Ten years closer to the future

A personal reflection on the ten year anniversary of the Future of Humanity Institute

Anders Sandberg, Oxford Martin Senior Fellow

2015 marks the ten year anniversary of the Future of Humanity Institute at Oxford University. It was founded in late 2005 as part of the original batch of institutes of the Oxford Martin School[1]. The FHI began as little more than a research group within the philosophy department but has blossomed into something unique.

Early years

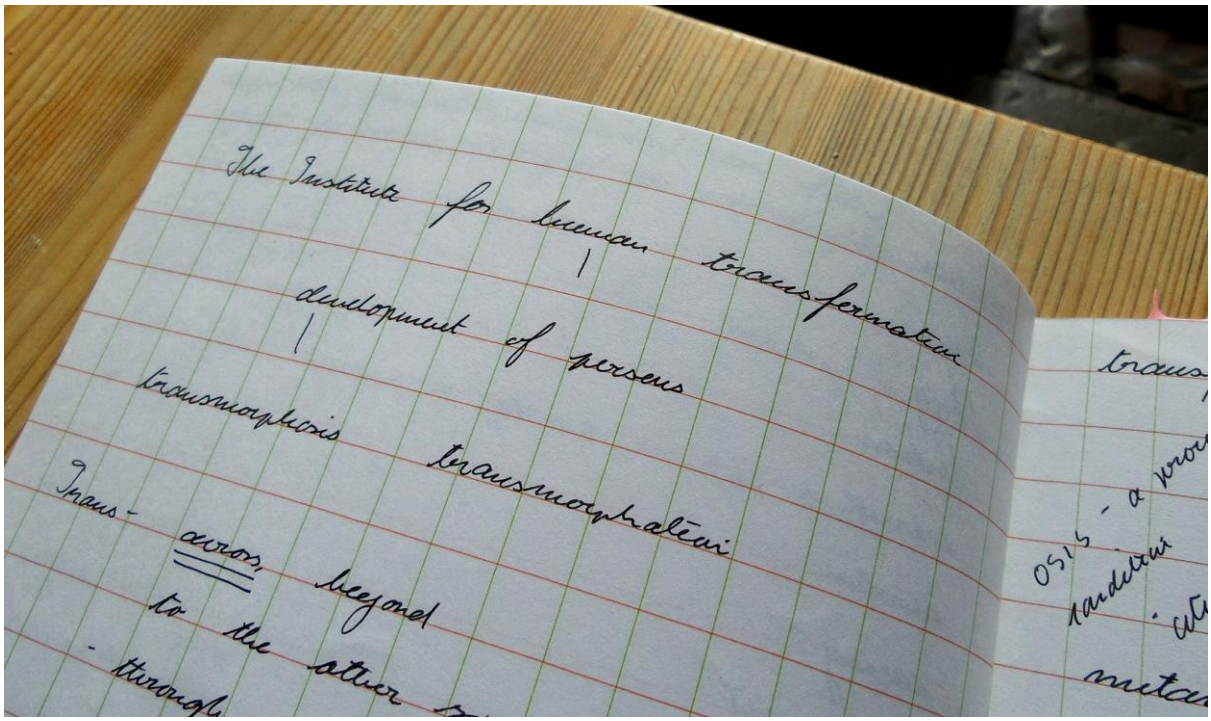

Figure 1: Heather Bradshaw-Martin (then at the Uehiro Centre) helped set up the institute. Here is a page in her notebook where she and Nick brainstorm for names. We were lucky they choose the right one.

Myself, I arrived in Oxford on 1st of January 2006, taking up a position in the EU ENHANCE project. 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, the Maastricht node mood enhancement, Milano life extension, Stockholm bodily enhancement, and Bristol coordinating us.

[1] Back then the James Martin 21st Century School – a name wisely changed, since nothing ages faster than a "futuristic" name referring to a particular time.

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

During this early period FHI produced an book on global catastrophic risks that still remains a core reference in the field, held a workshop on brain emulation that led to the formation of a small but active research field, and worked with researchers in cognitive enhancement to understand what forms of enhancement truly matters and how policies ought to be updated to deal with it.

We also gained the black diamond logo. 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. We were happy until a fellow professor 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. Over the years the edges have become subtly concave and star-like as the institute and logo matured.

Figure 2: The Petrov seminar room is named after Stanislav Petrov, who is credited with preventing an accidental nuclear war between the Soviet Union and the US 26 Steptember 1983. Across the hallway is the Arkhipov room, named after Vasili Arkhipov, who prevented a nuclear launch from submarine B-59 during the Cuban Missile Crisis.

Maturation and networking

Figure 3: Research seminar where Eric Drexler, David Deutsch and Nick Bostrom discuss possible ways of analysing the limits of technological progress. May 2012.

When Nick Bostrom began to work on a book on existential risk in 2009 we soon found one of the chapters 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 on its own, evolving into a long-running research project and eventually a monograph on its own, Bostrom's 2015 *Superintelligence*.

