civilization, attempting to shield them from ordinary academic pressures and create an organizational culture conducive to creativity and intellectual progress.

How did this come about?

The most surprising thing is that the institute ever started. While there is no shortage of institutes proclaiming that they are going to pursue the big questions in a transdisciplinary way, in practice this rarely happens. Indeed, it is rare to see funding for such topics: as is well known, academic funders prefer research that already has high prestige, a known track record, a clear target, and a clear disciplinary home. At the beginning FHI had none of these.

The role of Dr. James Martin cannot be overstated. He not only provided the initial funding via the Oxford Martin School (then known as the James Martin 21st Century School) but actively advocated for FHI during his visits to Oxford and took a special interest in FHI's research. His perspective of the current century as a "make or break century" with numerous interwoven challenges fit very well with the FHI interdisciplinary mindset. While FHI was growing up the Martin School was also finding its place, becoming an interdisciplinary structure cutting through many of the institutional walls at Oxford between faculties, colleges, and other institutions.

In the School FHI's role was to scan the horizon for the long-term, overarching issues, rather than the more specific topics the other School Institutes dealt with. Besides encouraging interdisciplinary networking with the other Institutes, the School under Prof Ian Goldin promoted policy-relevance: it was not enough to produce excellent research, but also to be able to answer the "so what?" question of how that research could actually change the world. If it could, the School was happy to support policy outreach globally, something that helped support the level of ambition.



Littlegate House, 2007. A labyrinthine office building inhabited by various small University centers and research groups. While not the most inspiring architecture, we had an enjoyable view of the garden of Pembroke College across the street… while they had to put up with our facade.

Another natural ally was the The Oxford Uehiro Centre for Practical Ethics (founded in 2003). For a long time FHI incubated inside and alongside the Centre. The aim of helping to understand, evaluate, and

respond to radical change was very compatible with the Centre's goal to support debate and deeper rational reflection on practical ethics. While FHI was typically focused on more futuristic topics there was a shared understanding that this all fell within the very broad scope of ethical practice. Practical ethics does not have to be localized to the medical bedside, but could be about global policy and helping future generations.

For the first years, FHI and the Centre shared offices in Littlegate House on St. Ebbe's Street and remained close until FHI moved to Trajan House in 2021.

Ambitions

Research seminar where Eric Drexler, David Deutsch and Nick Bostrom discussed possible ways of analyzing the limits of technological progress. May 2012. (L to R: Eric Drexler, David Deutsch, Owain Evans, Nick Bostrom, Owen Cotton-Barratt, Toby Ord, and a visitor. Professor Vincent Müller just outside the frame to the left.)

One of the key parts of the ethos of the institute was to actually try to work on what is important for humanity. Not what is recognized, accepted, or fashionable: our job has been to find things deserving of being recognized, show that they matter, invent the theoretical and conceptual tools needed to start to do useful work on them—and then (hopefully) hand it off to others as the topic matures. This is very much in line with Richard Hamming's famous questions ("What are the important problems in your field? Why don't you work on them?")[1].

[1] Hamming, R., & Kaiser, J. F. (1986, March). You and your research. In *Transcription of the Bell Communications Research Colloquium Seminar* (Vol. 7). Morristown, NJ: Bell Communications Research.

As Nick Bostrom put it in the first Bimonthly Progress Report in April 2006[2]:

> "We pursue an opportunistic research agenda, in that we focus on those questions that are amenable to analysis, and where new insights would bring the greatest payoff in terms of improving humanity's ability to navigate its future."

That approach obviously did not optimize for standard academic reward, but it did favor both field-building and a form of advocacy once important insights came about. Rather than just spread the word in academia, the institute ethos favored reaching out to relevant groups - publics, policymakers, entrepreneurs, particular research fields - and try to make them aware of the issue. One welcome consequence was that over time a vast bilateral network of interdisciplinary connections emerged as people in these other groups began collaborating on the topics.

Field-building involves convincing other practitioners that a new topic is worthy of study, demonstrating that progress can be made on the topic, developing research agendas, establishing organizations, and—if all goes well—helping to direct funding and interest to the field.

There is an infinite number of potentially important questions that can be explored and only a finite amount of time and resources to do so. Hence one of the earliest discussions became how to decide on research priorities, leading to regular discussions of the philosophy of priority-setting that influenced much of subsequent work. We recognized that many worthy topics may be best left to others, or to the future, and that our focus must always be on where our marginal benefit was greatest.

(See Appendix A for a survey of how the formal goals on the website changed over the years)

[2] https://web.archive.org/web/20060512085807/http://www.fhi.ox.ac.uk:80/Papers/FHI%20Newsletter%201%20-%20April%20200611.pdf

History

Early days

Nick directing a workshop on global catastrophic risk, with Milan Ćirković and Rebecca Roache. 2008

The first major project of FHI was the EU-funded ENHANCE project, 2006-2008. ENHANCE explored the ethics and social impact of human enhancement, with the Oxford node (FHI and the Uehiro Centre of Practical Ethics) investigating cognitive enhancement. We sought to produce the necessary "deliverables" while also creating a protected space where we would pursue the core questions that we were really interested in.

For the first few years, FHI had a major focus on the ethics of human enhancement, but gradually three main interests crystallized: human enhancement and other emerging technologies that could fundamentally change the human condition, global catastrophic and existential risks threatening humanity's future, and methodology: how to think well about these highly uncertain things.

Beside the standard academic research, the institute also engaged in significant outreach. Of particular note was the start of the (initial) group blog Overcoming Bias in November 2006.[3] It later became the personal blog of professor Robin Hanson, FHI research associate, but for many years acted as a forum devoted to investigation of rationality, futures studies and other topics; in turn it became a seed for the forum LessWrong. FHI research associate Peter Taylor (in conjunction with the FHI program on global catastrophic risk) started the Lighthill Risk Network with the aim of bringing world-wide scientific expertise in various aspects of risk to the financial services sector and (re)insurance industry.

During this early period one of the key achievements was the book *Global Catastrophic Risk*[4] and the 2008 GCR conference (plus workshops) bringing together a field of global catastrophic risk and existential risk studies.[5] The conference was an important milestone in building an academic community around reducing risks to humanity's future. AI was one of several topics under discussion, and Eliezer Yudkowsky was among the speakers.[6]

Another field building exercise was the 2007 workshop on whole brain emulation that led to the 2008 roadmap for WBE[7]. While this has remained a small and somewhat specialized field, activity and capacity has steadily grown since then. In 2023, a follow-on workshop in Oxford brought together many of the original participants and new people to take stock of over a decade of progress: many of the blue-sky visions of 2007 are now practical reality.

These early years were a funding-constrained period. The Institute remained small (only 3 salaried research staff aside from the director) and members felt strongly constrained to produce papers that were academically legible. Research might have to be done about less relevant topics just to pay the bills and fit in with the surrounding academic environment. (Some of our research hires were also constrained by the requirement to produce outputs comprehensible to the academic community, and to outside faculty members on the appointment panels.) Yet much significant research was done.

Indeed, FHI was very prolific in this period. From November 2005 to July 2007, it produced 40 journal papers (+ 32 in preparation); 43 articles and book chapters; 5 books; 20 reprints and translations.[8] While a

---

[3] https://web.archive.org/web/20070705000635/http://www.fhi.ox.ac.uk:80/updates.html#blog

[4] Bostrom, N., & Ćirković, M. M. (Eds.). (2011). *Global catastrophic risks*. Oxford University Press, USA.

[5] https://web.archive.org/web/20080914211545/http://www.global-catastrophic-risks.com/aboutconf.html

[6] http://www.global-catastrophic-risks.com/reports.html

[7] Sandberg, A., & Bostrom, N. (2008). *Whole brain emulation: A roadmap*. Technical Report #2008-3, Future of Humanity Institute, Oxford University

[8] https://web.archive.org/web/20110908095411/http://www.fhi.ox.ac.uk/__data/assets/pdf_file/0003/19902/Final_Complete_FHI_Report.pdf

fair amount might have been research done to fulfill grants rather than being relevant (causing some concern about lack of clear focus[9]), it also published many widely cited references in the core fields.

Maturation

Nick Bostrom at the United Nations. October 7 2015.

By 2010 the cognitive enhancement workstream began to conclude. After publishing numerous research papers, reports, and books on the topic as well as hosting several workshops it was increasingly clear that while there was still great academic and public interest in the ethics of enhancement, the likely impact of the field over the coming years was looking minor compared to other considerations such as global

[9] "Much of the dispersion is *caused* by the lack of unrestricted funds (and lack of future funding guarantees). Since we don't have enough funding from private philanthropists, we have to chase academic funding pots, and that then forces us to do some work that is less relevant to the important problems we would rather be working on. It would be unfortunate if potential private funders then looked at the fact that we've done some less-relevant work as a reason not to give."

https://www.lesswrong.com/posts/oJx2Qguf8EasLSet2/safety-culture-and-the-marginal-effect-of-a-dollar#HTMSNsQyjebdymu4h

catastrophic risks and artificial intelligence. While some members like Anders Sandberg continued neuroethical research, FHI's focus moved elsewhere.

This was a fairly typical approach for FHI: we attempted to notice overlooking topics deserving of research and attention early, germinating it in the sheltered FHI greenhouse; showing that progress could be made; coalescing a field and setting research directions; attracting bright minds to it; and once it's established enough, setting it free, and moving onto the next seedlings.

This was when artificial intelligence safety became a major focus. Existential risk from AI had been discussed by FHI members for many years,[10] and was one of the risks that most concerned us, but it wasn't yet a core part of our academic research. However, when Nick Bostrom began work on a book on existential risk in 2009 he soon found one of the chapters, the one on AI, getting out of hand. The issue of risks from superintelligent systems, especially self-improving artificial intelligence, turned out to be much deeper and wider than initially expected. The chapter began to take a life of its own, evolving into a long-running research project and eventually a monograph on its own: Bostrom's 2014 bestseller *Superintelligence: Paths, Dangers, Strategies*. During the writing of the book the AGI study group discussed the work in progress, spinning off parallel research papers. Stuart Armstrong and Daniel Dewey were hired as technical AI safety researchers in 2011 and 2012. AI safety and governance became one of the signature topics for FHI research.

[10] https://nickbostrom.com/old/predict

Demis Hassabis, from the newly founded DeepMind, presenting at the 2011 Winter Intelligence conference[11].

Several conferences were held on AI-related topics. The January 2011 *Winter Intelligence Conference* was FHI's first conference explicitly focused on AI, and coincided with the start of the organization's pivot towards AI as a primary research topic. As part of the conference, an early manuscript of *Superintelligence* was circulated amongst attendees.[12] The second Winter Intelligence Conference, hosted by FHI in December 2012, was organized in parallel with the 2012 AGI conference. AGI-12 was hosted in Oxford, and was followed by FHI's AGI Impacts conference, a successor to the 2011 Winter Intelligence conference.[13] In September 2013 Philosophy and Theory of AI, was hosted in Oxford by FHI, and brought together another impressive group of academics working in philosophy and computer science, most notably Stuart Russell (who gave the keynote address), and Daniel Dennett.[14] Later FHI participated in the influential 2015 Puerto Rico and 2017 Asilomar AI conferences organized by the Future of Life Institute.

Another project that developed in unexpected directions was the Oxford Martin Programme on the Impacts of Future Technology which began in 2011. Carl Frey and Michael Osborne found a new way of

[11] https://www.youtube.com/watch?v=IjG_Fx3D0o0

[12] https://www.fhi.ox.ac.uk/winter-intelligence-conference-2011/

[13] https://web.archive.org/web/20120815232147/http://www.winterintelligence.org/

[14] http://www.pt-ai.org/2013

analyzing the potential automation of different jobs, leading to the widely cited (and misunderstood) claim that 47% of jobs may be at risk from automation[15]. This work was continued at the Oxford Martin Programme on Technology and Employment, later the Oxford Martin Programme on the Future of Work. A key member of the FHI team, professor Vincent Müller, responsible for much of the success with the AI conferences, joined because of this project.

As the investigations into AI risk expanded, more related projects were started. The Alexander Tamas Initiative on Artificial Intelligence Safety was followed by and complemented by the Strategic Artificial Intelligence Research Centre and the Leverhulme Centre for the Future of Intelligence, both starting in 2016.[16] [17] All of them connect FHI deeply into the world of AI engineering and policy. As a response to skepticism from AI practitioners that there was any real science to AI safety (plus, genuine interest from our side), FHI members began to regularly participate (and publish) in AI conferences, helping establish technical AI safety as a credible field.

[15] Frey, C. B., & Osborne, M. A. (2017). The future of employment: How susceptible are jobs to computerisation?. *Technological forecasting and social change*, 114, 254-280.

[16] https://web.archive.org/web/20240407050858/https://www.fhi.ox.ac.uk/research/research-areas/strategic-centre-for-artificial-intelligence-policy/

[17] https://www.fhi.ox.ac.uk/10-million-grant-for-new-centre-for-the-future-of-intelligence/

FHI's Governance of AI Program was launched in 2017 (L to R: Tanya Singh, Ben Garfinkel, Nick Bostrom, Baobao Zhag, Carrick Flynn, Jade Leung, Allan Dafoe, Roxanne Heston, Laura Pomarius)

It became increasingly clear that even if AI safety were to be somehow technically solved, it might not be enough as implementation of safe AI development and deployment were an entirely separate set of problems. So, AI governance joined AI safety as a key focus of FHI. This was spearheaded by Professor Allan Dafoe, who joined FHI in 2017 to establish the Governance of AI Program. This eventually grew to become FHI's largest team before spinning out of the university in 2021 to escape bureaucratic restrictions imposed by the Philosophy Faculty administration, and going on to have great success as an independent organization, the Centre for the Governance of AI.

In 2012 the Centre for the Study of Existential Risk (CSER) at Cambridge University was started. Seán Ó hÉigeartaigh, our very capable administrator and the only person known to be able to work on the Oxford-Cambridge bus, was instrumental in making it real and became the founding executive director. FHI and CSER were in many ways sibling institutes; sharing a core focus (on existential risk), but each forging our own path. Through the years we have benefited greatly from each other's different approaches to similar questions.

Meanwhile FHI also began (as far as I know) the first industry collaboration in the history of the Oxford philosophy department. The reinsurance company Amlin brought an important problem to the institute: given that the leading complex, imperfect mathematical models of catastrophe risk are shared across the insurance business, might this create a systemic risk like the one that had brought down the financial crisis of 2007-8? Together Amlin and FHI began exploring the cognitive biases, systemic risks and social epistemology of the world of insurance, culminating in a scoring system for systemic risk in 2015. Still, while useful and indirectly linked to FHI interests, it was not a natural fit. This research was later handed over to a further collaboration with the Institute for New Economic Thinking at Oxford.

After the success of *Superintelligence*, FHI's funding levels increased substantially.[18] In 2015 FLI's first round of grants, FHI researchers received $1.8m.[19] The European Research Council.[20] awarded a €2m grant for the ERC Advanced Project on Uncertainty and Precaution, which attacked the problem of how decision-making can be guided better than the precautionary principle in domains of radical uncertainty. Eventually, Open Philanthropy became FHI's most important funder, making two major grants: £1.6m in 2017,[21] and £13.3m in 2018.[22] Indeed, the donation behind this second grant was at the time the largest in the Faculty of Philosophy's history (although, owing to limited faculty administrative capacity for hiring and the subsequent hiring freezes it imposed, a large part of this grant would remain unspent). With generous and unrestricted funding from a foundation that was aligned with FHI's mission, we were free to expand our research in ways we thought would make the most difference.[23]

---

[18] The initial funding was provided by James Martin, via the James Martin 21st Century School, in the form of a multi-year funding grant and subsequent extensions, totaling £1.3m between 2006-2012. In 2011, FHI secured a second major round of funding from the Martin School, via their Future of Technology program, with the mission to "analyze possibilities related to long-range technological change and the potential social impacts of future transformative technologies." This provided £1.2m funding between 2012-2014, with supplementary funding of £225k from a private donor.

Aside from the Martin School, FHI received £0.9m from Amlin, as part of the collaboration on systemic risk between 2013-2016. FHI also received donations from a small number of private individuals, totaling £760k prior to 2016. Between 2005-2016, FHI's funding sources were split as follows: 60% from the Martin School, including via the FutureTech program; 20% from Amlin; 20% from private donors.

[19] https://futureoflife.org/first-ai-grant-recipients/

[20] https://web.archive.org/web/20230928082643/https://www.fhi.ox.ac.uk/fhi-awarded-prestigious-e2m-erc-grant/

[21] https://www.openphilanthropy.org/focus/global-catastrophic-risks/potential-risks-advanced-artificial-intelligence/future-humanity-institute-general-support

[22] https://web.archive.org/web/20240407050843/https://www.fhi.ox.ac.uk/grant-announcement/

[23] While these were the major grants, we also started to receive a significant number of donations from ordinary people who reached out to the university to donate small sums to FHI - a sign that the research was well received. We also received numerous letters and emails, from people all over the world, thanking us for the work that was being done at FHI.

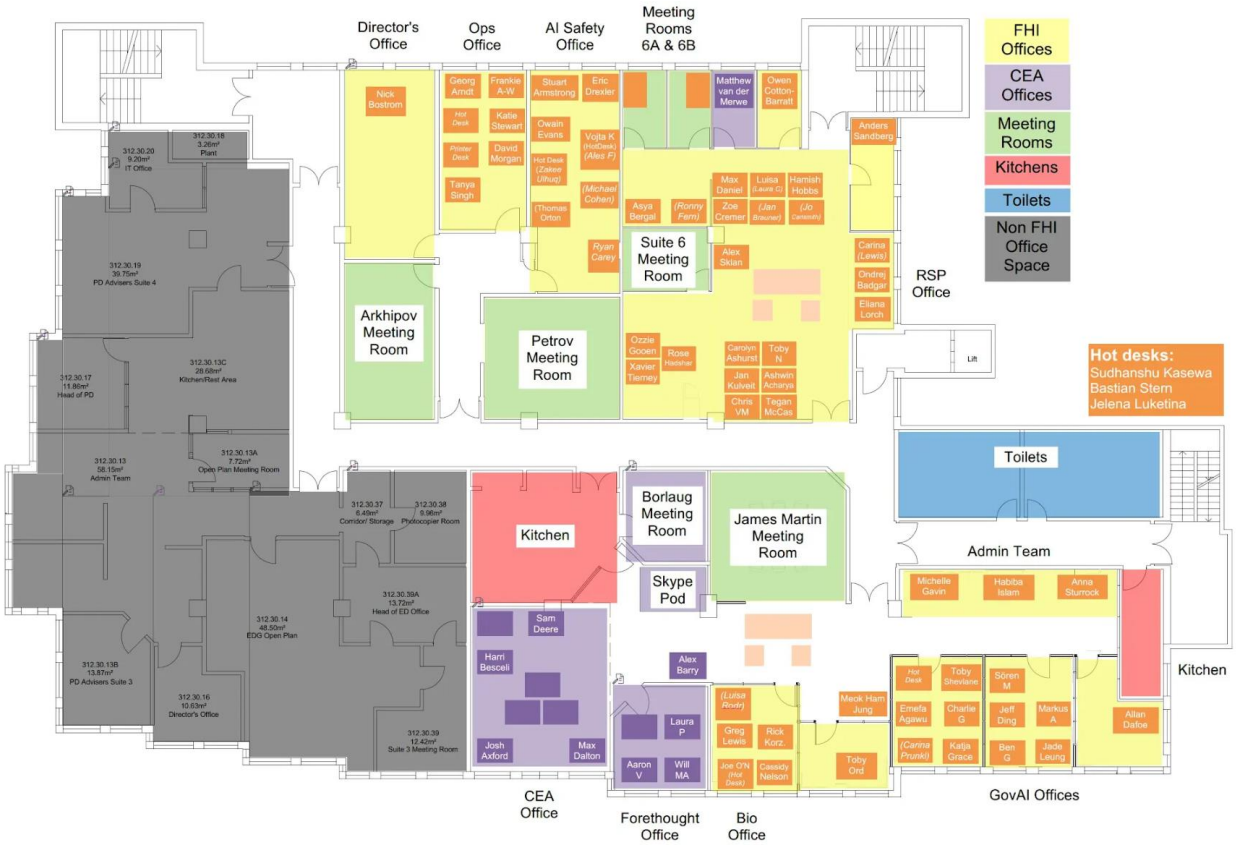
FHI expanded into several directions. People now had the opportunity to create their own subgroups within FHI, with wide autonomy. A Biosafety team was set up. The Global Priorities Project emerged as a collaboration between the Centre for Effective Altruism and FHI, analyzing how to prioritize policy and handling unprecedented technological risks. The Population Ethics: theory and practice team approached the questions of how to think about value in respect to different and future populations.

