[Excerpt from *The Anti-Catastrophe League*, by Tom Ough. Available at Waterstones (UK) and Amazon (US), and reproduced here with the author's permission.]

CHAPTER 8

THE MISFIT PRODIGY

Thought bubbles
Inflating multiplying drifting
Rearranging and annexing
Little rainbows of wishful thinking
Enclosing gases of concern

Nick Bostrom, 'Bubble Ontology'

On 16 April 2024, the website of the Future of Humanity Institute was replaced by a simple landing page and a four-paragraph statement.[1] The institute had closed down, the statement explained, after 19 years of existence. The statement briefly sketched the institute's history, appraised its record, and referred to 'increasing administrative headwinds' blowing from the University of Oxford's Faculty of Philosophy, in which the institute, generally referred to as FHI, had been housed.

Thus died one of the quirkiest and most ambitious academic institutes in the world. FHI's mission had been to study big-picture questions for humanity: our direst perils, our range of potential destinies, our unknown unknowns. By the time it was founded, we knew about the natural risks, climate change and nuclear war. But what else did we need to know? FHI's researchers pondered this matter deeply. The concept of super-intelligent AI is one of several that they ushered into best-

seller lists, serious academic study and the headquarters of the United Nations.

To its many fans, the closure of FHI was startling. This group of polymaths and eccentrics, led by the visionary philosopher Nick Bostrom, had seeded entire fields of study, alerted the world to grave dangers, and made academia's boldest attempts to see into the far, far future. As well as fans, though, the institute had detractors. The one thing FHI had not foreseen, those detractors commented archly, was its own demise.

Why would the university shutter such an influential institute? Was this a whodunnit? What was FHI like on the inside? To what extent did its office harbour medieval weaponry? And, to invert the organisation's mission somewhat, how, if at all, will our descendants look back on FHI? For an institute that set out to find answers, FHI left observers with a lot of questions.

In 1989, a teenage Swede borrowed a library book of 19th-century German philosophy, took it to a favourite forest clearing and experienced what *The New Yorker* would later describe as 'a euphoric insight into the possibilities of learning and achievement'.[2] Damascene moments are rare in real life, but this seems to have been one of them. Niklas Boström dedicated himself to a lifetime of intensive study. He withdrew himself from school, rattling through syllabuses at home, and he read widely and manically. At the University of Gothenburg, he took a BA in philosophy, mathematics, mathematical logic and artificial intelligence.[3] After that, he took postgraduate degrees in philosophy and in physics, in astrophysics and general relativity, and in computational neuroscience. In what little spare time he had, Boström emailed, and met up with, fellow transhumanists: an assembly of vitamin-popping free spirits, united by their enthusiasm for radically improving human biology and lifespans via methods such as cryopreservation.

In 1998, Boström co-founded the World Transhumanist

Association.[4]* As early as 2001, he was studying little-known phenomena called 'existential risks', writing that nanotechnology and machine intelligence could one day interfere with our species' ascent to a transhuman future.[5] He also, in 2003, formulated the 'simulation hypothesis', advancing in *Philosophical Quarterly* the theory that we might be living in a computer simulation run by humanity's hyper-intelligent descendants.[6]† Most famously, he posited the thought experiment of the 'paperclip maximiser', illustrating the risk posed by powerful AI by imagining such a system being tasked with improving the output of a paperclip factory. If the system is able to improve itself in service of its goal, wrote Bostrom, what follows is 'the consequence that it starts transforming first all of earth and then increasing portions of space into paperclip manufacturing facilities'.[8]

By 2003, Bostrom‡ had arrived at Oxford as a postdoctoral fellow at the Faculty of Philosophy. Many years later, the faculty would be his *bête noire*. But it was Bostrom's membership of it that enabled a stroke of luck that would dramatically change his life. He had already come across James Martin, an IT entrepreneur. Following his success in the business world, Martin had become a prescient futurist. He wrote more than 100 books, including a Pulitzer-nominated exploration of how computerisation would change the world.[9] Martin also produced a documentary featuring Bostrom.