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 and juicy problem to the team: given the use of complex, imperfect mathematical models of risk that are shared across the insurance business, was there a systemic risk like the one that had brought down the financial crisis of 2006? And what could be done about it? 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. This research is now being taken to the next level by a further collaboration with the Institute for New Economic Thinking.

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 analysing the potential automation of different jobs, leading to the widely cited (and misunderstood) claim that 47% of jobs may be at risk from automation. This work is now continued at the Oxford Martin Programme on Technology and Employment.

FHI expanded into several directions. The Global Priorities Project emerged as a collaboration between the Centre for Effective Altruism and FHI, analysing 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. The ERC Advanced Project on Uncertainty and Precaution is currently starting up, attacking the problem of how decision-making can be guided better than the precautionary principle in domains of radical uncertainty.

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 Centre

for Artificial Intelligence Policy and the Leverhulme Centre for the Future of Intelligence, both starting in 2016. All of them connect FHI deeply into the world of AI engineering and policy.

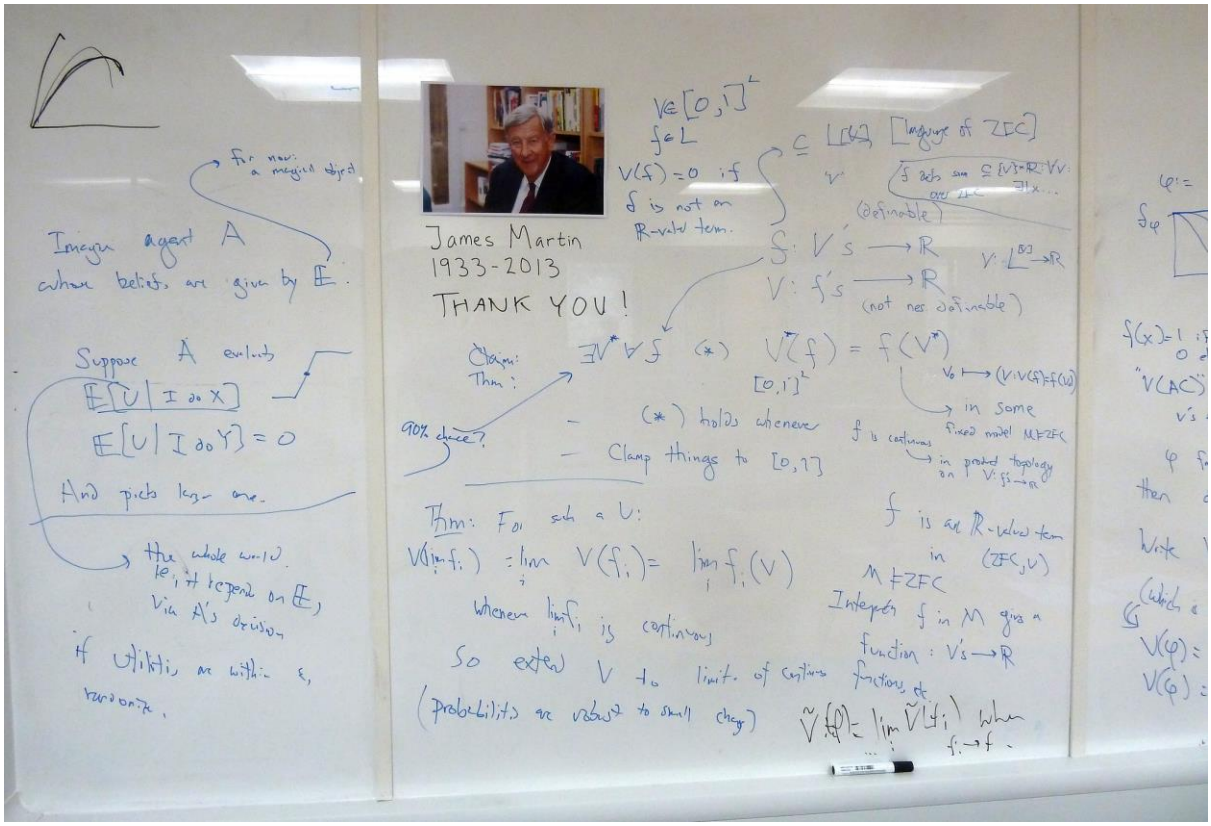

Figure 4: Typical state of the main whiteboard (in this case dealing with self-reflexive AI). We decided that the best way to recognize James Martin was to always have his picture in the middle of our work.

The Oxford Martin School has always been an important matrix for FHI. While FHI was growing up the 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. The truly unifying theme however was ambition: if the research we did was not going to change the world it was likely not worth doing. Being in an environment where you were rewarded for trying to think big and then tell the key people what they needed to know was not only unusual and inspiring, but it sharpened focus and sense of responsibility tremendously.

James Martin was always interested in the work at FHI and tended to visit whenever he was in Oxford; I remember having good-natured quarrels with him about the feasibility of brain emulation and the safety of alien software intelligence, as well as inspiring conversations about where the levers of global change can be found. His death in 2013 was a great loss, and I am happy his widow Lillian still visits and grills us.