AI safety discussion in the Petrov Room. Jan 20, 2018. Vladimir Mikulik, Joar Skalse, Shahar Avin, Jelena Luketina, Jan Leike, Ben Garfinkel

One of the insights that had developed was the need to stimulate the growth of talent in the interdisciplinary space FHI resided in. While there was great interest in applying for jobs at FHI whenever they were announced, we often felt the candidates were too focused on narrow topics and too uninterested in linking up their research with other research. To remedy this, in 2018 the Research Scholars Programme launched, led by Owen Cotton-Barratt. This allowed promising young researchers to work for two years at the Institute, learning and contributing to the broad intellectual milieu. In 2018 we launched our DPhil scholarship program, providing funding to promising PhD students at Oxford whose work aligned closely with FHI's core areas of focus, and encouraging them to contribute to our

research. We welcomed scholars from nearly every academic division in the university, pursuing degrees in law, engineering, computer science, clinical medicine, zoology, to name just a few.

In many respects, the years just before the pandemic would mark a high point for FHI. At our peak, the Institute reached around 40 staff, alongside a rotating cast of interns, summer fellows, and academic visitors. Desks were crammed into every conceivable space with increasing ingenuity; research meetings were standing room only; the tiny lunchroom was filled with exciting discussions over vegan burritos and the great whiteboards showed ever changing mental maps. I often failed at reaching my office on arrival because I got dragged into research conversations. That the operations staff managed to keep things on track as well as they did was a miracle.

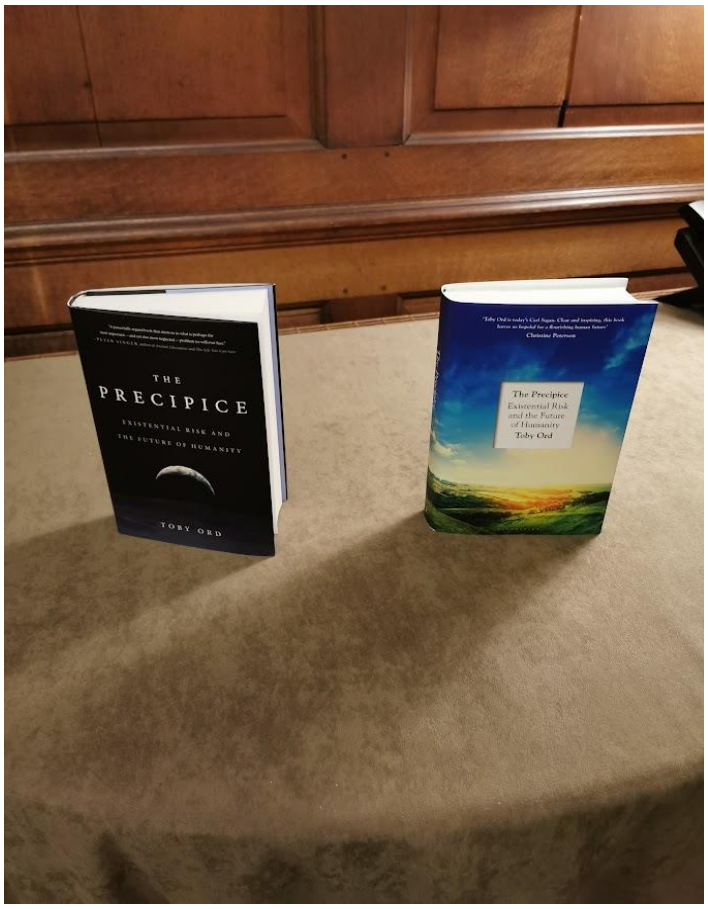


October 2019 floorplan of FHI in Littlegate House, at the point of maximal density. Beside the "locals" there were almost always visitors working in the main meeting rooms.

During this period the amount of policy work and its impact became larger. The world was increasingly coming around to recognize the relevance of some of the research topics. In the UK, numerous FHI staff have been invited to present to the UK parliament: Nick Bostrom, Toby Ord, Allan Dafoe, Gregory Lewis, Cassidy Nelson, Owen Cotton-Barratt, Michael Osborne, and Michael Cohen. Nick Bostrom was invited to address the UN 70th General Assembly to share expertise on challenges of international security and

emergence of AI. Allan Dafoe presented to the Sub-Committee for Security and Defence of the European Parliament.

Final Years



The Precipice (American and UK editions), at the release event two weeks before lockdown. March 6 2020.

In early 2020, just as Toby Ord's book *The Precipice* was set to launch, the biosafety team began to take note of an emerging outbreak of a coronavirus, and warned us that it was a good idea to prepare to work from home. This proved to be good foresight. During the SARS-CoV-2- pandemic the institute left its old premises in Littlegate House and moved into Trajan House, although few had expected how long it would take before people could return to the office. The institute still functioned, and the biosafety team worked on medical practicalities, attempts at improving forecasting and uncertainty management in the pandemic, and the wider implications of the pandemic for lab safety and global stability. FHI researchers set up the Epidemic Forecasting Project[24], providing bespoke policy analysis for officials from several governments, and reaching 10,000 users per day.

FHI also began a program on digital minds led by Robert Long, starting to explore the implications of AI sentience for ethics and policy. Anders Sandberg began to expand macrostrategy research into a Grand Futures program, attempting to put rigorous limits on what future intelligence can achieve and estimating how far we currently are from these limits.

[24] https://web.archive.org/web/20200430201156/http://epidemicforecasting.org/

FHI kept on producing policy outputs. The pandemic had made global risks far more salient to decision-makers, and past work showed FHI to be a useful partner. In 2021, FHI co-wrote the *Future Proof* report on the UK's resilience strategy[25], mentioned favorably in the Paymaster General's speech. Toby Ord advised the UN Secretary General's Office on existential risk and future generations, and contributed to the UN's 2020 Human Development Report. His book was quoted by the UK Prime Minister in his speech to the UN General Assembly. Several junior researchers undertook secondments in government to advise on technology policy.

While FHI had achieved significant academic and policy impact, the final years were affected by a gradual suffocation by Faculty bureaucracy. The flexible, fast-moving approach of the institute did not function well with the rigid rules and slow decision-making of the surrounding organization. (One of our administrators developed a joke measurement unit, "the Oxford". 1 Oxford is the amount of work it takes to read and write 308 emails. This is the actual administrative effort it took for FHI to have a small grant disbursed into its account within the Philosophy Faculty so that we could start using it - *after* both the funder and the University had already approved the grant.)

Starting in 2020, the Faculty imposed a freeze on fundraising and hiring. Unfortunately, this led to the eventual loss of lead researchers and especially the promising and diverse cohort of junior researchers, who have gone on to great things in the years since. While building an impressive alumni network and ecosystem of new nonprofits, these departures severely reduced the Institute. In late 2023, the Faculty of Philosophy announced that the contracts of the remaining FHI staff would not be renewed. On 16 April 2024, the Institute was closed down.

Nick Bostrom has stated: "I wish it were possible to convey the heroic efforts of our core administrative team that were required to keep the FHI organizational apparatus semi-performant and dynamic for all those years until its final demise! It is an important part of the story. And the discrepancy between the caliber of our people and the typical university administrators - like Andrew carpet bombing his intray with pomodoros over the weekends, or Carrick coming over with a J.D. from Yale Law School to volunteer to help out with initially the lowest-level office tasks, or Tanya putting in literal 21 or 22 hour workdays (!) for weeks at an end. Probably not even our own researchers fully appreciate what went on behind the scenes."

---

[25] Ord, Toby; Mercer,Angus; Dannreuther, Sophie; Belfield, Haydn; Whittlestone, Jess; Leung, Jade; Anderljung, Markus; Nelson, Cassidy; Lewis, Gregory; Millett, Piers; Hilton, Sam (2021) *Future Proof: the opportunity to transform the UK's resilience to extreme risks*. The Centre for Long-Term Resilience. https://www.cser.ac.uk/resources/future-proof/

Research topics: what we found out

A few of FHIs deliverables in book form. 2014.

Over its history FHI researched many topics, and added and removed them depending on how useful further work appeared to be for the future of humanity.

In many cases the work was exploratory: checking if a topic was even amenable to investigation, demonstrating some nontrivial results to prove this and entice further work, and - if the topic appeared important - starting the process of turning it into a generally accepted concept. It is worth noting that many of the topics discussed below once elicited sneers of derision since they were so obviously "out there" ... yet some years later had become part of the intellectual Overton window.

While the list below may look like claiming credit, for many of the topics FHI was not the only early actor: we maintained a broad network of intellectual associates, many of whom contributed the lion's share of field building. One of the key roles of FHI was to be a stable organizational home for investigation, often visited by researchers in these nascent fields for discussion and networking.

Existential Risk

The Petrov seminar room at Littlegate House was named after Stanislav Petrov, who is credited with preventing an accidental nuclear war between the Soviet Union and the US 26 September 1983. Across the hallway was the Arkhipov room, named after Vasili Arkhipov, who prevented a nuclear launch from submarine B-59 during the Cuban Missile Crisis.

One of the core FHI topics throughout its existence was existential risk: ways humanity can fail by going prematurely extinct or lose its long-term potential. This leads to numerous important questions: What makes existential risk bad? What kinds of trans-generational high-severity risks exist? How can these questions usefully be approached from an axiological, ethical, epistemic and practical standpoint? How likely are different risks? How does one prioritize between them? What policies can be implemented to reduce them?

While the threat of human extinction and its weightiness had been considered before, FHI pioneered the systematic study and field-building of existential risk (and the closely related global catastrophic risks). It

began in Bostrom's early papers[26], then was solidified in the 2008 book *Global Catastrophic Risks*[27] and Bostrom's 2013 paper on existential risk as a global priority[28], his Vulnerable World Hypothesis paper (2019), followed by much research. It can be said to have culminated at FHI in Toby Ord's *The Precipice* (2020)[29], the first book-length treatment of existential risk for a wide audience, which in turn has influenced policy in the United Kingdom and at the UN.

Some papers and themes worth mentioning:

An early paper by Nick Bostrom and Max Tegmark[30] demonstrated a way of bounding a speculative risk (indeed, an entire vast category of unknown risks) and how a corrected bound could be established by properly taking into account observation selection effects. The same basic idea was generalized and applied to a wider set of cases in a second paper by Bostrom, Ćirković, and Sandberg, which introduced the concept of "anthropic shadow"[31] (an important potential bias that will systematically distort our estimates of many global catastrophic risks unless corrected for). Later research applied this to risks of nuclear war.

Andrew Snyder-Beattie, Toby Ord and Michael Bonsall estimated an upper bound on the background rate of human extinction, from natural hazards, given paleontological evidence, making a strong case that anthropogenic risk is higher[32]. This estimation was again deeply informed by the work on observer selection FHI had pursued over the years. The recognition that anthropogenic existential risk appears to represent most of the existential risk we face has significant implications for policy.

Owen Cotton-Barratt and Toby Ord worked on refining the concept of existential risk beyond mere extinction, elucidating how it was tied to the expectation of future value (and opening the door for the concept of existential hope, that the future can be better than expected)[33].

---

[26] Bostrom, N. (2002). Existential risks: Analyzing human extinction scenarios and related hazards. *Journal of Evolution and technology*, *9*. https://nickbostrom.com/existential/risks.pdf

[27] Bostrom, N., & Cirkovic, M. M. (Eds.). (2011). *Global catastrophic risks*. Oxford University Press, USA.

[28] Bostrom, N. (2013). Existential risk prevention as global priority. *Global Policy*, *4*(1), 15-31.

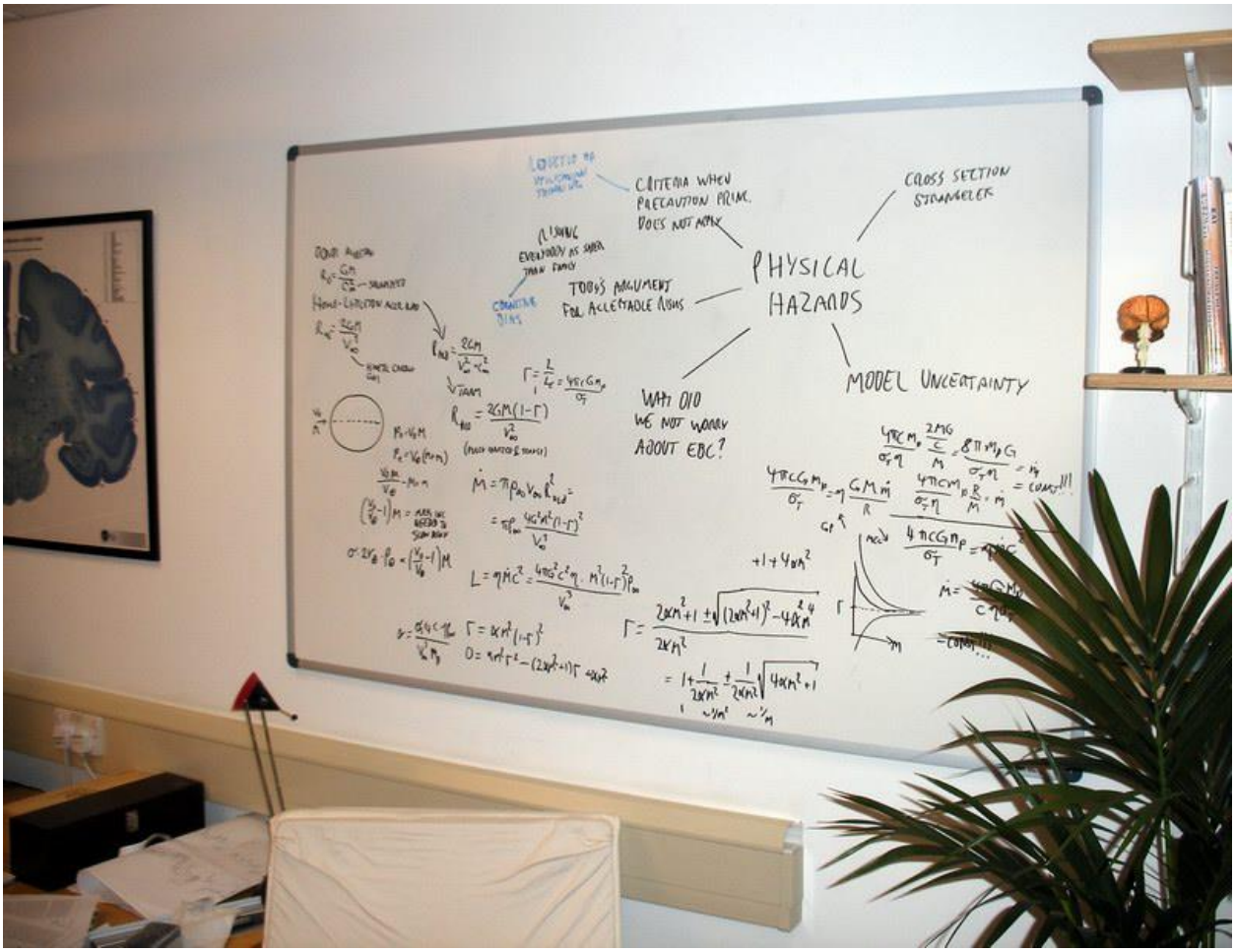
[29] Ord, T. (2020) *The Precipice: Existential Risk and the Future of Humanity*, (London: Bloomsbury).

[30] Tegmark, M., & Bostrom, N. (2005). Is a doomsday catastrophe likely?. *Nature*, *438*(7069), 754-754.

[31] Ćirković, M. M., Sandberg, A., & Bostrom, N. (2010). Anthropic shadow: observation selection effects and human extinction risks. *Risk Analysis: An International Journal*, *30*(10), 1495-1506.

[32] Snyder-Beattie, A. E., Ord, T., & Bonsall, M. B. (2019). An upper bound for the background rate of human extinction. *Scientific reports*, *9*(1), 11054.

[33] Cotton-Barratt, O., & Ord, T. (2015). Existential risk and existential hope: definitions. *Future of Humanity Institute Report*, #2015-1.



Anders' whiteboard when estimating constraints on strangelet and black hole accretion rates inside planets. April 2008.

Toby Ord, Anders Sandberg and Rafaela Hillerbrand analyzed the problem that when attempting to bound low probability severe risk the chance of errors in theory, modeling or calculations become larger than the probability that is being estimated, leading to subtle methodological challenges that need handling[34]. This was applied to the hypothetical risks of particle accelerators possibly creating planet-threatening phenomena, showing that previous risk analysis was somewhat flawed and suggesting how future versions could be made more watertight.

[34] Ord, T., Hillerbrand, R., & Sandberg, A. (2010). Probing the improbable: methodological challenges for risks with low probabilities and high stakes. *Journal of Risk Research*, *13*(2), 191-205.

Nick Bostrom proposed the "Vulnerable World Hypothesis" in a widely debated paper[35], arguing that there may exist some level of technological development at which civilizations almost certainly get devastated by default unless they achieve the right kind of global coordination.

Leopold Aschenbrenner (then interning at FHI) and Philip Trammell explored how economic growth interacts with existential risk and risk reduction using economic modeling[36], opening for further research in how to strategize risk reduction on a global scale.

While much of this was theory, it fed an interest in investigating the particular issues of different kinds of risk (see below). It also led to work on how to improve humanity's resilience against risks and the governance of global risks, both on a theoretical level[37] and practical policy work. For example, a workshop organized by the Global Priorities Project and the Finland Ministry for Foreign Affairs joined by the FHI produced the report "Existential Risk: Diplomacy and Governance"[38]. Anders Sandberg joined the board of the non-profit ALLFED[39] to support work on making the global food system resilient to the largest disasters. FHI staff contributed and produced numerous policy reports intended to help governments manage large risks sensibly[40].