But Martin, who lived on his own private Bermudan island,

* The transhumanists deserve a much fuller treatment than I am able to give them. I recommend that those interested read Elise Bohan's PhD thesis, which is a rollicking history of this group.[7]

† Armchair philosophers will identify that this theory is itself a descendant of similar thought experiments. Plato proposed that our relation to reality might be that of cave-dwellers to the outside world. Descartes posited, in place of hyper-intelligent descendants, an evil demon. And we must not ignore the contribution made to this field by *The Matrix* (1999).

‡ Time and Anglicisation have sanded off his umlaut.

was not only a producer; he was also a deep-pocketed philanthropist. Thanks to Julian Savulescu, who was another young philosopher interested in human enhancement, Bostrom learned that Martin was planning to fund future-minded research at Oxford. Hoping that this could encompass work on his specialisms, Bostrom made his case to the university's development office.

After two decades, the details are hazy. Bostrom emphasises that there was no single turning point. But FHI lore has it that the university liked to host dinners where such donors sat next to exciting academics, creating the perfect conditions for what we now call a nerdsnipe. Bostrom spoke earnestly and intelligently about existential risk and transhumanism, goes the story;* Martin was deeply engaged. 'That's Nick's greatest charm, actually,' Bostrom's FHI colleague Anders Sandberg told me. 'He can be nice, but he's even better when he's just going at an important question, explaining why it is the most important thing in the universe.' Martin was, according to an FHI old-stager, 'blown away'.

In 2005, Martin made what was then the biggest benefaction to the University of Oxford in its nine-century history. It came to more than £70 million,[10] of which a small portion funded what Bostrom termed the Future of Humanity Institute. 'It was a little space,' Bostrom told me of this nascent FHI, 'where one could focus full-time on these big-picture questions.'

That seed grant was small: enough for a few people for three years. Narrow thinkers were not welcome. With such a small team, and such an unconventional brief, Bostrom wanted multidisciplinarians. He was looking, he told me, for 'brainpower especially, and then also a willingness and ability to work in areas where there is not yet a very clear methodology or a clear

* Bostrom can no longer recall which dinner happened when, and Martin is no longer with us. This leaves us with only the second-hand anecdote to rely on.

paradigm. So, it's a generalised thinking skill, often combined with some kind of polymathy.'

One of his earliest hires was Anders Sandberg, whom you might recall from my chapter on super-eruptions. Sandberg, also a Swede, was a fellow member of the Extropians, an online transhumanist community that Bostrom had joined in the 1990s.* This community also included Eliezer Yudkowsky, a brilliant, disagreeable young autodidact† who often corresponded with Bostrom, as well as individuals who helped initiate the cryptocurrency revolution.

Where Bostrom is generally ultra-serious, Sandberg is ebullient and whimsical. These differences in personality belied the two men's similarity in outlook. Sandberg, too, was an unorthodox thinker who was interested in transhumanism and artificial intelligence. (Within transhumanism, Sandberg was particularly interested in the theoretical practice of whole-brain emulation; that is, the uploading of a human mind to a digital substrate. This might, one day, be humanity's best long-term solution to biological death.)

After finishing a PhD in computational neuroscience at the University of Stockholm, Sandberg had embarked on an exhibition project in which he tried to explain the human brain 'to everyone,' he told me, 'from kindergarten kids to professors'.

* The name 'Extropian' is a play on the word 'entropy'. Entropy, put simply, is the natural tendency toward decay. An Extropian believes that living beings ought to be able to defy entropy and remain indefinitely vital.

† It would be impossible to do justice in this footnote to the singular mind that is Yudkowsky's. I will return to him later, but for now, know this. As a teenager, he quit high school in order to teach himself, scoring preposterously highly on his exams. As an adult, his works on rationality and AI have made him one of the most influential thinkers of our time. He is a lordly essayist, sometimes given to administering brutal put-downs to those in his intellectual warpath. As he told one such victim in 2023: 'You are trying to solve the wrong problem using the wrong methods based on a wrong model of the world derived from poor thinking and unfortunately all of your mistakes have failed to cancel out.'

With that project finished too, Sandberg was free to be interviewed at Oxford, arriving on a wet day in July 2005. Even a futurist, it turns out, can appreciate the timelessness of the dreaming spires. 'I was walking along Queen's Lane in the rain, and I realised that I couldn't tell, from this vantage point, which century I was in.'