FHI today: tackling the big picture

Figure 5: A few of FHIs deliverables in book form.

The FHI research focus today still encompasses the old interests, but now in a framework aiming to answer important questions about humanity's future that are unduly neglected, yet might be tackled using available tools. The core interest is macrostrategy: where should we be going? What actions today truly matters? This is complemented by the AI safety research, dealing with a particular subset where we identified good leverage points. There is also much work on how to do technology forecasting and risk assessment better. Finally, the aim is very much practical results that can be used, whether for policymakers or as mental tools for better decisions.

What makes FHI unique is not so much the interdisciplinary work but the focus on looking for the most enormous, important problems and then attacking them, with no regard for traditional tractability.

This might appear Quixotic[2], but many of the truly important problems appear under-researched and there are hence low-hanging fruits to be picked; by identifying the problems and doing a preliminary survey FHI can increase the amount of knowledge about them significantly despite a small staff and time-budget. If a problem is truly important, even making a microscopic dent in it means a large utility gain. Often our role is more of explorers than colonizers: we are happy to leave many domains to better equipped teams once they are discovered and mapped. Like explorers we also cannot rely on a fixed preconception of what tools will be required: FHI is methodologically flexible and agnostic, using anything from analytic philosophy to large-scale agent based models to examine the problems. Having a large interdisciplinary toolkit is essential for exploring the outer reaches of the future.

What have we achieved? We certainly have produced some useful ideas: tools such as the reversal test and evolutionary heuristic for judging proposed enhancements, noticed and formalized problems that thinking in these fields must get around (such as the "probing the improbable" issue, the

[2] Piet Hein was right in that "A problem worthy / of attack / proves its worth / by hitting back."

unilateralist curse), and turned loose speculations on futuristic topics into somewhat more cohesive research topics (brain emulations, technological singularity or large-scale space colonization). More importantly – but harder to measure as deliverables – we have catalysed or participated in the formation of communities or research directions that approach important problems: effective altruism, AI safety, brain emulation, the applied rationalism movement, and so on. We have helped bring about at least one new research institute, the Centre for the Study of Existential Risk in Cambridge[3], and now exist within a larger community of other institutes worldwide devoted to big picture problem solving and reducing global risks. None of these things did exist ten years ago.

FHI has existed ten years. A simple Copernican argument would suggest it hence has a good chance of lasting ten more years. But the aim has never been to *persist* as an institution but to *improve* the future of humanity. Radically.

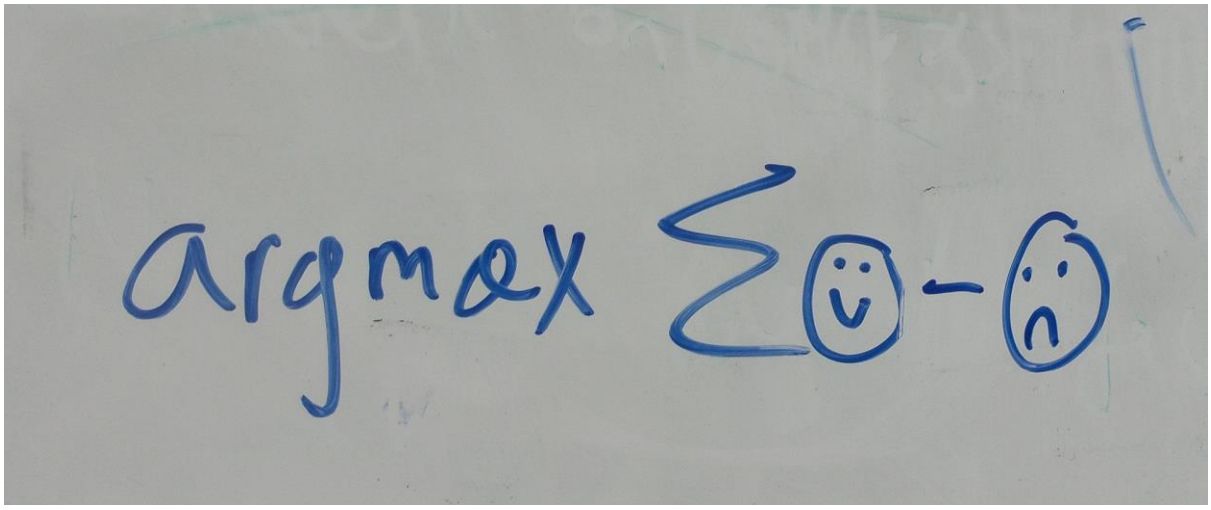

Figure 6: Equation found on the whiteboard. Let us maximize the amount of good (integrated across the future light-cone of humanity) minus the amount of bad. Of course, some further filling in of detail is needed...

[3] Having a sibling institute in "the other place" is an excellent way of avoiding groupthink and get the benefits of intellectual competition.