---

[35] Bostrom, N. (2019). The vulnerable world hypothesis. *Global Policy*, *10*(4), 455-476.

[36] Aschenbrenner, L. (2020). Existential risk and growth. Global Priorities Institute, GPI Working Paper, 0-84.

[37] Cotton-Barratt, O., Daniel, M., & Sandberg, A. (2020). Defence in depth against human extinction: Prevention, response, resilience, and why they all matter. *Global Policy*, 11(3), 271-282. ; Fisher, L., & Sandberg, A. (2022). A safe governance space for humanity: necessary conditions for the governance of global catastrophic risks. *Global Policy*, 13(5), 792-807.

[38] Farquhar, Sebastian; Halstead, John; Cotton-Barratt, Owen; Schubert, Stefan; Belfield, Haydn; Snyder-Beattie, Andrew (2017). "*Existential Risk: Diplomacy and Governance*". Global Priorities Project https://web.archive.org/web/20170211130009/https://www.fhi.ox.ac.uk/wp-content/uploads/Existential-Risks-2017-01-23.pdf

[39] https://allfed.info/

[40] Some examples are: Beckstead, N.; Bostrom, N.; Bowerman, N.; Cotton-Barratt, O.; MacAskill, W.; Ó hÉigeartaigh, S. & Ord, T.(2014) Unprecedented technological risks. Future of Humanity Institute report 2014. http://amirrorclear.net/files/unprecedented-technological-risks.pdf ; Becksted, N. & Ord, T. (2014) Managing existential risk from emerging technologies. In M Walport (ed.) Annual report of the government chief scientific advisor 2014. Innovation: Managing Risk, Not Avoiding It. Evidence and Case Studies. (London: Government Office for Science), 115–120, 2014. ; Ord, T. (2020) Existential risks to humanity. In P Conceicao (ed.) Human Development Report 2020. (New York: The United Nations Development Programme), 106–11. 2020. ; Ord, T.; Mercer, A.; Dannreuther, S.; Belfield, H.; Whittlestone, J.; Leung, J.; Anderljung, M.; Nelson, C.; Lewis, G.; Millett, P.; Hilton, S.(2021) Future Proof: the opportunity to transform the UK's resilience to extreme risks. The Centre for Long-Term Resilience. https://www.cser.ac.uk/resources/future-proof/ ; Ord, T. (2021). Proposal for a New 'Three Lines of Defence'Approach to UK Risk Management. Future of Humanity Institute report #2021-1.

Biological Risk

Andrew Snyder-Beattie representing the FHI at the UN Biological Weapons Convention, Geneva, December 2017.

The threat of natural or artificial biological risks was an active topic within the larger global catastrophic risk and technology regulation themes. FHI had an active team working on biological risks, and was often engaged with policymakers (including the UN Convention on Biological Weapons), law enforcement, and biotech enthusiasts.

Much concern focused on emerging biotechnologies that threaten a massive democratization of the ability to create pathogens. One way of managing this issue may be to enhance the ability of attribution of novel pathogens to laboratories of origin[41].

[41] Lewis, G., Jordan, J. L., Relman, D. A., Koblentz, G. D., Leung, J., Dafoe, A., ... & Inglesby, T. V. (2020). The biosecurity benefits of genetic engineering attribution. *Nature communications*, 11(1), 6294.

The theoretical research on information hazards was applied to how to mitigate the information risks emerging from the dual-use nature of biotechnology[42]. A related paper by Anders Sandberg and Cassidy Nelson analyzed the issue of what kind of threat actors pose the most risk as technology advances[43]

During the SARS-CoV-2-pandemic the team worked on many aspects of mitigating the pandemic. Among the contributions, of particular note is the influential evaluation of non-pharmaceutical interventions against the pandemic by Jan Brauner et al.[44] becoming one of the ten most widely referenced papers in the history of the journal *Science*. While the status of the SARS-CoV-2 lab leak hypothesis remains murky, a dataset of the numerous historical escapes of deadly pathogens from laboratories was compiled by David Mannheim and Gregory Lewis[45] helping to inform debate and policy.

Macrostrategy

Macrostrategy can be described as "the study of how long-term outcomes for humanity may be connected to present-day actions" (Bostrom) or "investigating which crucial considerations are shaping what is at stake for the future of humanity" (Sandberg). This has been an important concept underpinning FHI's work. FHI was based on the idea that such long-term outcomes could be investigated with some rigor, and that such investigations could achieve decision-relevant insights. Investigating to what extent these assumptions were true and useful were a major long-term focus, although the term macrostrategy only emerged in the later years.

A foundational (and widely cited) pre-FHI paper was Nick Bostrom's 2003 Astronomical Waste paper[46]. In it he outlined an argument that under aggregative consequentialism or utilitarianism the vast "cosmic endowment" of resources that may produce value seem to imply a strong reason for immediate, rapid expansion across the universe to employ them. However, considering that existential risks may destroy humanity before it reaches the chance to do so, from this aggregative perspective it instead becomes even

---

[42] Lewis, G., Millett, P., Sandberg, A., Snyder-Beattie, A., & Gronvall, G. (2019). Information hazards in biotechnology. *Risk Analysis*, 39(5), 975-981.

[43] Sandberg, A., & Nelson, C. (2020). Who should we fear more: biohackers, disgruntled postdocs, or bad governments? A simple risk chain model of biorisk. *Health security*, 18(3), 155-163.

[44] Brauner, J. M., Mindermann, S., Sharma, M., Johnston, D., Salvatier, J., Gavenčiak, T., ... & Kulveit, J. (2021). Inferring the effectiveness of government interventions against COVID-19. *Science*, 371(6531), eabd9338.

[45] Manheim, D., & Lewis, G. (2021). High-risk human-caused pathogen exposure events from 1975-2016. F1000Research, 10.

[46] Bostrom, N. (2003). Astronomical waste: The opportunity cost of delayed technological development. *Utilitas*, 15(3), 308-314.

more important to reduce the risks to the future - our early position in history and the cheapness of time in the large implies that we ought to take far more care about risk.

One part of the macrostrategy approach was interest in decision-making about the future. What information and actions truly matter, and how should one handle profound ignorance about key facts? One guiding concept was *crucial considerations*[47]: those considerations that radically change the expected value of pursuing different courses of action, like realizing that one has been holding a map upside down or neglecting the moral importance of a large group of entities. Much of FHI macrostrategic work has consisted of trying to discover new crucial considerations.

Another abstract topic in macrostrategy is how to think of "the whole future": is it better seen as a sum of individual events and states, or better regarded as trajectories through the space of possibilities? A joint paper on the topic was the shared outcome of Anders Sandberg's tenure as distinguished chair at the Gothenburg Centre for Advanced Study in 2017, reviewing the case for thinking in terms of trajectories[48]. Toby Ord developed a general framework for how one could compare aggregated futures and trajectories[49].

The macrostrategy strand has worked on investigating civilizations (human or alien) as complex physical phenomena. Beside existential risks, are there "natural" limits to the stability of civilizations and other complex social organizations? These considerations matter both for direct risk, but also for what future projects can reliably be achieved. FHI researchers have explored theoretical strategies for indefinite survival through backups (implying that a surprisingly slow growth of backups can allow indefinite survival)[50], and the direct empirical question of whether there is any form of "aging" of societies[51].

---

[47] https://nickbostrom.com/lectures/crucial-considerations/

[48] Baum, S. D., Armstrong, S., Ekenstedt, T., Häggström, O., Hanson, R., Kuhlemann, K., ... & Yampolskiy, R. V. (2019). Long-term trajectories of human civilization. *Foresight*, 21(1), 53-83.

[49] Ord, T. (2023). Shaping humanity's longterm trajectory. In *Essays on Longtermism*, Jacob Barrett, Hilary Greaves & David Thorstad (eds), Oxford University Press.

[50] https://ora.ouls.ox.ac.uk/objects/uuid:81611331-c1fd-44c1-a3f0-b42637d3a260

[51] Sandberg, A. (2023). The Lifespan of Civilizations: Do Societies "Age," or Is Collapse Just Bad Luck?. In *How Worlds Collapse* (pp. 375-396). Routledge.

Longtermism

Anders Sandberg lecturing at the CNIO-Banc Sabadell Foundation Workshop on Philosophy, November 2022.

One hallmark of FHI thinking has been considerations of the very long-run future. If one takes a time-neutral perspective the importance of safeguarding the wellbeing of future people becomes as salient as safeguarding present people, an ethical view that has become known as Longtermism.

At FHI the Astronomical Waste paper gave the impetus for investigating just how vast the future could be (the start of Grand Futures exploration) but also debate about how to properly discount the future[52] and how to handle the often troubling implications of a longtermist perspective.

[52] Ord, T. *The Precipice* introduces longtermism pp. 43-46, and Appendix A discusses the discounting issue.

Anders Sandberg, Stuart Armstrong and Milan Ćirković bridged longtermism, grand futures, and the SETI research strand with the "Aestivation hypothesis"[53], an analysis based on the physics of information processing and megascale engineering, showing the benefits for advanced civilizations to use an extreme long-term strategy. This ties in closely with Toby Ord's work on shaping the long-term trajectory: is it better to speed up or delay civilizational activities, given the different tradeoffs due to risks and the physical properties of the universe?

Grand Futures

The grand futures research programme explored visions for what a spacefaring civilization could achieve. Part of the focus was on the future of humanity, sketching a picture of what we could eventually achieve were we to take to the stars and reach our full potential. Part of the focus was on spacefaring civilizations in general, theorizing about alien civilizations, the generic impact of intelligence on the physical structure of the universe at the large scale, and what we can learn from the absence of any evidence of such civilizations. The aim was to use solid arguments from physics, game theory, and other disciplines to find bounds on the attainable.

One of the key insights was the role of cosmological models for limiting what can be known and how many material resources (and hence survival) are available to future intelligence[54]. They also affect whether the longest-term future is offense or defense dominant, with implications for macrostrategy and survival.

Most of the results are forthcoming in Anders Sandberg's as yet unpublished book *Grand Futures*.

[53] Sandberg, A., Armstrong, S., & Cirkovic, M. M. (2016). That is not dead which can eternal lie: the aestivation hypothesis for resolving Fermi's paradox.*Journal of the British Interplanetary Society*, vol. 69, p. 406-415

[54] Ord, T. (2021). The edges of our universe. arXiv preprint arXiv:2104.01191.

Extraterrestrial Intelligence

Eric Drexler having tea while analyzing models for abiogenesis and the Fermi Paradox. Feb 23 2016

While the name Future of Humanity Institute may suggest that only humanity is the topic of the institute, we have also explored the question of what other intelligence may exist in the universe. One reason is that the presence or absence of extraterrestrial intelligence (ETI) gives information about our own future chances at avoiding existential risk[55], as per the Great Filter argument of FHI associate Robin Hanson[56]. The "Great Silence" is a profound interdisciplinary problem with deep philosophical and scientific roots, as explored by Milan Ćirković (another long-time FHI associate) in several books *(The astrobiological landscape, The Great Silence*)[57] and joint papers.

[55] Bostrom, N. (2008) Why I Hope the Search for Extraterrestrial Life Finds Nothing. *MIT Technology Review*, May/June issue (2008): pp. 72–77

[56] Hanson, R. (1998). The great filter-are we almost past it. preprint available at https://mason.gmu.edu/~rhanson/greatfilter.html

[57] Ćirković, M. M. (2012). *The astrobiological landscape*: Philosophical foundations of the study of cosmic life (Vol. 7). Cambridge University Press. ; Ćirković, M. M. (2018). *The great silence: Science and philosophy of Fermi's paradox*. Oxford University Press.

One approach is to use exploratory engineering to find bounds on how ETI can spread in the universe, finding that past analysis may have severely underestimated the severity of the Fermi Paradox[58], and the Fermi Paradox may give information about human technology potential[59]. Another model is how information physics may motivate extreme longtermism with possibly observable effects[60].

However, a more careful analysis of the uncertainties involved in estimating the density of ETI suggests a "dissolution" of the Fermi Paradox: given current uncertainties it is hard to rule out a very empty universe even if one starts with optimistic assumptions about the probability of life, intelligence and longevity of ETI[61]. This also makes the Great Filter bite less severely, and motivates more diligent search for independently evolved life in the solar system.

Using FHIs observer selection research, Andrew Snyder-Beattie et al. showed that given the assumption that there have been hard transitions in evolution and the timing of the transitions of Earth's biosphere gives a high probability to a fairly empty universe[62]. Building on previous considerations in self-selection, interstellar expansion and transitions, Robin Hanson et al. formulated the innovative "grabby aliens hypothesis"[63] suggesting that the universe is about halfway through a phase transition of settlement.

---

[58] Armstrong, S., & Sandberg, A. (2013). Eternity in six hours: Intergalactic spreading of intelligent life and sharpening the Fermi paradox. *Acta Astronautica*, 89, 1-13.

[59] Olson, S. J., & Ord, T. (2021). Implications of a search for intergalactic civilizations on prior estimates of human survival and travel speed. arXiv preprint arXiv:2106.13348.

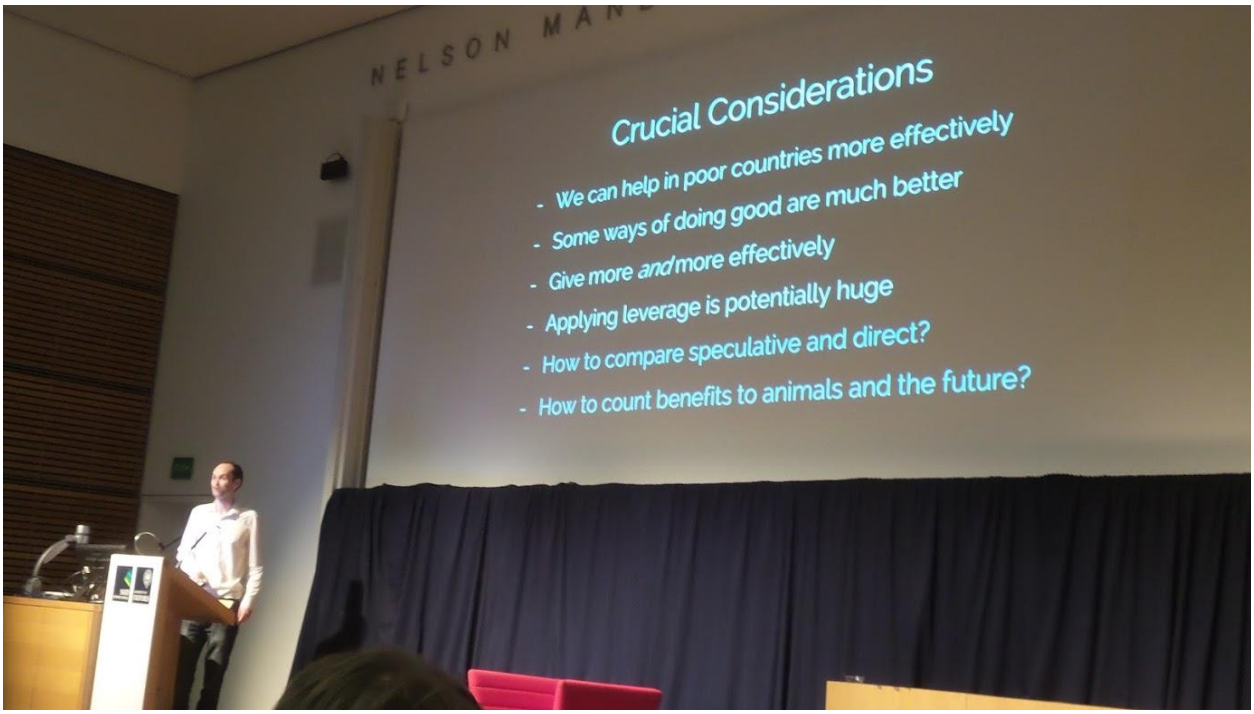
[60] Sandberg, A., Armstrong, S., & Cirkovic, M. M. (2016). That is not dead which can eternal lie: the aestivation hypothesis for resolving Fermi's paradox.*Journal of the British Interplanetary Society*, vol. 69, p. 406-415

[61] Sandberg, A., Drexler, E., & Ord, T. (2018). Dissolving the Fermi paradox. arXiv preprint arXiv:1806.02404.

[62] Snyder-Beattie, A. E., Sandberg, A., Drexler, K. E., & Bonsall, M. B. (2021). The timing of evolutionary transitions suggests intelligent life is rare. *Astrobiology*, 21(3), 265-278.

[63] Hanson, R., Martin, D., McCarter, C., & Paulson, J. (2021). If loud aliens explain human earliness, quiet aliens are also rare. *The Astrophysical Journal*, 922(2), 182.

Effective Altruism



Toby Ord lecturing at Effective Altruism 2015.

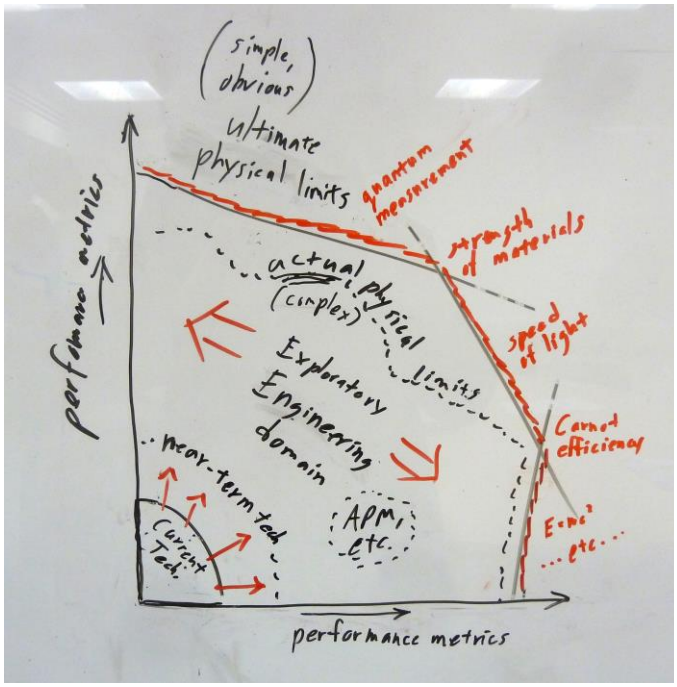
Effective altruism is a research field and practical community that tries to find the best ways of doing good in the world and put these into practice. Its theory was heavily informed by ideas in theoretical ethics, practical ethics, and economics — with a focus on finding rigorous tools of analysis for finding better ways of doing good. These threads were drawn together in Oxford through the work of Toby Ord and William MacAskill. A key early paper was Toby Ord's 'The moral imperative towards cost-effectiveness in global health'[64], which showed how different programmes for improving the health of people in poor countries varied in cost-effectiveness by many orders of magnitude, such that failure to prioritize could easily result in losing 99% of the benefits one could have achieved. Further work

[64] Ord, T. (2013) . The Moral Imperative toward Cost-Effectiveness in Global Health. Reprinted in Greaves and Pummer, eds. *Effective Altruism: Philosophical Issues*, (Oxford: OUP), 29–36, 2019.

included applications to charity[65], questions of whether effective altruism is too demanding[66], and a book length treatment[67].