Sandberg was interviewed in the Faculty of Philosophy's Ryle room, named for the philosopher Gilbert Ryle.* It is a room, Sandberg explained, where philosophers would have smoked pipes for decades. 'Smoking had probably been banned there for decades too, but the philosophical smoke was still in the walls. And it seemed like it worked.' Sandberg explained some neuroscience to the faculty staff who were assessing him and communicated his aptitude in another little-known area of human endeavour: web design. He was hired, and he returned in January 2006 to take up a desk at FHI and a 'silly little room in Derek's house'.

This was no ordinary Derek. Sandberg was lodging, with Bostrom, in the home of Derek Parfit, a wild-haired recluse who was also one of the most influential moral philosophers of the modern era. Bostrom had the master bedroom and collected rent from the rotating cast of lodgers.† Parfit, Sandberg recalled,‡ slept in 'a little cubby hole' of a bedroom and would scuttle at odd hours between the cubby hole and his office at All Souls, the highly selective graduate college seen as an Oxford within Oxford. For a time, the house's proverbial trousers were worn by a mouse that nibbled the philosophers' food and

* Ryle, famously, was critical of the view that the mind is a mysterious and non-physical entity that nevertheless controls the body. Perhaps he might have been sympathetic to the idea of whole-brain emulation.

† Sandberg and Bostrom were part of a succession of young intellectuals who lodged with Parfit. David Edmonds, in his biography of Parfit, describes 'a variant of the hit TV comedy *The Big Bang Theory*, except that, instead of being a household of physicists, it was chock-full of philosophers'.[11]

‡ Parfit died in 2017.

consistently evaded their ethical, but ineffective, attempts to trap it. 'The mouse was having the time of its life,' said Sandberg, 'up until the point where I accidentally killed it when throwing away some garbage. I had to bury it behind the Ashmolean Museum.'

Though the mouse did not live to see FHI take shape, the institute was developing nevertheless. Including Sandberg, Bostrom hired three researchers and began to sculpt a research agenda that, in these early years, was primarily concerned with the ethics of human enhancement. An EU-funded project on cognitive enhancement was one of FHI's main focuses in this period. The institute also organised a workshop that resulted in Sandberg and Bostrom authoring an influential paper that laid out the technological advances required to make whole-brain emulation feasible.[12] At the same time, FHI staff were producing work on the gravest perils facing humanity, a topic that was not yet an established academic discipline. In 2008, Bostrom co-edited the book *Global Catastrophic Risks*, which brought together analyses of such threats as asteroid impacts, nuclear war, totalitarianism and advanced nanotechnology. Bostrom would soon co-author research with Eliezer Yudkowsky, the brilliant Extropian whose excitement at the possibilities of AI had turned to terror.[13] Yudkowsky had hitherto existed outside the academic firmament; his co-authorship with Bostrom symbolised the entrance of his unorthodox ideas into the august halls of academia.

Those who worked for FHI had to accept temporary contracts and an ugly office building, Littlegate House. The building bore scant comparison with the beautiful quadrangles that were its near-neighbours. (Jaan Tallinn, an occasional visitor, is said to have joked that the windowlessness of Littlegate House had a hidden benefit – it would make FHI easier to simulate on the graphics cards of humanity's descendants.) In return for his staff's forbearance, Bostrom tried to lay over them a carapace that would shield them from many of the more tiresome

demands of academic life. There were no requirements to teach. As time went by, there was decreasing pressure to publish via traditional academic modes. Explaining this philosophy to me, Bostrom compared his staff to gems so brilliant that a jeweller would want to create bespoke casting for them. He wanted 'to pick these jewels and then create a kind of organisational fixture around them that would let them scintillate and do their thing with the smallest possible number of distractions'.

One of Bostrom's first selections was Toby Ord, a computer scientist turned philosopher. Ord was one of the key figures in the founding of Effective Altruism. His life had changed when he read the 'drowning child' argument advanced by the Australian philosopher Peter Singer. If you'd be happy to ruin a coat by leaping into a pond to save a drowning child, goes the argument – and many readers might nod their heads at this point – you should also be happy to spend the coat's worth in order to save the life of a child far away from you. At this point, readers often stop nodding; Ord did not. He researched the most cost-effective charities in the world, finding them to be a selection of those that addressed neglected diseases in the developing world. And he pledged to give away everything he earned over a modest inflation-adjusted allowance.

In 2009, Ord was put in touch with a similarly-minded graduate student called Will Crouch.* A coffee meeting in a graveyard, of all places, became a five-hour discussion of the necessity of giving, and the best ways in which to do it.[14] Crouch helped Ord found a charity, Giving What We Can, whose members give 10 per cent of their earnings to charities deemed to be effective. This meeting was a crucial moment in the development of effective altruism – a movement that would later provide considerable financial muscle to the thinking that emerged from FHI.

* Having changed his name on getting married in 2013, Will Crouch is now known as Will MacAskill.

Another early hire was Eric Drexler, an engineer, Extropian and futurist who has been called 'the undisputed godfather of nanotechnology'.[15] Like Ord and Sandberg, Drexler would stay with FHI until the very end. Each of the senior researchers had their own signature style of thinking. Bostrom would turn roiling, nebulous concepts into categories, like a cartographer mapping untouched continents. Ord would relentlessly quantify, often grappling with different forms of infinity. Drexler would approach things as an engineer, starting with a blueprint before moving on to practicalities. Together with Sandberg, they helped set the institute's tone: eclectic, curious and unabashedly vigorous in their pursuit of questions other academics would not touch.

Outside the carapace, the world was changing. It was the early 2010s and 'AI winter', as the field's period of moribundity is known, was easing. FHI staff, Bostrom in particular, began to spend more time considering the risks and opportunities that might result from the era's advances in machine learning. Bostrom began writing a book of his own on catastrophic risks; the work on human enhancement was largely wound down. 'This was a fairly typical approach for FHI,' Sandberg wrote in his retrospective of FHI.[16] The *modus operandi* was to find a neglected topic deserving of research, before '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'.

Two particular seedlings – AI risk and AI governance – were to become almighty forests. FHI itself was now a quickly growing sapling. In 2011 or so, it had been moved downstairs to a bigger set of rooms within Littlegate House. The original premises had whiteboards, but Sandberg had insisted on more. The new office had a central room that was encircled by whiteboards – a 'breathtaking 200-degree panorama!', as Sandberg put it.[17] It was nicknamed the 'whiteboard panopticon'. This

room, the James Martin Room, became the epicentre of the institute's collective thinking. Here, FHI staff would sketch out anything from potential solutions to AI safety (one researcher was briefly overjoyed when he mistakenly believed he had solved the problem), to a prediction of what interstellar war would involve, to a consideration of what music would be like if we had more than one time dimension. (The latter exploration was, of course, Sandberg's. Operations staff would sometimes groan at researchers' more flagrant indulgences of whimsy.) Particularly interesting whiteboard notes, especially those which involved collaboration, would be left up for days or weeks. On another whiteboard, Bostrom maintained a scorecard for FHI. He worked at night and was believed to update the number when nobody else was around. Over time, the number crept upward, though it would fall when there were setbacks. Bostrom, who can call upon a dry wit when he chooses to,* told me that the precise workings of the metric were 'shrouded in mystery'.

Martin died in 2013. FHI commemorated its most important early supporter by hanging a picture of him in the room that bore his name. Martin's death, to the sorrow of FHI staff, meant that he missed what was to be FHI's heyday. The institute was becoming rapidly more influential, and its staff's mistakes – historic and ongoing – were yet to catch up with them.

A crucial factor in FHI's ascent to prominence was its earlier decision to focus more intellectual energy on the risk posed by AI. Bostrom, you will recall, had been writing a book on catastrophic risks. A chapter of that book had taken on a mind of its own, so to speak, becoming the sole focus of the project. The chapter, naturally, was on artificial intelligence, and the book, the bestselling *Superintelligence*, warned humanity that the

* Internal FHI legend had it that Bostrom, as a younger man, did stand-up comedy. It was even claimed that there was extant footage, though few staff, if any, ever saw it.

creation of super-intelligent AI is likely to be our final act – for good or ill.