There was substantial cross-pollination between the ideas and methods of effective altruism and FHI. Effective altruism was influenced by many FHI ideas, including macrostrategy, existential risk, and taking humanity's position in the universe seriously. Effective altruism in turn influenced FHI's approach to choosing research topics — selecting for those where we could make the greatest counterfactual impact.

Technology



Eric Drexler's diagram of technology space, with accumulated annotations. November 2013.

FHI was always interested in emerging technology, due to its potential to affect the human condition directly or indirectly, cause or mitigate risk, as well as represent a challenging case of unpredictability. Overall, artificial intelligence dominated as a topic (see below), but FHI has pioneered investigation of Whole Brain Emulation (WBE) and often looked at other fields. We were in particular interested in understanding the capabilities of potential technologies that have not yet been developed, using methods such as exploratory engineering to get bounds on their capabilities.

Eric Drexler, the father of nanotechnology and a pioneer of exploratory engineering, joined the FHI 2011 and contributed greatly not just to considerations of molecular manufacturing but to general

[65] Ord, T. (2014). Global poverty and the demands of morality. In J Perry (ed.) *God, The Good, and Utilitarianism: Perspectives on Peter Singer*, (Cambridge: CUP), 177–91, 2014.

[66] MacAskill, W., Mogensen, A., and Ord, T. (2018). Giving isn't demanding. In Woodruff, ed. *The Ethics of Giving: Philosophers' Perspectives on Philanthropy,* (Oxford: OUP), 178–203, 2018.

[67] MacAskill, W. *Doing Good Better*. (2015). (New York: Gotham Books).

thinking about engineering, manufacturing, and technological systems. His book *Radical Abundance* (2013) sets out an updated strategy for achieving the as-yet unrealized potential for atomically precise manufacturing. Given recent advances in protein engineering and AI the importance of the field may well grow faster than many expect.

In 2016 we held a workshop on emerging cryptographic technologies (including Vitalik Buterin, Wei Dai and Gwern Branwen) exploring the implications of blockchain and other cryptographic technologies on AI, confidential computation and contracts[68]. In the debate on privacy versus surveillance[69] an emerging idea was the possibility of structured transparency: technology can, when applied right, improve the governability of information flows and thereby incentivize collaboration.[70]

AI Risk and Alignment

Artificial Intelligence (AI) has been a major focus since the earliest days. At first mostly seen as a potential radical amplifier of human capability the risks soon came into the forefront, as well as the problem of how to usefully predict where the emerging field was going.

In the early days, AI was not seen as progressing very fast by most mainstream thinkers and FHI was somewhat unusual in thinking it could become transformative in the not-too-far future and that it was worthwhile examining the most radical possibilities (even if they were in the long-term future) due to the potential existential risks. This exploration was vindicated to some extent by the surprising (even to field insiders) developments in machine learning in the 2010s, and perhaps even more by the dramatic growth of AI capabilities in the 2020s.

One topic was timelines: are there ways of estimating how far away transformative capabilities are in time? Nick Bostrom had a pre-FHI 1997 paper that predicted a 2004-2024 range of when enough computing power for human-level AI would become available[71]. This kind of compute-limited forecasting benefits from the relative predictability of Moore's law and its derivatives, but is sensitive to assumptions about brain computational power, and might fail if algorithmic improvements become significant. Another approach is expert judgment. FHI pioneered the practice of systematically surveying

---

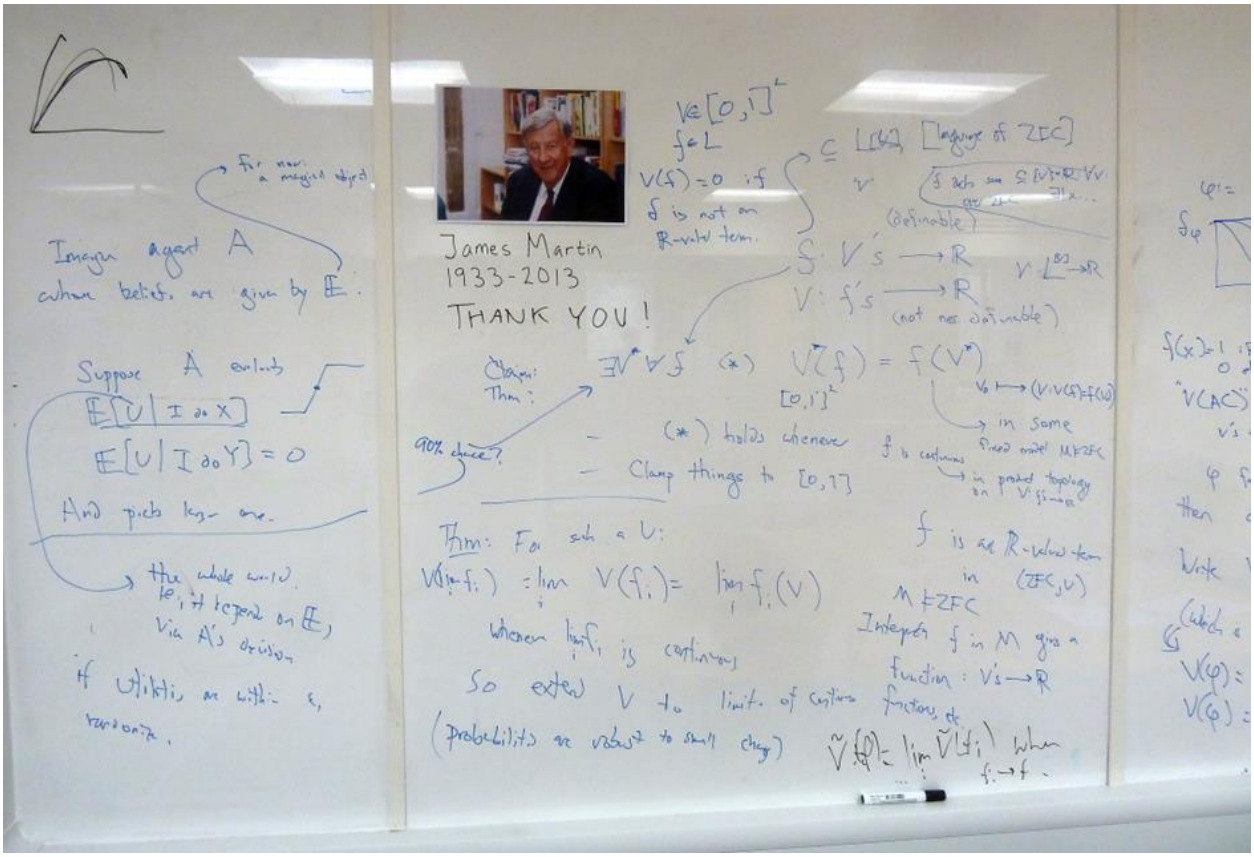
[68] Ben Garfinkel. Cryptographic Technologies: What They Are and How They Could Matter. Centre for the Governance of AI, Future of Humanity Institute, University of Oxford. May 2021

[69] Garfinkel, Ben (2020) "The Case for Privacy Optimism", Talk delivered at Google DeepMind.

[70] Trask, A., Bluemke, E., Garfinkel, B., Cuervas-Mons, C. G., & Dafoe, A. (2020). Beyond privacy trade-offs with structured transparency. arXiv preprint arXiv:2012.08347.

[71] Bostrom, N. (1998). How long before superintelligence. *International Journal of Futures Studies*, 2(1), 1-9. See https://nickbostrom.com/superintelligence for a version with updated postscripts till 2008.

expert communities on their beliefs about future developments in AI[72], something that is now common practice. Overall, the main finding is that there is little consensus (and, given known limitations on expertise, good reason to doubt individual and aggregate views as being accurate forecasts). Still, surveying the attitudes and values of the AI field may itself be a valuable data point even if one doubts its ability to forecast well, and later surveys have typically found shorter and shorter timelines[73].



Typical state of the main whiteboard (in this case dealing with self-reflexive AI). We decided that the best way to recognize James Martin was to always have his picture in the middle of our work.

Regardless of the speed of arrival, the issue of how controllable superintelligent systems could be emerged as a key topic. FHI formalized many discussions in the nascent AI safety community, for

[72] Sandberg, A., & Bostrom, N. (2011). Machine intelligence survey. FHI Technical Report, 2011-1. https://web.archive.org/web/20140409031128/http://www.fhi.ox.ac.uk/machine-intelligence-survey-2011.pdf ; Müller, V. C., & Bostrom, N. (2016). Future progress in artificial intelligence: A survey of expert opinion. Fundamental issues of artificial intelligence, 555-572. ; Grace, K., Salvatier, J., Dafoe, A., Zhang, B., & Evans, O. (2018). When will AI exceed human performance? Evidence from AI experts. *Journal of Artificial Intelligence Research*, 62, 729-754.

[73] E.g. see Grace, K., Stewart, H., Sandkühler, J. F., Thomas, S., Weinstein-Raun, B., & Brauner, J. (2024). Thousands of AI authors on the future of AI. arXiv preprint arXiv:2401.02843.

example the orthogonality thesis and the instrumental convergence thesis[74]. These theses suggest that there is no guarantee of beneficence or even moral behavior from a highly intelligent system, and that there are reasons to believe that dangerous behaviors are a common outcome if systems are not designed to avoid them. Hence risk became a major topic[75].

The research from the AGI study group at FHI led to Nick Bostrom's 2014 book *Superintelligence: Paths, Dangers, Strategies.* This book, which became a New York Times bestseller, helped spark a global conversation on the future of AI which continues to this date. In addition to laying out the case for AI risk, it also presented a rich framework for thinking about the impacts of superintelligence and for the impacts of transformative technologies more generally and the complex macrostragic considerations we face when trying to figure out how to make a positive difference.

Participants at the January 2016 FLI AI conference in Puerto Rico. A sizable number of FHI members are mixed in among the AI celebrities.

[74] Bostrom, N. (2012). The superintelligent will: Motivation and instrumental rationality in advanced artificial agents. *Minds and Machines*, 22, 71-85. ; Armstrong, S. (2013). General purpose intelligence: arguing the orthogonality thesis. *Analysis and Metaphysics*, (12), 68-84.

[75] Müller, V. C. (2014). Risks of general artificial intelligence. *Journal of Experimental & Theoretical Artificial Intelligence,* 26(3), 297-301.

FHI researchers explored different strategies for ensuring safe advanced AI, whether keeping them isolated[76], making them indifferent to certain facts about the world[77] , making them corrigible[78] and interruptible[79], learning values from humans using inverse reinforcement learning[80], human-in-the-loop reinforcement learning[81], as well as various theoretical approaches based on decision theory.

One strand of investigation by Eric Drexler criticized the agent-centric perspective of much AI safety research. In the influential *Reframing Superintelligence* (2019)[82] he showed that in many situations there is little need for an agent-like system with goals, and furthermore that alternative architectures are more similar to what engineering and corporate practices are likely to develop and value. A comprehensive AI services economy may reap the benefits of superintelligence without some of the risks linked to agential AI. Written just before the LLM revolution, it prefigured many of the later developments[83].

AI governance

In the 2010s, with the deep learning revolution underway and the shape of AI progress coming more clearly into view, questions around how humanity might govern the arrival of transformative technology became more pressing.

FHI noticed early on the dangers posed by 'race dynamics' between AI projects when there are trade-offs between safety and capabilities. Armstrong, Bostrom & Shulman's Racing to the Precipice sought to

---

[76] Armstrong, S., Sandberg, A., & Bostrom, N. (2012). Thinking inside the box: Controlling and using an oracle AI. *Minds and Machine*s, 22, 299-324.

[77] Armstrong, S. (2010). Utility indifference. FHI technical report 2010-1.

[78] Soares, N., Fallenstein, B., Armstrong, S., & Yudkowsky, E. (2015, April). Corrigibility. In Workshops at the Twenty-Ninth AAAI Conference on Artificial Intelligence.

[79] Orseau, L., & Armstrong, M. (2016, May). Safely interruptible agents. In *Conference on Uncertainty in Artificial Intelligence*. Association for Uncertainty in Artificial Intelligence.

[80] Evans, O., Stuhlmüller, A., Salvatier, J., & Filan, D. (2017). *Modeling agents with probabilistic programs*. URL: https://agentmodels.org/

[81] Abel, D., Salvatier, J., Stuhlmüller, A., & Evans, O. (2017). Agent-agnostic human-in-the-loop reinforcement learning. arXiv preprint arXiv:1701.04079. ; Saunders, W., Sastry, G., Stuhlmueller, A., & Evans, O. (2017). Trial without error: Towards safe reinforcement learning via human intervention. arXiv preprint arXiv:1707.05173.

[82] Drexler, K. E. (2019). *Reframing superintelligence*. Future of Humanity Institute. Technical report 2019-1.

[83] https://www.alignmentforum.org/posts/LxNwBNxXktvzAko65/reframing-superintelligence-llms-4-years

model this situation, showing that "winner-takes-all" situations might push actors to favor haste over prudence.[84]

Shahar Avin (CSER) led a design crunch at FHI that resulted in the game Intelligence Rising, aimed at educating about and exploring strategic AI futures. The Petrov Room was covered in post-it notes. September 2019.

2017 was a significant year. With the launch of FHI's Governance of AI program (led by Allan Dafoe) and the publication of the first research agenda,[85] the field had a name, a plan, and a home at FHI. A succession of important papers followed soon after.

[84] Armstrong, S., Bostrom, N., & Shulman, C. (2016). Racing to the precipice: a model of artificial intelligence development. AI & society, 31, 201-206.

[85] "AI Governance: A Research Agenda", Allan Dafoe (2018)

FHI organized a workshop on the potential for malicious uses of AI systems, culminating in an influential 2018 report.[86] Katja Grace's much-cited 2017 survey of AI experts found that they expected human-level AI within a few decades, and placed non-trivial probability on this leading to extremely bad outcomes (e.g., human extinction). Follow-up surveys were conducted in 2022 and 2023, and have shown increasingly shorter timelines to human-level AI.[87]

Malicious Use of AI workshop, 2017.

FHI was early to see the importance of international dimensions of AI competition. In 2018 Jeffrey Ding published a comprehensive report on China's AI capabilities, debunking several widespread misconceptions about US-China competition.[88]

While the focus was mainly on risk purely from AI systems with no human maliciousness, FHI did explore the topic of malicious AI use in an influential 2018 report. There were also several contributions

[86] Brundage, M., Avin, S., Clark, J., Toner, H., Eckersley, P., Garfinkel, B., ... & Amodei, D. (2018). The malicious use of artificial intelligence: Forecasting, prevention, and mitigation. arXiv preprint arXiv:1802.07228. https://maliciousaireport.com/

[87] 2022: https://aiimpacts.org/2022-expert-survey-on-progress-in-ai/
2023: Grace, K., Stewart, H., Sandkühler, J. F., Thomas, S., Weinstein-Raun, B., & Brauner, J. (2024). Thousands of AI authors on the future of AI. arXiv preprint arXiv:2401.02843.

[88] https://www.governance.ai/research-paper/deciphering-chinas-ai-dream-the-context-components-capabilities-and-consequences-of-chinas-strategy-to-lead-the-world-in-ai

to the ethics of AI[89] and AI development[90]. Cullen O'Keeffe and others made an influential proposal for a 'windfall clause', a mechanism to ensure the benefits of advanced AI are widely distributed amongst humanity rather than accruing to private actors.[91] Another strand of research looked to historical examples of transformative technology to draw lessons for AI governance.[92]

FHI/GPI meeting. L to R facing camera: Hilary Greaves, ???, Nick Bostrom, Teruji Thomas, Michelle Hutchinson, Anders Sandberg. Facing away: Owen Cotton-Barratt, Ben Garfinkel.

[89] Bostrom, N., & Yudkowsky, E. (2018). The ethics of artificial intelligence. In *Artificial intelligence safety and security* (pp. 57-69). Chapman and Hall/CRC.

[90] Prunkl, C. E., Ashurst, C., Anderljung, M., Webb, H., Leike, J., & Dafoe, A. (2021). Institutionalizing ethics in AI through broader impact requirements. *Nature Machine Intelligence*, 3(2), 104-110.

[91] O'Keefe, C., Cihon, P., Garfinkel, B., Flynn, C., Leung, J., & Dafoe, A. (2020, February). The windfall clause: Distributing the benefits of AI for the common good. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* (pp. 327-331). https://www.governance.ai/research-paper/the-windfall-clause-distributing-the-benefits-of-ai-for-the-common-good

[92] Zaidi, W., & Dafoe, A. (2021). International Control of Powerful Technology: Lessons from the Baruch Plan for Nuclear Weapons. *Centre for the Governance of AI*. https://cdn.governance.ai/International-Control-of-Powerful-Technology-Lessons-from-the-Baruch-Plan-Zaidi-Dafoe-2021.pdf ; Ord, T. (2022). Lessons from the development of the atomic bomb. Centre for the Governance of AI. https://cdn.governance.ai/Ord_lessons_atomic_bomb_2022.pdf

Nick Bostrom talking with Dragos Tudorache, the EU Chair of the Special Committee on Artificial Intelligence in the Digital Age (AIDA) and the LIBE rapporteur on the AI Act. 2024

One topic that has become hotly debated after ChatGPT is the role of openness in AI development. Is open research and technology sharing a way of reducing risk by reducing disparities and having more actors able to contribute, or does it spread dangerous capabilities? Nick Bostrom wrote a paper in 2017 that laid out some of the key issues.[93]

Since spinning out of FHI in 2021 to escape the strictures of university bureaucracy, the Centre for the Governance of AI (GovAI) has firmly established itself as one of the leading institutions in the field. In just a few years, AI governance has developed from a small team whose primary institutional foothold was in FHI, to a vibrant network of think tanks and research institutes around the world. Issues around

[93] Bostrom, N. (2017). Strategic Implications of Openness in Development. *Global Policy*, 8 (2), 135–148.

the governance of transformative AI, long regarded as an eccentric pursuit, are being taken seriously by world leaders and society at large.

Whole Brain Emulation

Whole Brain Emulation (WBE) is the potential approach to achieve software intelligence by copying the functional structure of biological nervous systems into software in an automated way. FHI has been pioneering in WBE field-building. In May 26-27 2007 FHI hosted a workshop that resulted in *Whole Brain Emulation: A Roadmap*.[94]

Subsequently, FHI researchers investigated issues of feasibility[95], ethics[96] and safety issues[97] of WBE. FHI associate Robin Hanson wrote *The Age of Em* (2016), exploring the implications of WBE given standard economic and sociological assumptions[98].

Outside FHI brain emulation related research has been pursued by computational neuroscientists, microscope technologists, and the connectomics field. Since 2007 the field has advanced significantly and a research community seeded by the roadmap is now increasingly pursuing the goal of translating scanned data into computational models. A new workshop held in Oxford 2023 by the Foresight Institute and FHI[99] stimulated research further. There are now ambitious projects related to the vision.