Bostrom likes to adorn his work with fables of his own devising and the book's cover, a beady-eyed owl, refers to a story he included in the book. A group of sparrows, wishing to build their nests with the help of a bigger bird, set out to find an owl egg. In doing so, they defy the warnings of Scronkfinkle, 'a one-eyed sparrow with a fretful temperament'. Scronkfinkle tells his fellow sparrows that, without an attempt to master owl-taming, the plan would be their ruin. The sparrows do not listen and the fable is marked unfinished.[18]

Thanks in part to Bostrom's knack for storytelling, the book was a wild success. Elon Musk, who would soon become a donor to FHI, publicly praised it, as did Bill Gates. (Musk's donation was the first major funding for technical work on AI safety, though his enthusiasm would later be used against FHI.[19]) Derek Parfit was said to have received the book as a 'work of importance', while Sam Altman, then a 29-year-old who had just been put in charge of the Silicon Valley start-up accelerator Y-Combinator, wrote that *Superintelligence* was 'the best thing I've seen' on the topic of AI risk.[20] In October 2015, Bostrom briefed a United Nations committee on the dangers posed by future technologies.[21]

When academics become famous, they can sometimes be regarded with suspicion by colleagues. But this was not true of Bostrom at FHI, where he was deeply respected. 'From the outside,' said a former colleague, 'I wouldn't have been able to see the difference between Nick and the other researchers. It's only when you watch them in discussion that you see it. Oh my God, the long tail of intelligence really is long.'

Littlegate House, that cramped, ugly building with a dearth of windows, now felt like one of the brightest intellectual scenes in the world. The work on AI was more than talk: FHI researchers were among the first people in the world to do empirical, rather than just theoretical, work on the problem of AI alignment. After

a period at FHI, Jan Leike and Paul Christiano helped create the method we now know as reinforcement learning from human feedback, a method which today undergirds every major large language model. With two research scholars, the alignment specialist Owain Evans provided an important benchmark of AI truthfulness; this benchmark is still used by major developers.[22] And Katja Grace, with Evans and others, began the project that became AI Impacts.[23] (AI Impacts gathers and synthesises experts' views on what we can expect from AI development, providing useful data for decision-makers.)

By the mid-2010s, FHI was attracting visits not only from technology heavyweights such as Demis Hassibis (co-founder of DeepMind) and Vitalik Buterin (inventor of Ether), but also curious hacks from the mainstream media. The writer of Bostrom's *New Yorker* profile found the FHI office to be 'part physics lab, part college dorm room', noting the posters of the film *Brave New World* and of HAL 9000, the computer that goes rogue in *2001: A Space Odyssey*.

Visiting in the same year, a breathless young newspaper reporter* described FHI as being akin to 'the School of Athens taking place in an IT support office'. There were split keyboards, homemade keyboards, Dvorak keyboards.† Embedded in furniture were loose Nerf gun pellets, the remnant of a day in which the young daughter of Stuart Armstrong, an AI safety specialist, had gone hunting for her father's fellow researchers. ('The way of getting a little girl out of her shell is to give her a gun,' Sandberg observed to me, 'or at least a Nerf gun.') Two of the rooms bore names that might be familiar. One was named after Stanislav Petrov, the Soviet lieutenant colonel who averted a nuclear war by correctly guessing that what appeared to be a hail of incoming American missiles was in fact a computer

* Me.

† Dvorak keyboards, unlike Qwerty keyboards, have their keys arranged for optimal typing speed.

glitch. Another was named after Vasili Arkhipov, the submariner who vetoed the use of a nuclear torpedo in the Cuban Missile Crisis.*

Ord's office was beside a thoroughfare. When I visited, I knocked on his door and he gamely chatted to me about the universe's deep future, in which our descendants might harness entire galaxies to power our computation. Sandberg's was a grotto of curios. He had a collection of chemical elements, a *Terminator* skull and many other trinkets. (Bostrom joked that he had saved Sandberg from becoming a museum curator.) Having heard a rumour that there had once been a sword in Littlegate House, I asked Sandberg if it was his. It wasn't, he said, but the rumour might have had something to it. Stuart Armstrong had a collection of katana swords at home and Bostrom, in the days of the Parfit house share, dabbled in swordsmanship himself. On hearing a ruckus from the street one night, Sandberg looked out of his window to see Bostrom and his future wife, Susan, duelling with foam swords they had won at St Giles Fair.

At FHI, Bostrom's ludic side was less visible. In the tiny kitchen – which wasn't really big enough for more than six – more than a dozen researchers, typically, would convivially cram in for vegan lunches brought in from a nearby supermarket. Bostrom would come in later, having designed his nocturnal working pattern so that he could better stay in touch with Susan, who lived and worked in Montreal, and their young son. (Bostrom spends six months per year with his wife and child. For the other six months, the nocturnal working pattern has the additional advantage of quietness and solitude.) Arriving at FHI in the afternoon, he would work into the small hours. But first, in the afternoon, he could be spotted in the kitchen. Here he would concoct the vegetable-based smoothie that he wryly

* I told these stories in more detail in Chapter 4.

called his 'elixir'.* This was often the only time that staff would bother him before he disappeared into his office. Knowing that Bostrom liked to descend deep into the halls of concentration, staff would seldom disturb him. Tanya Singh, whose five-year stint at FHI encompassed several senior operations roles, as well as periods of being Bostrom's executive assistant, said she knocked on his door only seven or eight times in all that time.

On the rare occasions that she did enter Bostrom's brightly lit room,† Singh would find him sitting and thinking in near-perfect stillness. 'There was a palpable intensity in that stillness,' she said. 'I have never seen anything like it. You could drop something next to him – a bomb could go off – and he wouldn't move, he wouldn't register it at all.' (Another member of FHI staff recalled entering the whiteboard room with a group to find Bostrom sitting deep in thought. The group quietly left the room.) Interruptions, Singh believed, would ruin Bostrom's day 'because he would have to yank himself out of that mind state'. Bostrom was as protective of his own freedom to sit and think undisturbed as he was of his staff's. It was a carapace within a carapace, and Bostrom, by all accounts, spent little time maintaining relations with the philosophy faculty.

For a while, this didn't seem to matter. FHI's work was becoming even more relevant to the outside world, even if it wasn't much appreciated within Oxford. In March 2020, Ord had published *The Precipice*, a book that examined the perils facing humanity. 'As the gap between our power and our wisdom grows,' he wrote, 'our future is subject to an ever-increasing level of risk.' The book examined perils such as man-made pandemics, to which FHI had begun to devote more time and attention. FHI brought together policy leaders and experts to discuss such threats, and was beginning to map this new landscape of biological risk. Soon after FHI's biological

* Known ingredients: cabbage, cauliflower, carrot, lime and olive oil.

† A *New Yorker*-certified 14 lamps.

threats team had begun drawing attention to an epidemic in Wuhan, much of the world was plunged into the first COVID lockdowns. It was a vindication of a sort, if a grim one. As with artificial intelligence, FHI had been early to sound the alarm.

Philanthropic funders admired this and other lines of work on catastrophic risks. Thanks to these funders' munificence, the institute was expanding. Its administrative duties were taken on by bright young minds who could otherwise have been earning vast corporate salaries or taking on high-status research work at other institutes. As Bostrom put it: 'We had some guy who came with a jurisprudence doctorate from Yale Law School, which is roughly the best law school in the world, initially to work as an unpaid intern to do menial office chores – replace the paper in their coffee machine and stuff like that. He eventually became like a low-level administrator, a job that would pay next to nothing, because of university salary rules. He could have just walked into the most prestigious law firms in Manhattan.'

And recruits kept coming. FHI's Research Scholars Programme, led by the mathematician Owen Cotton-Barratt, was a revolving door of young talent; Cotton-Barratt was renowned at FHI for his ability to spot people with potential and to encourage them into productive lines of research. Other entry points were the DPhil Scholarship and the summer programme for undergraduates. 'Desks were crammed into every conceivable space with increasing ingenuity,' wrote Sandberg. Sometimes Sandberg struggled to make it to his office without being attracted to an intellectual discussion. If novel ideas and conversations are pollen, then Sandberg was a bumble bee, dragged this way and that by colourful stimuli. 'Quite often,' he told me, 'I stood around with my raincoat still dripping as someone was enthusiastically explaining something.' Multiple staff were invited to present their work to the British parliament; Toby Ord would soon to be quoted in the address that Boris Johnson, then the British prime minister, made in 2021 to the UN General Assembly.

But the writing, appropriately for FHI, was already on the wall. In late 2020, and to the shock of FHI, the university imposed on it a hiring freeze. No new staff, no new research scholars. This was one of several measures, including a fundraising freeze, that FHI staff believed were designed to throttle their work. If this was bureaucratic animus, as FHI believed it to be, where did it come from? FHI staff, for their part, often expressed frustration at the delays and paperwork that university membership entailed. As Sandberg wrote: '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.'