---

[94] Sandberg, A., & Bostrom, N. (2008). Whole brain emulation: A roadmap. Technical Report #2008-3, Future of Humanity Institute, Oxford University

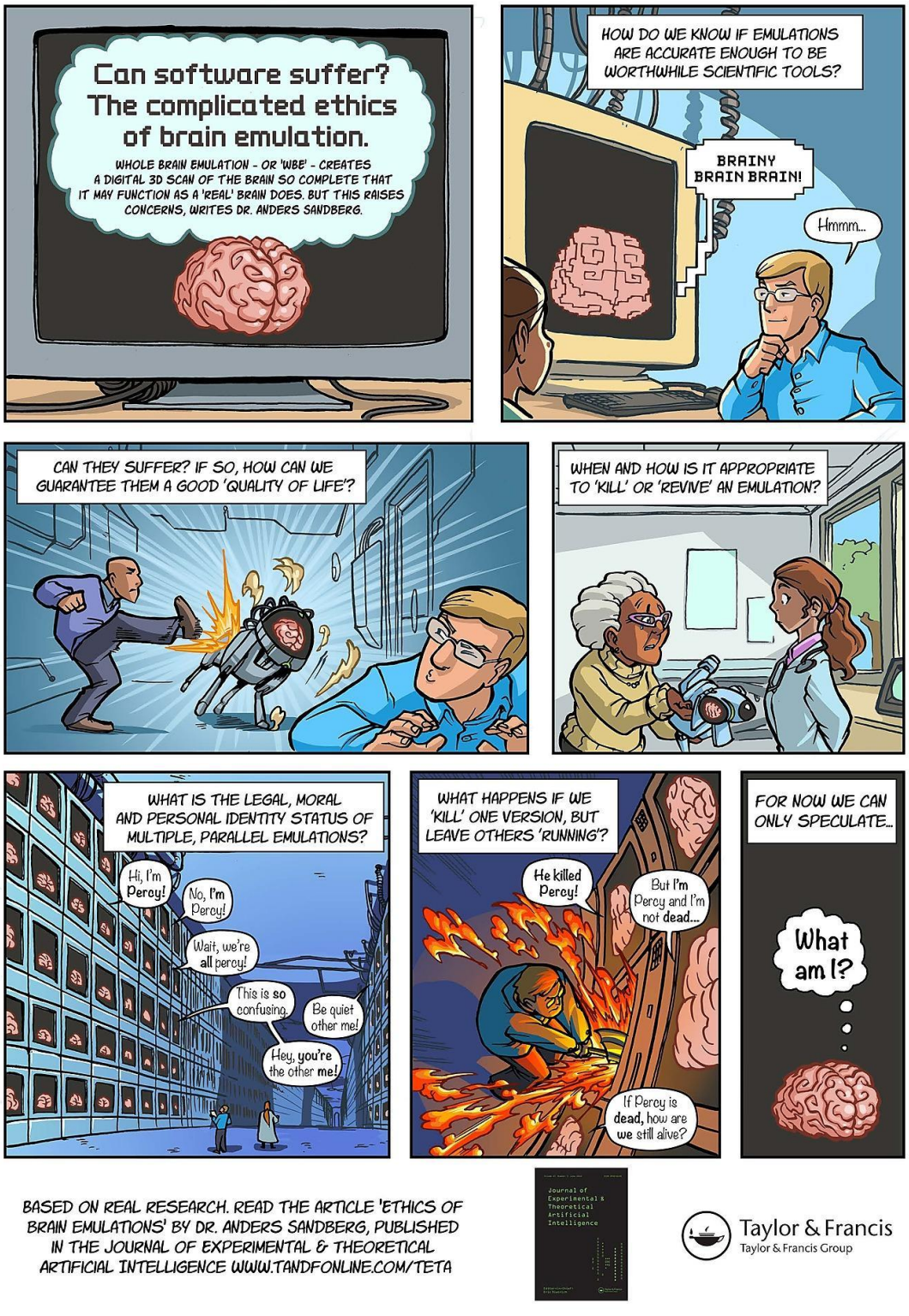
[95] Sandberg, Anders (2013), 'Feasibility of whole brain emulation', in Vincent C. Müller (ed.), *Theory and Philosophy of Artificial Intelligence* (SAPERE; Berlin: Springer), 251-64.

[96] Sandberg, A. (2014). Ethics of brain emulations. *Journal of Experimental & Theoretical Artificial Intelligence*, 26(3), 439-457.

[97] Eckersley, P., & Sandberg, A. (2013). Is brain emulation dangerous? *Journal of Artificial General Intelligence*, 4(3), 170-194.

[98] Hanson, R. (2016). *The age of Em: Work, love, and life when robots rule the earth*. Oxford University Press.

[99] https://foresight.org/whole-brain-emulation-workshop-2023/



Taylor & Francis made a cartoon abstract for the 2013 paper about ethics of brain emulations, featuring Anders in

various roles.

Digital Minds

A topic which has been of interest for a long time but became an increasing focus in the last few years of FHI is that of the ethics of digital minds—questions related to the potential moral patienthood of AIs and digitally instantiated intellects, and more broadly how to conceive of a positive and cooperative future that may include vast numbers of digital minds along with human minds and animal minds.

FHI has explored the moral and political status of digital minds[100]. A particular problem might be that the possibility that some digital minds could become "super-beneficiaries" and/or "super-patients"[101]. Bostrom and Shulman set out an ambitious attempt to stake out a preliminary comprehensive position concerning digital minds and society. It touches on issues ranging from the metaphysics of consciousness, the moral status of AIs, security and social stability, persuasion and social manipulation, epistemology, and much else, with the aim of adumbrating a general framework that could enable a future to go well for both its digital and biological constituencies.[102]

An earlier paper, from 2006[103], investigated the metaphysical question of what happens to mental experience (from a computationalist perspective) when the same computation is run multiple times—and discovered an interesting non-trivial structure: the answer depends on how the multiple instantiations are implemented. Also concluded that there is a coherence notion of "fractional" amounts of experience of qualitatively identical content:

The issue of what kinds of systems can be moral patients may hinge on whether they are conscious or not. In a major paper (even covered in the *New York Times*[104]) Patrick Butlin et al. described a rigorous and empirically grounded approach to AI consciousness based on assessing existing AI systems in detail in light of our best-supported neuroscientific theories of consciousness[105].

---

[100] Bostrom, N., & Yudkowsky, E. (2011). The ethics of artificial intelligence. *The Cambridge handbook of artificial intelligence*.

[101] Shulman, C., & Bostrom, N. (2021). Sharing the world with digital minds. In *Rethinking moral status*, 306-326.

[102] Bostrom, N., & Shulman, C. (2022). Propositions Concerning Digital Minds and Society. Nick Bostrom's Webpage, 1, 1-15.

[103] Bostrom, N. (2006). Quantity of experience: brain-duplication and degrees of consciousness. *Minds and Machines*, 16, 185-200.

[104] https://www.nytimes.com/2023/09/18/science/ai-computers-consciousness.html

[105] Butlin, P., Long, R., Elmoznino, E., Bengio, Y., Birch, J., Constant, A., ... & VanRullen, R. (2023). Consciousness in artificial intelligence: insights from the science of consciousness. arXiv preprint arXiv:2308.08708.

Human Enhancement Ethics

Elise Bohan and Anders Sandberg discuss transhumanism and the ethics of enhancement in a panel with professor Steve Fuller and Luke Robert Mason (outside frame) 16 February 2023.

FHI ran a long-running investigation into the ethics and possible social impact of human enhancement. This was a natural fit given the interest in any technology that could change the human condition.

During and after the ENHANCE project FHI researchers produced many widely cited papers, such as Nick Bostrom and Anders Sandberg's overview of the field of cognitive enhancement[106], Bostrom and

[106] Bostrom, N., & Sandberg, A. (2009). Cognitive enhancement: methods, ethics, regulatory challenges. *Science and engineering ethics*, 15, 311-341.

Rebecca Roache's overview of the ethical issues[107], the review by Sandberg et al. of non-pharmacological cognitive enhancement[108], Bostrom and Sandberg's evolutionary heuristic for decisionmaking about enhancement[109], and Bostrom and Ord's reversal test for removing status quo bias in bioethical decisionmaking[110]. Several volumes collecting the research have become standard reference works[111]. Other paper explored key foundational ethical questions like the normativity of memory modification[112] or how powerful genetic selection could be in cognitive enhancement[113].

At the end of the 2000s, FHI gradually focused more on other topics. One reason was priority: existential risk and AI appeared more urgent and high-impact. Another was that by this time the academic debate on enhancement ethics had largely settled down; not so much into a consensus, but into a number of relatively fixed positions.

Sandberg, together with researchers at the Oxford Uehiro Centre for Practical Ethics, opened up investigations into love enhancement[114]. Recent results in the neuroscience of pair bonding gave rise to a fruitful topic where plausible near-future technology could have direct effect on romantic attachment,

---

[107] Bostrom, N., & Roache, R. (2008). Ethical issues in human enhancement. *New waves in applied ethics*, 120-152.

[108] Dresler, M., Sandberg, A., Ohla, K., Bublitz, C., Trenado, C., Mroczko-Wąsowicz, A., ... & Repantis, D. (2013). Non-pharmacological cognitive enhancement. *Neuropharmacology*, 64, 529-543.

[109] Bostrom, N., & Sandberg, A. (2009). The wisdom of nature: an evolutionary heuristic for human enhancement. In *Human enhancement*, eds. Julian Savulescu and Nick Bostrom. Oxford, Oxford University Press, pp. 375-416.

[110] Bostrom, N., & Ord, T. (2006). The reversal test: eliminating status quo bias in applied ethics. *Ethics*, 116(4), 656-679.

[111] Savulescu, J. & Bostrom N. (eds) (2009) *Human Enhancement.* Oxford University Press.

Savulescu, J., ter Meulen, R., & Kahane, G. (2011). *Enhancing Human Capacities*. Wiley.

[112] Liao, S. M., & Sandberg, A. (2008). The normativity of memory modification. *Neuroethics*, 1, 85-99.

[113] Shulman, C., & Bostrom, N. (2014). Embryo Selection for Cognitive Enhancement: Curiosity or Game-changer?. *Global Policy*, 5(1), 85-92.

[114] Savulescu, J., & Sandberg, A. (2008). Neuroenhancement of love and marriage: The chemicals between us. *Neuroethics*, 1, 31-44.

Earp, B. D., Wudarczyk, O. A., Sandberg, A., & Savulescu, J. (2013). If I could just stop loving you: Anti-love biotechnology and the ethics of a chemical breakup. *The American Journal of Bioethics*, 13(11), 3-17.

posing interesting challenges to the individual-centered typical approaches to human enhancement, how to socially frame such technology[115], and where the limits of medicalization lie[116].

From the FHI perspective, the current human species and condition is not unchanging nor the only or best state of being. While we were generally optimistic that were humanity to survive it could develop into wondrous and desirable posthumanities[117], the possibility of losing value or potential is a real risk requiring careful consideration of future trajectories[118].

Applied Epistemology, Rationality and Decision-theory

Studying the future raises challenging questions about methodology: what is the epistemic status of claims about the future? How to deal with different forms of uncertainty? When trying to act rationally about potential future events, what are the best practices? How severely does cognitive bias prevent us from doing the right thing?

Already before FHI Nick Bostrom had been working on issues related to observer selection effects - when the existence of an observer may bias observations and predictions in nontrivial ways, "anthropic bias"[119]. Over the history of the institute FHI researchers explored different ways of interpreting or controlling for such bias, especially in regard to existential risk and extraterrestrial intelligence.

FHI and associates were involved in the question of how to improve forecasting since the early days. Robin Hanson was an early proponents of prediction markets[120]. FHI alumnus Jason Matheny began

---

[115] Earp, B. D., Sandberg, A., & Savulescu, J. (2014). Brave new love: The threat of high-tech "conversion" therapy and the bio-oppression of sexual minorities. *AJOB neuroscience*, 5(1), 4-12.

[116] Earp, B. D., Sandberg, A., & Savulescu, J. (2015). The medicalization of love. *Cambridge Quarterly of Healthcare Ethics*, 24(3), 323-336.

[117] Bostrom, N. (2005). In defense of posthuman dignity. *Bioethics*, 19(3), 202-214. ; Bostrom, N. (2008). Letter from utopia. Studies in *Ethics, Law, and Technology*, 2(1).; Bostrom, N. (2013). Why I want to be a posthuman when I grow up. *The transhumanist reader: Classical and contemporary essays on the science, technology, and philosophy of the human future*, 28-53.

[118] Bostrom, N. (2004). The future of human evolution. Death and anti-death: Two hundred years after Kant, fifty years after Turing, 339-371.

[119] Bostrom, N. (2013). Anthropic bias: Observation selection effects in science and philosophy. Routledge.

[120] Arrow, K. J., Forsythe, R., Gorham, M., Hahn, R., Hanson, R., Ledyard, J. O., ... & Zitzewitz, E. (2008). The promise of prediction markets. *Science*, 320(5878), 877-878.

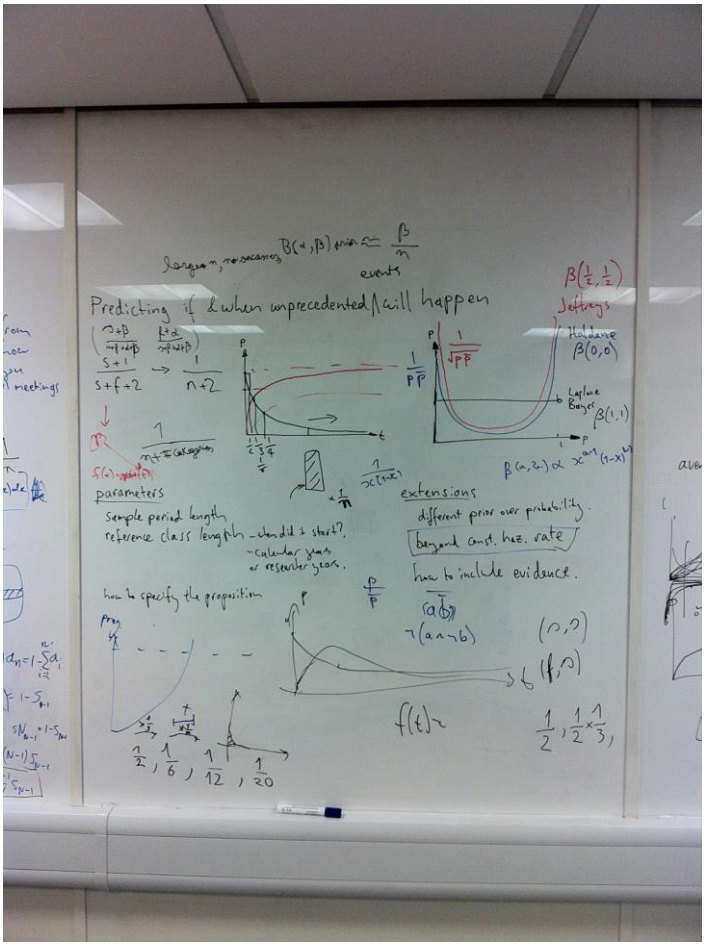
forecasting tournaments, eventually leading to "superforecasting", which is now used to understand disagreements on AI risk[121].

On the other hand, not all information is helpful. FHI also explored information hazards[122], true but in some way harmful information, and how it might be managed[123]. FHI researchers have helped other organizations consider their information hazard policies and their approach to specific information hazards.

[121] https://forecastingresearch.org/news/results-from-the-2022-existential-risk-persuasion-tournament

[122] Bostrom, N. (2011). Information hazards: A typology of potential harms from knowledge. *Review of Contemporary Philosophy*, (10), 44-79.

[123] Lewis, G., Millett, P., Sandberg, A., Snyder-Beattie, A., & Gronvall, G. (2019). Information hazards in biotechnology. *Risk Analysis*, 39(5), 975-981.



Predicting unprecedented events using non-informative priors, 2014.

Another counterintuitive result was the "unilateralist curse": situations where well-intentioned agents might still act against their own and everybody's interest[124]. This curse strikes when it is enough that anyone performs a certain action for it to have full consequences for everyone (e.g. releasing a potential information hazard, performing geoengineering), and judgments about its net utility are uncertain. Then the more agents there are, the more risk for an action with negative consequences since each agent can make a mistake. To counteract this requires adopting a principle of conformity, or preferably, constructing institutions that can coordinate the actions.

One area of interest was reasoning under extreme uncertainty. Estimating the time to or probability of unprecedented events, and the cost-effectiveness of working on problems of unknown difficulty[125] were practically relevant issues for FHI strategy. Could one use the Lindy Effect — the tendency for things with longer pasts behind them to have longer futures ahead — to make estimates? Toby Ord explained and formalized the effect[126]. This work was also closely related to observer selection effects and anthropic reasoning, ideas that offer tantalizing hope of reducing uncertainty in many such situations but are also philosophically controversial.

---

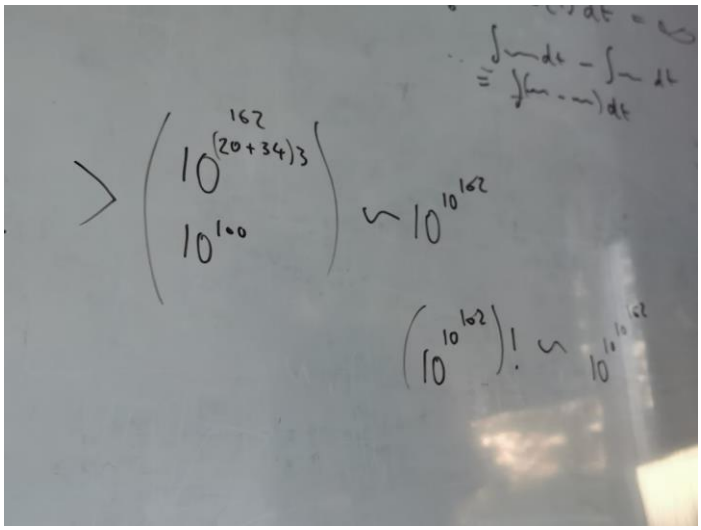
[124] Bostrom, N., Douglas, T., & Sandberg, A. (2016). The Unilateralist's Curse and the Case for a Principle of Conformity. *Social epistemology*, 30(4), 350-371.

[125] https://web.archive.org/web/20240407050810/https://www.fhi.ox.ac.uk/how-to-treat-problems-of-unknown-difficulty/ ; https://web.archive.org/web/20240407050809/https://www.fhi.ox.ac.uk/estimating-cost-effectiveness/

[126] Ord, T. (2023). The Lindy Effect. arXiv preprint arXiv:2308.09045.

While most of this work was done in the abstract, FHI did look for applications. In the FHI-Amlin collaboration on systemic risk we explored how the use of models in reinsurance might cause systemic risks by acting as hidden correlations between company actions[127], how biased error-checking could produce subtly biased models[128], underwriter bias, and various ways the insurance industry could improve model diversity[129]. The ERC UNPREDICT project examined alternatives to the precautionary principle in cases of strong uncertainty and heuristics useful for policy making.

Ethics



Some of the numbers that show up in FHI research were unusually large even by scientific standards. In this case it was a rough comparison of bounds on the number of possible futures, leading to a profound challenge to value theory. 2023, Toby Ord's whiteboard.

As mentioned earlier, FHI always regarded itself as doing a form of applied ethics by evaluating future possibilities and whether current decisions could affect them. Still, these investigations often led to questions reaching into metaethics and more theoretical considerations of consequentialism.

Nick Bostrom opened up the exploration of the problems of "infinite ethics": the consequences of aggregative consequentialism seem to be paradoxical in the context of infinitely large sets of possibilities[130]. This topic has hence been investigated by others inside and outside FHI, finding a rich vein of complexity.

Another problem in consequentialism is that low-probability, high (but finite) value situations might produce apparently irrational decision-

[127] Sandberg, A. & Zhou, F. (2016) *Meta-modelling of the systemic effects of catastrophe modelling model diversity*, report of the FHI-Amlin collaboration.

[128] Sandberg, A. Beckstead, N., & Armstrong, S. (2014) *Systemic Risk of Modelling*, report from the FHI-Amlin Systemic Risk of Risk Modelling Collaboration. Oxford University.

[129] (2016) *Did your model tell you all models are wrong*? White paper by the Systemic Risk of Modelling Working Party..

[130] Bostrom, N. (2011). Infinite ethics. *Analysis and Metaphysics*, (10), 9-59.

making, "Pascal's mugging" [131]. This has obvious implications for the rationality of investigations into exotic existential risks.

One common challenge in applying ethics in real life is normative disagreement. How to act given that not just the public and the philosophy community disagree on what the proper course of action is, but also that oneself is uncertain about what moral framework is correct? Nick Bostrom proposed a parliamentary model for resolving the issue[132]. Further work inspired by this model and exploring reasoning and decision-making under moral uncertainty have been produced[133], culminating in the 2020 book *Moral Uncertainty* by Will MacAskill, Krister Bykvist and Toby Ord[134].

One innovative use of the moral uncertainty proposed by Toby Ord[135] is moral trade, where agents who disagree on their moral frameworks can make mutually beneficial trades that make both parties feel that the world is a better place or that their moral obligations are better satisfied.

---

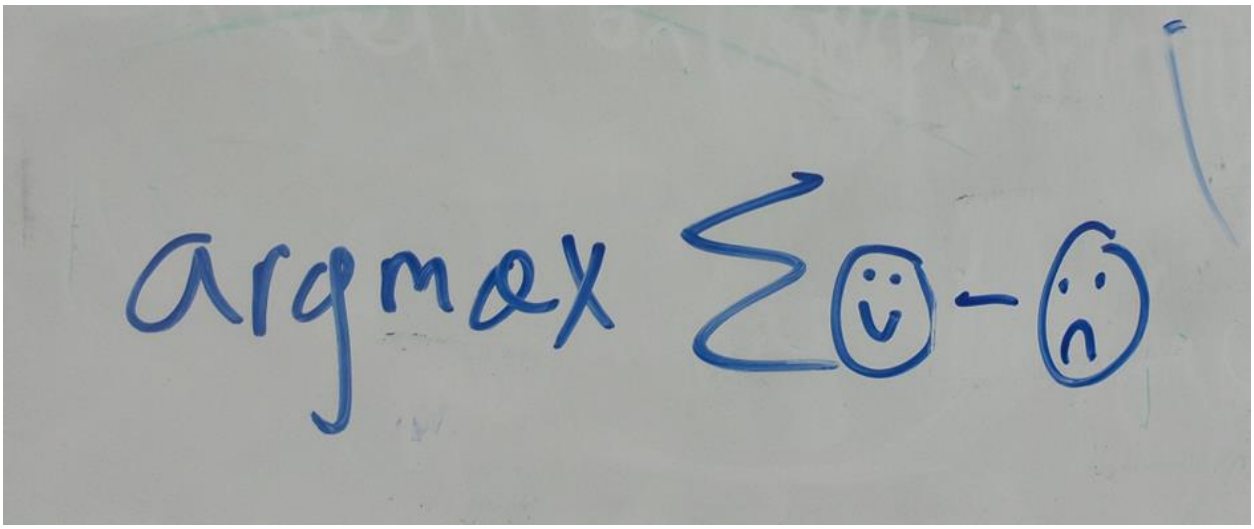
[131] Bostrom, N. (2009). Pascal's mugging. Analysis, 69(3), 443-445.

[132] https://www.overcomingbias.com/p/moral-uncertainty-towards-a-solutionhtml ; Newberry, T., & Ord, T. (2021). The parliamentary approach to moral uncertainty. Future of Humanity technical report #2021-2.

[133] MacAskill, W., & Ord, T. (2018). Why maximize expected choice-worthiness, *Noûs*. 54(2), 327-353. ; Cotton-Barratt, O., MacAskill, W., & Ord, T. (2020). Statistical normalization methods in interpersonal and intertheoretic comparisons. *Journal of Philosophy*, 117(2). ; Greaves, H., & Ord, T. (2017). Moral uncertainty about population axiology. *J. Ethics & Soc. Phil*., 12, 135.

[134] MacAskill, M., Bykvist, K., & Ord, T. (2020). *Moral uncertainty*. Oxford University Press.

[135] Ord, T. (2015). Moral trade. *Ethics*, 126(1), 118-138.

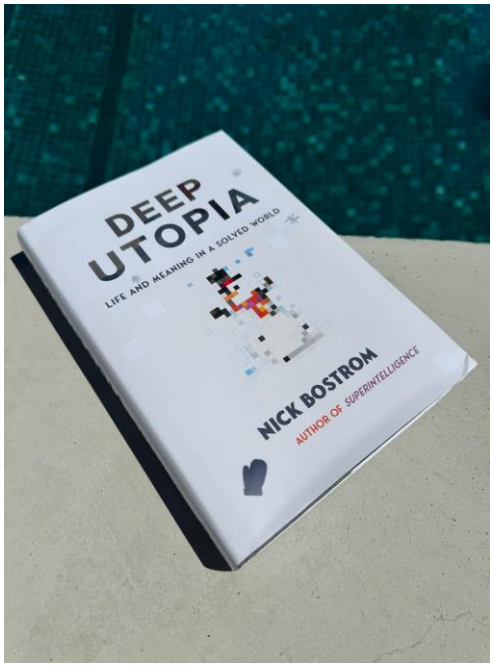


Equation found on the whiteboard. Let us maximize the amount of good (integrated across the future light-cone of humanity) minus the amount of bad. Of course, some further filling in of detail is needed...

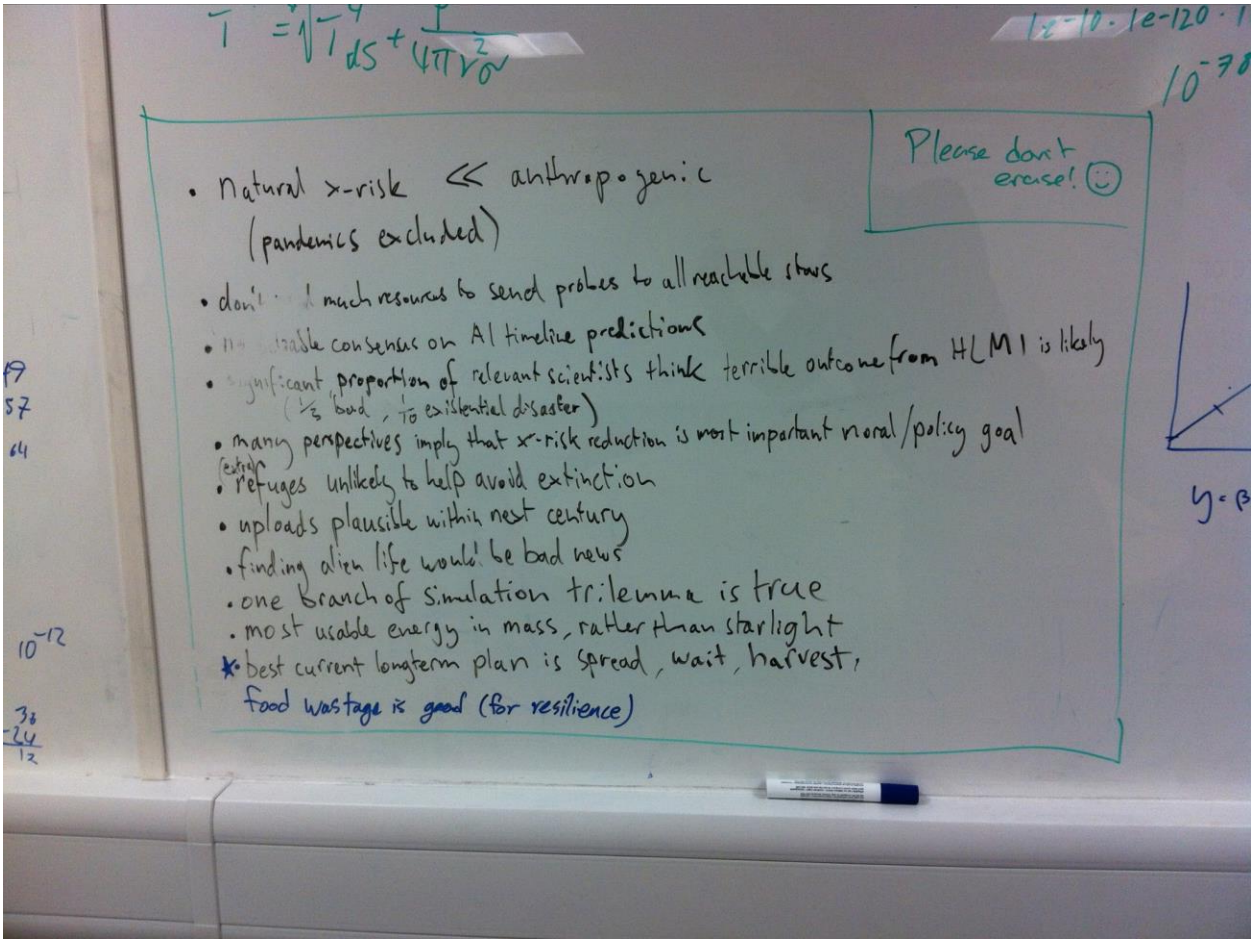
Deep Utopia

As of this writing, the book *Deep Utopia: Life and Meaning in a Solved World* by Nick Bostrom has just been released. In a sense it represents the opposite of existential risk, existential hope, but also the challenges that hope entails: how can humanity find meaning if things go well? Is this state worth aiming for?

It may be a hopeful coda to the research at FHI.



Concepts



Some conclusions of FHI research, 2014

One of the most striking things about FHI's legacy is how many basic concepts originally invented at FHI have since become almost commonplace in contemporary discourses about humanity's future and macrostrategic analyses.

Some of the foundational concepts perhaps worth mentioning are: existential risks, singletons[136], astronomical waste, information hazards[137], the "unilateralist's curse"[138], differential technological development[139], crucial considerations[140], exploratory engineering, whole brain emulation, macrostrategy, grand futures, structured transparency, as well as numerous concepts in technical AI safety.

Quantitative methods for qualitative insight: one useful method FHI championed was to use (approximate) quantitative methods to inform qualitative analysis. As demonstrated in the Astronomical Waste paper, if one can show that one factor is many orders of magnitude larger than another factor even significant uncertainty is not enough to make the conclusion non-robust. Many FHI projects and papers exploited quantitative modeling for approaching qualitative challenges in ethics, macrostrategy and epistemology.

This is closely related to exploratory engineering, the use of standard engineering and physics principles to describe technology that does not exist (and may never exist), yet were it to exist we have good confidence that it would have the estimated properties. Exploratory engineering, pioneered at FHI by Eric Drexler, gives lower bounds on what future technology can achieve.

One guiding concept was the insight that in many domains, importance has a heavy-tailed distribution. The most important action may have a value many times the second most important action. This means that prioritizing the right action becomes far more important and worth spending effort on (up to half of the value difference between the best and second-best action) than is intuitively obvious. Also, detecting whether a situation has such a skew importance distribution or not becomes a priority.

---

[136] Bostrom, N. What is a Singleton? *Linguistic and Philosophical Investigations*, Vol. 5, No. 2 (2006): pp. 48-54

[137] Bostrom, N. (2011). Information hazards: A typology of potential harms from knowledge. *Review of Contemporary Philosophy*, (10), 44-79.

[138] Bostrom, N., Douglas, T., & Sandberg, A. (2016). The Unilateralist's Curse and the Case for a Principle of Conformity. *Social epistemology*, 30(4), 350-371.

[139] Originally introduced in (Bostrom 2002) and discussed in *Superintelligence*, later extended in Sandbrink, Jonas and Hobbs, Hamish and Swett, Jacob and Dafoe, Allan and Sandberg, Anders, Differential technology development: An innovation governance consideration for navigating technology risks (September 8, 2022). Available at SSRN: https://ssrn.com/abstract=4213670 or http://dx.doi.org/10.2139/ssrn.4213670

[140] https://nickbostrom.com/lectures/crucial-considerations/

Outreach and Popular Science

Article in Nikkei about existential risk research. Diagram based on loose sketch by Anders Sandberg. Note the references to FHI in the text. April 2015.

Unlike many parts of the humanities division, FHI never had a problem with outreach or capturing the interest of the public. The topics FHI dealt with naturally interested people despite their sometimes outré appearance. Indeed, we often were concerned with too much media exposure - it cost time, attention, and could sometimes lead to cascades of misunderstanding of our work[141]. Some FHI members were also

[141] British tabloids being what they are, we gave rise to some amusing headlines: "WAKE UP LAZY 'BOTS! Robotic aliens are in hibernation and preparing for a dramatic return to our universe, boffins say"

https://www.thesun.co.uk/tech/3709158/robotic-aliens-are-hibernating-and-could-return-to-our-universe-one-day-oxford-boffins-say/ , "Wanted: three boffins to save the world from the 'AI apocalypse"

https://www.telegraph.co.uk/men/thinking-man/wanted-three-boffins-to-save-the-world-from-the-ai-apocalypse/, "Human extinction fears as Oxford boffin says world-changing AI could spell end of species"

https://www.express.co.uk/news/world/1727046/artificial-intelligence-doomsday-humanity-extinction-terminator , "BOMBSHELL REPORT Fears over Skynet-style AI that 'controls our NUKES' as Oxford prof warns rogue robots

enthusiastic about outreach, seeing it as both an academic obligation and a way to explore their research from unexpected angles.

There were many essays in *The Conversation*, *BBC Future*, *New Scientist*, *Quartz*, *Io9*, *Practical Ethics*, and other popular science journals by FHI members, and even more written based on interviews with us. There was also no shortage of popular science and overviews of the Institute and its members[142].

As for popular science books, Stuart Armstrong wrote *Smarter Than Us: The Rise of Machine Intelligence* (2014) and Elise Bohan, *Future Superhuman: Our transhuman lives in a make-or-break century* (2022). Of course, one can argue that *Superintelligence* and *The Precipice* also count.

Anders Sandberg was a mainstay in UK media whenever issues of human enhancement, cryonics or extraterrestrial intelligence came up, and was a regular on many panels at the *How The Light Gets In* philosophy festival. He also contributed advice and ideas to the National Geographic *Year Million* television series[143], the V&A museum exhibition "The Future Starts Here"[144], showed up as a talking head exhibit in the Welcome Collection 2012 exhibition "Superhuman", and in countless radio programs, TV shows, podcasts, art projects and media ventures.

pose 'greater risk than Covid-19'" https://www.thesun.co.uk/tech/science/15138658/skynet-ai-controls-nukes-oxford-prof-robots-covid-19/

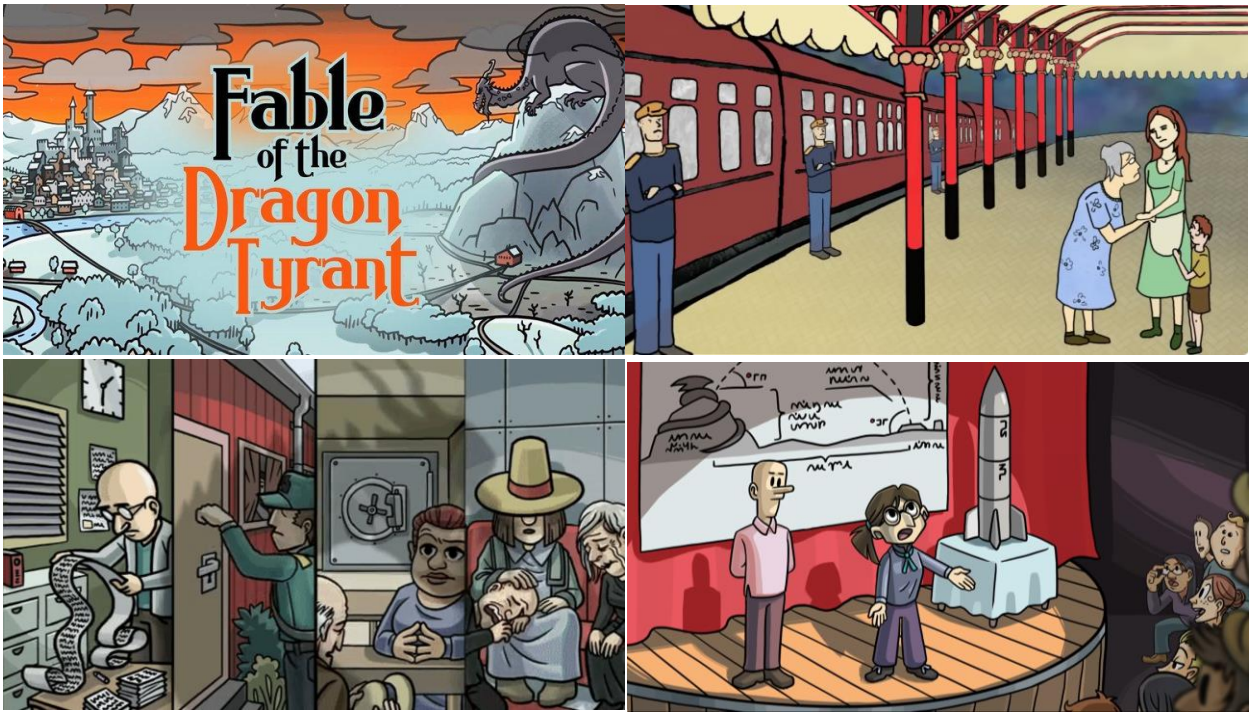
[142] Too many to mention, but one that comes to mind is Katchadourian, R. (2015). The Doomsday Invention: Will Artificial Intelligence Bring Us Utopia or Destruction?. *New Yorker*, *23*, 64-79. https://www.newyorker.com/magazine/2015/11/23/doomsday-invention-artificial-intelligence-nick-bostrom

[143] https://www.imdb.com/title/tt6945708/

[144] https://www.vam.ac.uk/exhibitions/the-future-starts-here

FHI's work also inspired a great deal of "grassroots" interest and cultural production. To mention just a few examples, Bostrom's Philosophical Quarterly paper that introduced the simulation argument inspired an off-Broadway play, *The World of Wires*, by Jay Scheib (2012) and *L'Anomalie* (The Anomaly), novel by Hervé Le Tellier (2020), which won the prestigious Prix Goncourt and became an international bestseller.