Few people relish paperwork, but to those convinced of the merits of Bostrom's 'astronomical waste' argument, inefficiency is a profound wrong. (In the paper of the same name, Bostrom argues that the future could contain such profound amounts of happiness that any delay constitutes a loss of value – wastage on an astronomical scale, in other words – that 'boggles the mind'.*) Bostrom wanted to hire people quickly, work with businesses and non-profits, and host conferences without having

* Bostrom's introduction to this paper is the most arresting you'll read for some time. 'As I write these words, suns are illuminating and heating empty rooms, unused energy is being flushed down black holes, and our great common endowment of negentropy is being irreversibly degraded into entropy on a cosmic scale. These are resources that an advanced civilization could have used to create value-structures, such as sentient beings living worthwhile lives. The rate of this loss,' Bostrom continues, 'boggles the mind. One recent paper speculates, using loose theoretical considerations based on the rate of increase of entropy, that the loss of potential human lives in our own galactic supercluster is at least ~10^{46} per century of delayed colonization.' As he notes in *Deep Utopia*, we lose three galaxies per year to the expansion of the universe. Send a signal to those galaxies and it will never reach them. We are sundered forever.

to parlay any of this through the university's bureaucratic machinery. But the Faculty of Philosophy, by Bostrom's description, had 'a very different cultural mindset'. Its attitude to hiring, Bostrom told me, was rooted in a culture of teaching the same sort of philosophy – Aristotle, Plato et al – for centuries. '"We have this person who should teach ancient philosophy,"' he said, approximating the faculty view, '"and then when they retire, 40 years from now, we'll hire another person to teach ancient philosophy" … whereas our research agenda was very much designed to be flexible.' On one occasion, a former member of Bostrom's team told me, the faculty's slow movement cost the institute a new hire. Another alumnus wondered whether, in an academic world in which status is often zero-sum, FHI became a target of jealousy or resentment.

FHI might have been bureaucratically sinned against, but it was also a bureaucratic sinner. It seems to have had a reputation of being difficult to deal with, and – depending on whom you ask – of having management that thought itself above the petty demands of university bureaucracy. 'During my time,' a former senior staff member, Seán Ó hÉigeartaigh, wrote on the Effective Altruism forum, 'FHI constantly incurred heavy costs for being uncooperative.'[24] Its misdeeds, though minor, irritated the university. Staff were reproved for using Gmail instead of Outlook, for travelling without risk assessments, and so on.

The pairing of the institutions was an increasingly unwieldy one. FHI had become larger than the body that housed it, attracted more attention and funding, and employed many more non-philosophers than it did philosophers. Nor, it seems, was the faculty particularly enamoured of the institute. 'The impression I got,' I was told by a don from a different department, 'was that the philosophers' – that is, those within the faculty – 'didn't have much regard for it.'

Fundamentally, it seems, there was a mismatch between the way the organisations assessed the value of their work. 'The philosophy faculty's currency is peer-reviewed papers in

prestigious journals that get cited a lot,' said Niel Bowerman, who was assistant director of FHI at the time I visited, 'whereas that wasn't the currency of FHI. The currency was cool ideas that could improve the world.' The mismatch only increased as FHI became more high-profile and started to attract funding with fewer obligations.*

Multiple FHI staff told me that relations had worsened when a new faculty chair, Chris Timpson, arrived in 2018. Timpson chose not to comment when I attempted to speak to him. When I put a detailed set of questions to the university, I was issued with the same vague statement it had released on FHI's closure.† The eternal bureaucratic logjam was, in the view of FHI staff, having real consequences: not just in the more faraway astronomical waste sense, but in the sense of costing present-day lives. Jan Kulveit, who worked at FHI between 2018 and 2023, had led a successful COVID-19 forecasting project in the virus' first wave. He wanted to expand the project and provide medium-range forecasts for the whole world, warning about the potential second wave. The project was offered philanthropic funding, but it turned out it would be too bureaucratically difficult for the university to accept the grant, not least because it would have been used to hire external software engineers. The project expansion didn't happen. 'This was quite frustrating,' Kulveit told me. 'We had the best forecasters and some idea about what was coming. Overall, I think there was some

* In the early days, FHI was obliged to carry out insurance-related research as a result of its partial funding by the re-insurer MS Amlin. I'm told that Sandberg, true to character, managed to find the project ravishingly interesting.