His *Fable of the Dragon-Tyrant*, a paper published in the *Journal of Medical Ethics* (2005) spawned a considerable amount of grassroots cultural production, including music tracks, artworks, and several animated cartoon retellings.



Animations of the Fable of the Dragon Tyrant by CGP Grey and Roz Francis

Indeed, several FHI papers have been turned into popular science presentation animations by the Rational Animations and Kurzgesagt YouTube channels. For example, "How to Take Over the Universe (in Three Easy Steps)"[145] describes the Sandberg and Armstrong "Eternity in Six Hours" scenario (with cartoon portraits of the authors advising the adorable would-be universal conquerors).

[145] https://www.youtube.com/watch?v=fVrUNuADkHI

A running joke around the institute was that certain webcomics creators[146] and scriptwriters must have placed bugs in our offices. The reality was, of course, that the Institute and its research was not a separate ivory tower but densely interconnected with the outside world. We acted as public intellectuals and spread ideas through all available media - and were often delighted by seeing echoes of our ideas showing up in unexpected places.

SMBC, 2013-04-23 (punchline cropped out). About a year before Superintelligence.

[146] Zach Weinersmith, it is time to confess!

Learnings

David Denkenberger and the ALLFED team discuss global food crises such as nuclear winter with FHI. October 2021

What we did well

One of the most important insights from the successes of FHI is to have a long-term perspective on one's research. While working on currently fashionable and fundable topics may provide success in academia, aiming for building up fields that are needed, writing papers about topics before they become cool, and *staying in the game* allows creating a solid body of work that is likely to have actual meaning and real-world effect.

The challenge is obviously to create enough stability to allow such long-term research. This suggests that long-term funding and less topically restricted funding is more valuable than big funding.

Many academic organizations are turned towards other academic organizations and recognized research topics. However, pre-paradigmatic topics are often valuable, and relevant research can occur in non-university organizations or even in emerging networks that only later become organized. Having the courage to defy academic fashion and "investing" wisely in such pre-paradigmatic or neglected domains (and networks) can reap good rewards.

Having a diverse team, both in terms of backgrounds but also in disciplines, proved valuable. But this was not always easy to achieve within the rigid administrative structure that we operated in. Especially senior hires with a home discipline in a faculty other than philosophy were nearly impossible to arrange. Conversely, by making it impossible to hire anyone not from a conventional academic background (i.e., elite university postdocs) adversely affects minorities, and resulted in instances where FHI was practically blocked from hiring individuals from under-represented groups. Hence, try to avoid credentialist constraints.

In order to do interdisciplinary work, it is necessary to also be curious about what other disciplines are doing and why, as well as to be open to working on topics one never considered before. It also opens the surface to the rest of the world. Unusually for a research group based in a philosophy department, FHI members found themselves giving tech support to the pharmacology department; participating in demography workshops, insurance conferences, VC investor events, geopolitics gatherings, hosting artists and civil servant delegations studying how to set up high-performing research institutions in their own home country, etc. - often with interesting results.

It is not enough to have great operations people; they need to understand what the overall aim is even as the mission grows more complex. We were lucky to have had many amazing and mission-oriented people make the Institute function. Often there was an overlap between being operations and a researcher: most of the really successful ops people participated in our discussions and paper-writing. Try to hire people who are curious.

Where we failed

Any organization embedded in a larger organization or community needs to invest to a certain degree in establishing the right kind of social relationships to maintain this embeddedness. Incentives must be aligned, and both parties must also recognize this alignment. We did not invest enough in university politics and sociality to form a long-term stable relationship with our faculty.

There also needs to be an understanding of how to communicate across organizational communities. When epistemic and communicative practices diverge too much, misunderstandings proliferate. Several times we made serious missteps in our communications with other parts of the university because we misunderstood how the message would be received. Finding friendly local translators and bridgebuilders is important.

Another important lesson (which is well known in business and management everywhere outside academia) is that as an organization scales up it needs to organize itself differently. The early informal structure cannot be maintained beyond a certain size, and must be gradually replaced with an internal structure. Doing this gracefully, without causing administrative sclerosis or lack of delegation, is tricky and in my opinion we somewhat failed.

So, you want to start another FHI?

Did FHI become humanity's best effort at understanding and evaluating its own long-term prospects? We leave that to the future to evaluate properly, but we certainly think we did unexpectedly well for a "three-year project".

FHI is ending and we are sad to see it go. We think it could have achieved far more than it did, but circumstances made it impossible to continue. On the plus side, we know FHI did not live past its time. There is a real risk that organizations lose their mission and become self-perpetuating users of resources that could better be used for other things, preventing the flowering of the new.

In fact, as mentioned above, FHI has seeded a number of new organizations, fields and topics. In a biological or memetic sense, it would count as having had great fitness in propagating successors, although many are not FHI-like or have different goals.

What would it take to replicate FHI, and would it be a good idea? Here are some considerations for why it became what it was:

- Concrete object-level intellectual activity in core areas and finding and enabling top people were always the focus. Structure, process, plans, and hierarchy were given minimal weight (which sometimes backfired - flexible structure is better than little structure, but as organization size increases more structure is needed).
- Tolerance for eccentrics. Creating a protective bubble to shield them from larger University bureaucracy as much as possible (but do not ignore institutional politics!).
- Short-term renewable contracts. Since firing people is basically impossible within the University, only by offering short-term contracts (two or three years) was it possible to get rid of people who turned out not to be great fit. It was important to be able to take a chance on people who might not work out. Maybe about 30% of people given a job at FHI were offered to have their contracts extended after their initial contract ran out. A side-effect was to filter for individuals who truly loved the intellectual work we were doing, as opposed to careerists.
- Valued: insights, good ideas, intellectual honesty, focusing on what's important, interest in other disciplines, having interesting perspectives and thoughts to contribute on a range of relevant topics.
- Deemphasized: the normal academic game, credentials, mainstream acceptance, staying in one's lane, organizational politics.

- Very few organizational or planning meetings. Most meetings were only to discuss ideas or present research, often informally.

A comment from a member:

> "I think there's no cookie-cutter template for replicating FHI because it depends critically on having the right (rare) people and a particular intellectual culture. The secret sauce was not an organizational structure or some kind of management process. But with the right people and culture, then shielding from other constraints can become enabling.
>
> To the extent that it can be replicated, I think it is because (a) it was an existence proof of organization, template ideas and research results, and legitimization, and (b) the intellectual culture has spread (e.g. in the wider rationalist and EA networks). But it could probably not be replicated by having some random administrator or manager trying to reproduce the same organizational structure - that would be like the cargo cult."

So, the conclusion may be that while the above considerations give a recipe to aim for, the key question for any replication should be: "What are the important topics this organization should aim at?" Pursuing those topics must always be at the center of what is being done (both in research and administration), even when new knowledge and developments change them and their priorities.

FHI AT OXFORD

> the big creaky wheel
>
> a thousand years to turn
>
> thousand meetings, thousand emails, thousand rules
>
> to keep things from changing
>
> and heaven forbid
>
> the setting of a precedent
>
> yet in this magisterial inefficiency
>
> there are spaces and hiding places

for fragile weeds to bloom

and maybe bear some singular fruit

like the FHI, a misfit prodigy

daytime a tweedy don

at dark a superhero

flying off into the night

cape a-fluttering

to intercept villains and stop catastrophes

and why not base it here?

our spandex costumes

blend in with the scholarly gowns

our unusual proclivities

are shielded from ridicule

where mortar boards are still in vogue

- Nick Bostrom, 2018

This was recently set to music by the Fooming Shoggoths, on their AI-generated album *I Have Been a Good Bing*, as a "farewell to one of the greatest intellectual institutions to have ever existed"[147].

[147] https://twitter.com/ohabryka/status/1774982480896065885

Acknowledgments

To everybody who contributed to making FHI possible - funders, researchers, associates, staff, PhD students, visitors, and friends: a huge big thank you!

We hope that the future is as bright, diverse and hopeful as you dreamed.

And if it is not, we will change it into a bright future – together.

∞

Appendix A: Formal goals over time

The first version of the FHI website in October 2005 stated[148]: "The Institute aims to become humanity's best effort at understanding and evaluating its own long-term prospects. FHI will study how anticipated technological developments may change human beings and transform the human condition."

By October 2007[149] it had been updated to: "FHI's mission is to pursue big picture questions for humanity. We study how anticipated technological developments may affect the human condition in fundamental ways, and how we can better understand, evaluate, and respond to radical change."

October 2008[150] had the blander: "FHI's mission is to bring excellent scholarship to bear on big picture questions for humanity."

October 2010[151] went a bit more informal: "The Future of Humanity Institute's mission is to bring careful thinking to bear on big-picture questions about humanity and its prospects."

October 2011[152] went boastful: "The Future of Humanity Institute is the leading research centre looking at big-picture questions for human civilization." However, the blurb also concluded: "Our goal is to clarify the choices that will shape humanity's long-term future."

At a strategy meeting in 2011 we said "FHI is primarily about improving the future of humanity." In 2014 we noted that maybe we should say "FHI is primarily about improving and understanding the future of humanity." Still, the focus was always about doing this through research,

In 2016[153], after a major website redesign it was changed to: "The Future of Humanity Institute's mission is to shed light on crucial considerations for humanity's future. We seek to focus our work where we can make the greatest positive difference." This was further explicated as:

"The Future of Humanity Institute's mission is to bring excellent scholarship to bear on big-picture questions for humanity. We seek to focus our work where we can make the greatest positive difference.

---

[148] https://web.archive.org/web/20051013060521/http://fhi.ox.ac.uk/

[149] https://web.archive.org/web/20071013010744/http://fhi.ox.ac.uk/

[150] https://web.archive.org/web/20081013194244/http://fhi.ox.ac.uk/

[151] https://web.archive.org/web/20101012053829/http://fhi.ox.ac.uk/

[152] https://web.archive.org/web/20111028193829/http://www.fhi.ox.ac.uk/about

[153] https://web.archive.org/web/20161019072801/http://fhi.ox.ac.uk/

This means we pursue questions that are (a) critically important for humanity's future, (b) unduly neglected, and (c) for which we have some idea for how to obtain an answer or at least some useful new insight. Through this work, we foster more reflective and responsible ways of dealing with humanity's biggest challenges."

In 2017[154] this was shortened to: "FHI explores what we can do now to ensure a long flourishing future."

In 2020[155] the boasting was back: "The Future of Humanity Institute is a unique world-leading research center that works on big picture questions for human civilisation and explores what can be done now to ensure a flourishing long-term future."

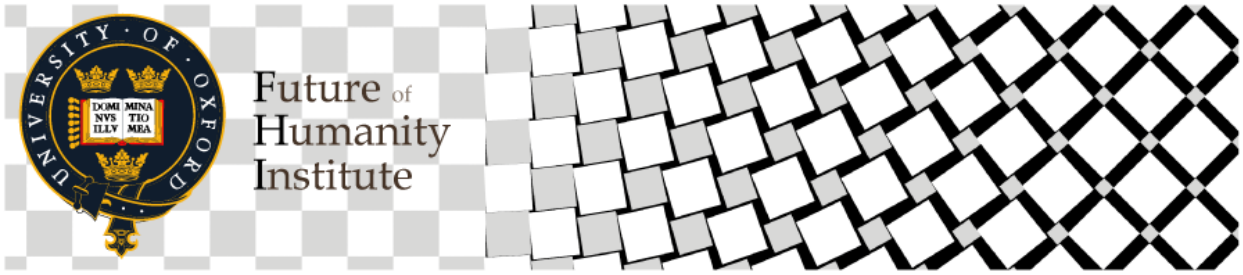
That became the final public version of a goal, since we lost the ability to update the site in 2022.

Overall, the goals remained fairly clear throughout this timespan. There was a focus on research to uncover crucially relevant considerations for the future and elucidating their empirical and philosophical basis. There was also always an interest in making them policy relevant if possible.

[154] https://web.archive.org/web/20171026093136/http://fhi.ox.ac.uk/

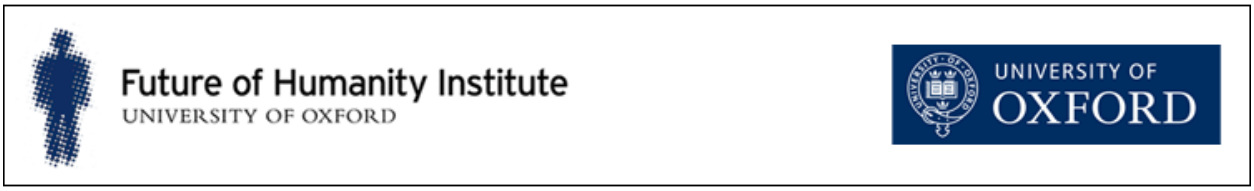
[155] https://web.archive.org/web/20201101102006/https://www.fhi.ox.ac.uk/about-fhi/

Appendix B: The FHI Logo Across Time



First website logo, 2006.

When the original home-made "logo" image on the homepage became too embarrassing, we asked an artist to design a new one. After a long process with many interesting concepts, we ended up with a simple human silhouette seen through a raster.



The very briefly used diffuse person logo, December 2008

We were happy until professor Janet Radcliffe-Richards exclaimed "it looks like the sign on the men's room door!" Once seen, it could not be unseen. We had to find a new one. Fortuitously Nick went on the ski trip and noticed the black diamond symbol for advanced or expert slopes – it was simple, pure, and dynamic. In modal logic the diamond denotes possibility. We had our new logo.

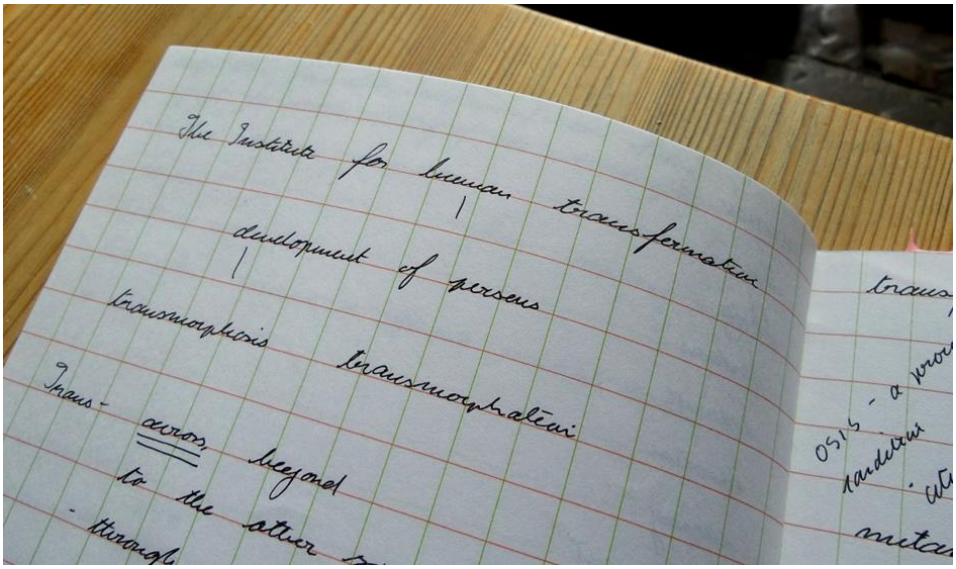
The original black diamond logo. 2009

After a few adventures with framing the diamond with text in different colors, eventually things and typeface settled down.

Over the years the edges have become subtly concave and star-like as the institute and logo matured. Exactly how concave and whether to mathematically describe it exactly was a running trivial bone of contention between Nick and Anders.

2014 Logo

Appendix C: Various Pictures



Heather Bradshaw-Martin (then at the Uehiro Centre) helped set up the institute. Here is a page in her notebook where she brainstormed for names. We were lucky she and Nick chose the right one.

Peeking Through the Veil, 2007. The windows at Littlegate House had plastic insulation that saved energy at the expense of seeing the beauty outside.

Toby Ord, 2007

Nick Bostrom, 2007

WBE workshop, St. Hilda's College May 2007. Peter Passaro, Rebecca Roache, and Bruce McCormik try to get the projector to work. As a general rule, the more high-tech the workshop topic, the more likely annoying practical technical trouble is to happen.

Filling out a survey of risk estimates at the Existential Risk workshop 2008. Doom and gloom.

Existential Risk workshop, 2008. Nick presents some typologies and considerations of the concept.

Existential Risk workshop, 2008.

Dr. James Martin attending the Global Catastrophic Risk conference in 2008.

Global Catastrophic Risk conference, 2008. Milan Ćirković, Nick Bostrom and James Martin.

Punting after the GCR conference 2008. Nick was aiming for reaching the picnic spot faster than anybody else. One person later fell in.

James Hughes at GCR conference 2008.

Discussion meeting, 2008. Anders Sandberg, Toby Ord, Rafaela Hillerbrand, Nick Bostrom.

Winter Intelligence attendees queuing for lunch at Jesus College. Who cares about the length of the queue when there are interesting concepts to discuss with interesting people? January 2011.

The Winter Intelligence Conference 2012 fittingly coincided with unusually cold and snowy weather in Oxford.

James Martin celebrated his graduation anniversary by giving a lecture. November 2012.

FHI and the Machine Intelligence Research Institute think about future plans. And have a post-conference breakfast. December 2012.

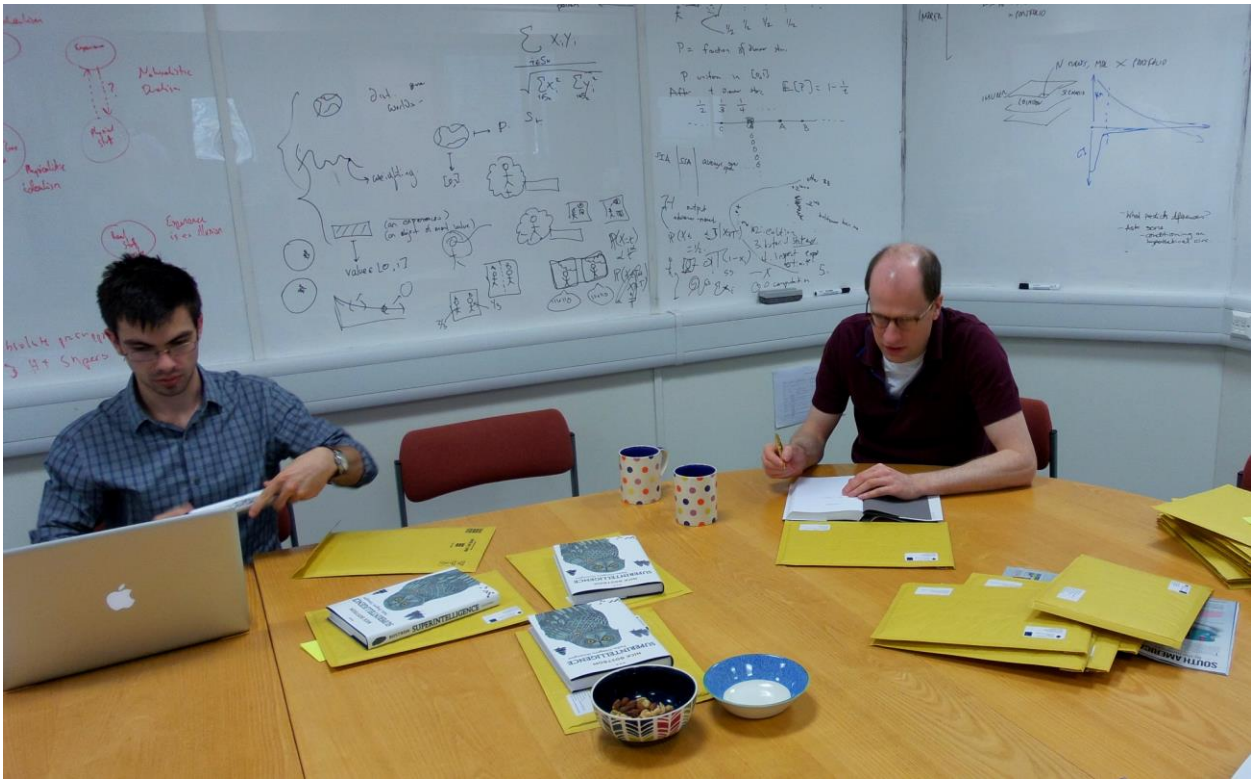
The FHI-Amlin systemic risk of risk modeling team: Nick, Nick, Sean, Anders, Vincent, Andrew and Stuart. Together we are figuring out how to think about risk without increasing risk. January 2014.

Nick delivering the Crucial Considerations talk at Good Done Right (All Souls), 2014

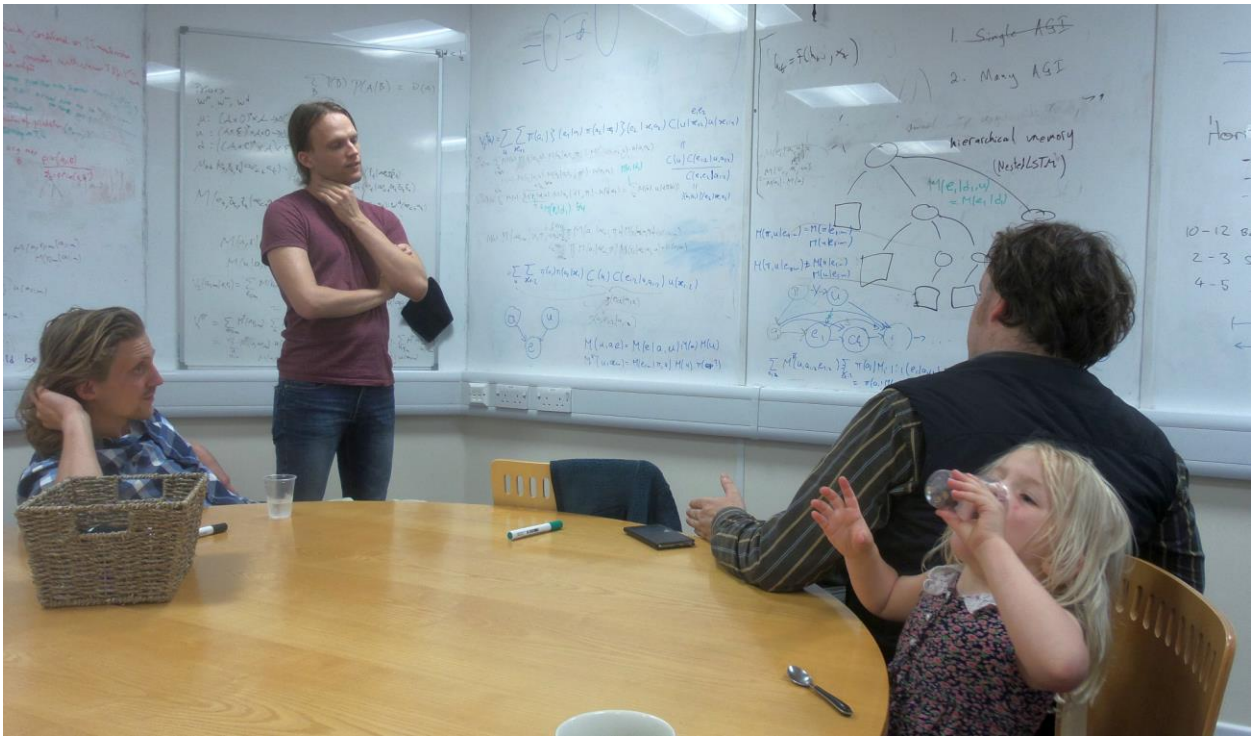
FHI and CSER in Berlin after a biosecurity workshop at the German Foreign Ministry. September 19 2014. Left to right: William Sutherland, Nick Bostrom, Sean Ó hÉigeartaigh, Stuart Russell, Huw Price.

Daniel Dewey made a joke bingo game based on common FHI phrases and activities. State of my board after one week. A problem is that some items are tied to a person and hence cannot be gained by that person. Daniel added the extra rule that we could exchange such boxes for another one (hence my conference box being exchanged for a nootropics box). October 2014.

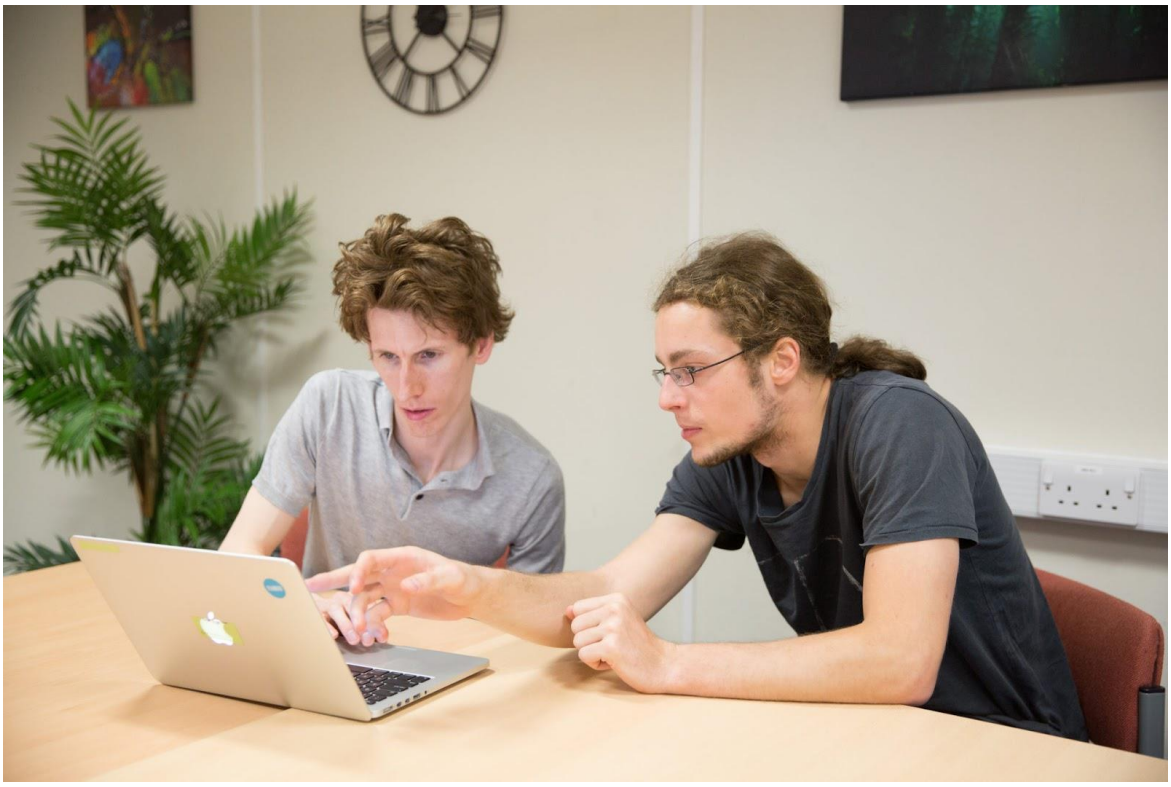
Halloween, 2014



Andrew Snyder-Beattie and Nick Bostrom mailing copies of Superintelligence to various people. In the background anthropic analysis of the Sleeping Beauty Problem and metamodelling of insurance risk. 2014



Tom Everitt, Jan Leike, Stuart Armstrong and Maya Armstrong working on technical AI safety (2016)



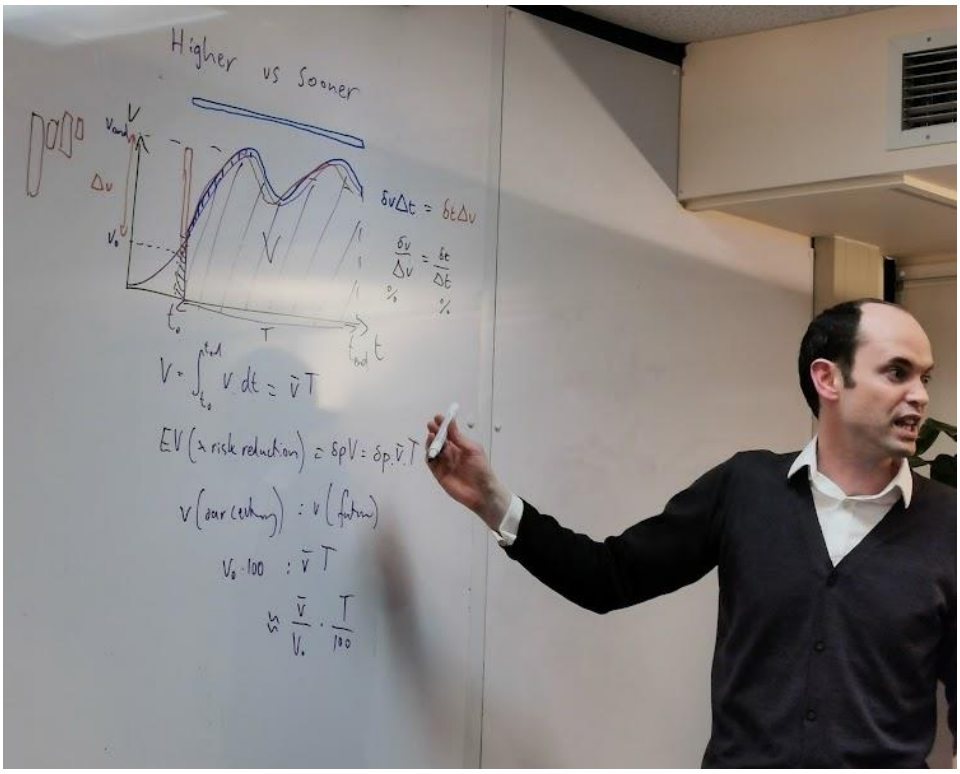
Debugging. Owain Evans and Sebastian Schulze.

View from Littlegate House towards Tom Tower in Christ Church, across the rooftops of Pembroke College.


KEEP
CALM
AND
REDUCE
XRISK

A reading group, 2017

FHI present as NGO at the Convention on Biological Weapons in the United Nations in Geneva, December 2017.

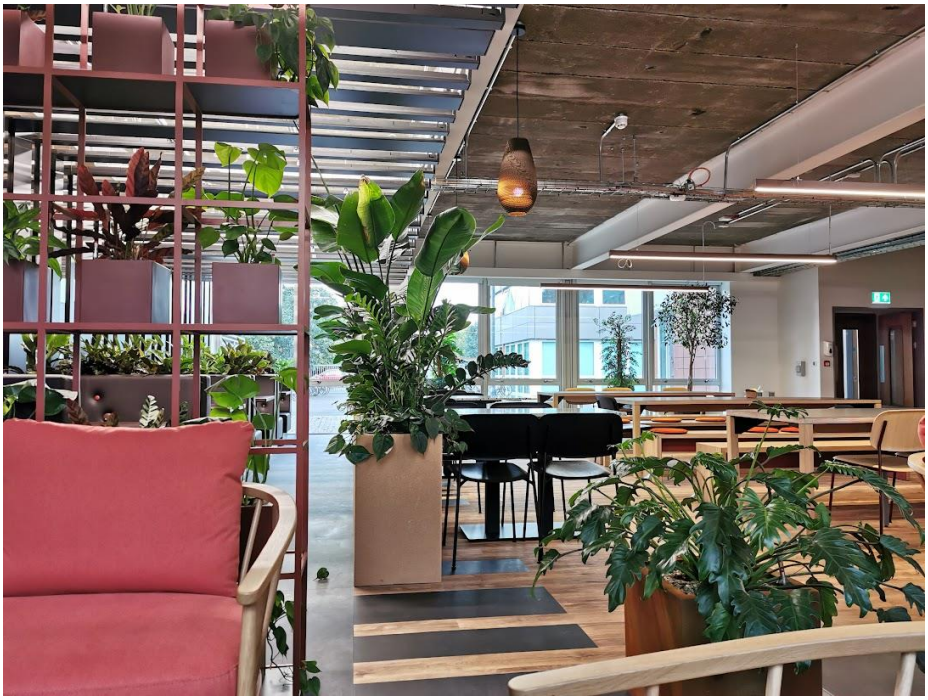


Toby Ord presents different approaches for quantifying "improving the future". Feb 11 2020.

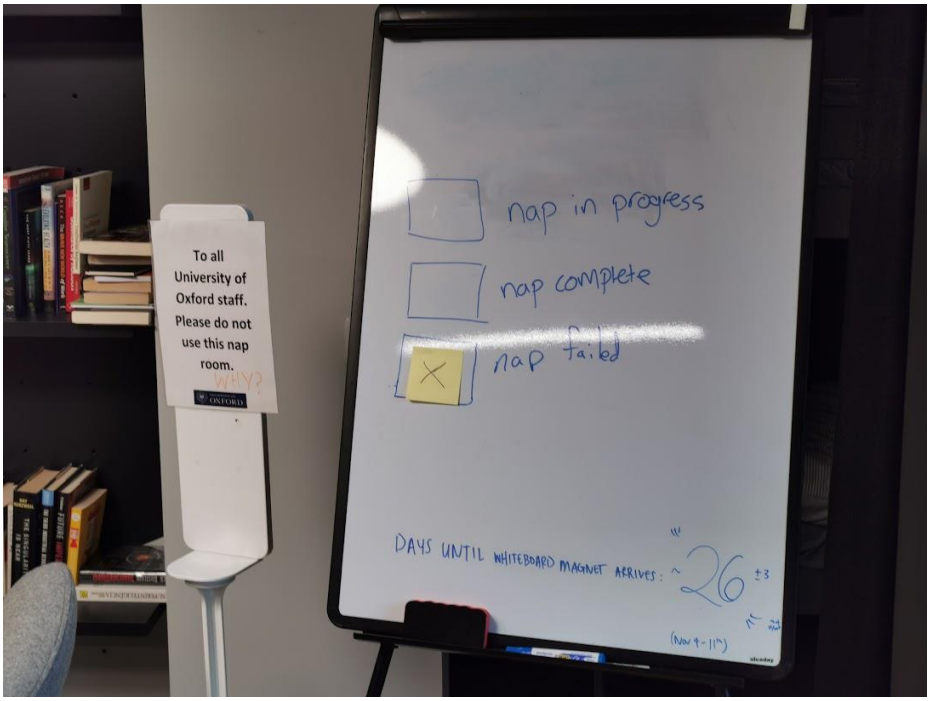
Anders giving a speech to Toby at the book launch for The Precipice, 2020. Corpus Christi College.

FHI's new offices in Trajan House on Mill St. May 2021. It was a former school building converted into offices for

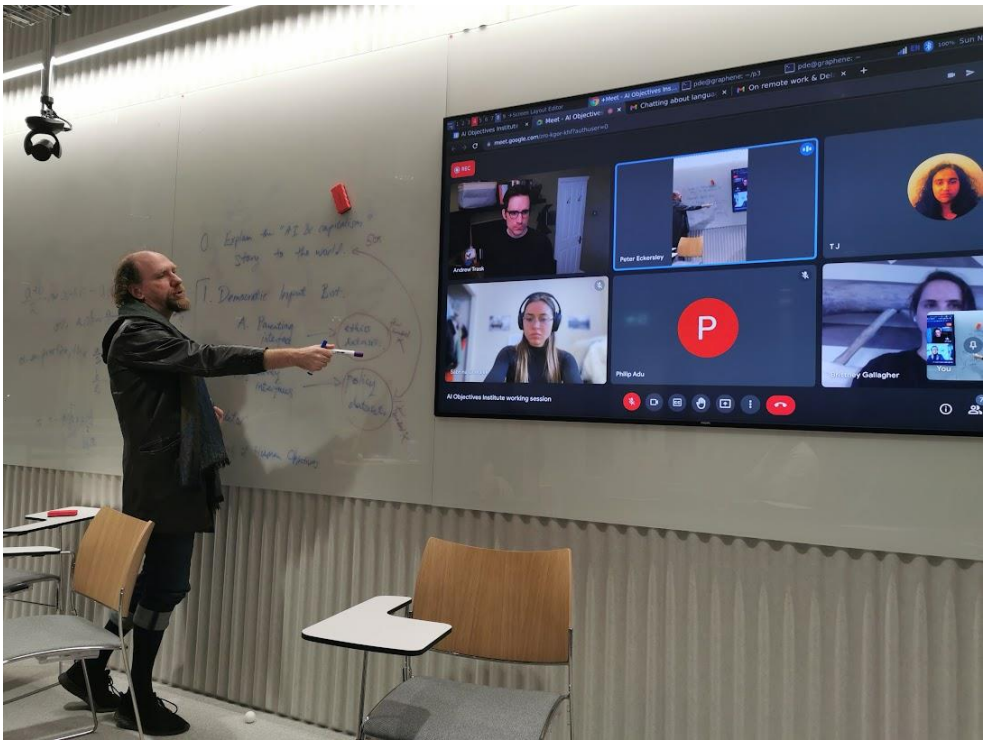
FHI, GPI and various non-profit organizations.



The Trajan House lunch room, September 2021. One of the key ideas was to ensure that people from the different organizations and visitors to the building naturally mingled and had conversations. Still, the vast whiteboards of Littlegate House were missed.



The nap room saga. Due to regulations FHI members could not use the FHI nap room. October 2021.



The late research associate Peter Eckersley led a joint FHI/AI Objectives Institute hybrid meeting. Nov 2021. Peter and Anders Sandberg started the very vital AOI while procrastinating on a joint paper.

Anders Sandberg at the Nobel Week Dialogue 2022 with Helga Nowotny (President of the ERC), Frances Arnold (Chemistry 2018), Steven Chu (Physics 1997).

The Duck of Existential Risk. "Rubber duck debugging" is a practice in programming where one explains one's problem to a literal rubber duck. This is surprisingly often helpful. Anders donated a big black duck to the institute, where it has been moving around (here in the biosecurity room). 2023.

Anders Sandberg discussing astrophysics with Lord Martin Rees. Cambridge, April 2023.

Anders Sandberg pointing dramatically at a crucial idea. FHI/Foresight Whole Brain Emulation Workshop 2023.