† That vague statement is as follows: 'We regularly consider the best structures for conducting our academic research, as part of the University's governance processes. After such consideration, the decision was made to close the Future of Humanity Institute. The University recognises the Institute's important contribution to this emerging field, which researchers elsewhere across the University are likely to continue.'

mismatch between the timescale of the crisis and timescales on which things like processing grants in academia operate, and I became worried this would be true for some other risks FHI studied.'

FHI asked its biggest donor, Open Philanthropy, to put it to the university's vice-chancellor that the institute have more autonomy.* This, FHI staff told me, did not go down well with Oxford. Within FHI, there was plenty of frustration with the university, but there was frustration with Bostrom, too. In August 2021, Owen Cotton-Barratt, the architect of the Research Scholars Programme, quit FHI. In his resignation letter, addressed to Bostrom and shared with FHI staff and some allies of the institute, Cotton-Barratt praised Bostrom's intellectual leadership, but criticised his management.

On my request, Cotton-Barratt showed me the letter. Its tone was gentle and conciliatory, but its substance was serious. Bostrom was a bad delegator, Cotton-Barratt wrote, and disinclined to invest in communication with staff. By Cotton-Barratt's account, Bostrom prioritised his own research over FHI's relationship with the university; as a result, the institute had suffered. To the dismay of FHI staff, the letter eventually reached the philosophy faculty. (This was not Cotton-Barratt's doing.) Reflecting on the letter, Cotton-Barratt told me in 2024: 'I think Nick did a great job building FHI and the world has lost somewhere special. I wrote that letter in the hope that it might help FHI to iterate towards the best version of itself.' Bostrom viewed the letter as well-meaning, but 'a bit of a facepalm'. It arose from an effective altruist culture of extreme openness, he said, but it 'set us back administratively by about a year'.

After the worst of the pandemic, FHI staff re-assembled at new premises. The institute's home was now Trajan House, an office

* Open Philanthropy is the effective altruism-inspired philanthropic vehicle of Dustin Moskovitz and his wife Cari Tuna. Moskovitz was one of the co-founders of Facebook.

on the outskirts of Oxford. (It overlooked a graveyard, prompting Bostrom, true to his transhumanist roots, to call the building 'Corpsewatch Manor'.[25]) FHI was to share the building with the Centre for Effective Altruism (CEA) and similarly minded non-profits. There were Huel shakes in the fridge* and a nap room that university employees, such as FHI staff, were banned, for bureaucratic reasons, from using. A former CEA employee told me that the university and the Faculty of Philosophy had rarely shown much interest, or pride, in Effective Altruism, despite it being 'one of the biggest success stories in applied philosophy in maybe 100 years'. In the view of this employee, the university had 'a weird beef that has to be motivated by a personal grudge'. (It's also worth considering that EA's uncompromising utilitarianism has been known to get people's backs up.)

FHI's relocation to EA HQ, therefore, might not have been helpful for its relations with the university, but it made intellectual sense. When Toby Ord had helped found what we now know as EA, its concern had been poverty in the developing world. (Fin Moorhouse, who worked alongside him for two years as a young researcher, told me that Ord remains, in moral terms, 'the real deal' – ascetic, kind and intellectually earnest.) As technology such as AI advanced, EA had, like its co-progenitor Ord, embraced the idea that the key moral priority of our time is the protection and improvement of the long-run future.[26] This school of thought – longtermism – is indebted to the work of Derek Parfit, the economist John Broome and Bostrom.† Close to its heart is the endeavour to reduce existential risk. EA, a

* Huel is a meal replacement shake: beige in colour, chalky in constitution and near-impeccable in its environmental credentials. Its name is a portmanteau of 'human fuel'. As a younger man, I was forced by a newspaper to live off Huel for five days. I do not remember that experience particularly fondly.

† In 'The Case for Strong Longtermism', Hilary Greaves and Will MacAskill, two leading EA philosophers, refer specifically to Bostrom's paper 'Astronomical Waste'.

community whose existence was indebted to FHI staff, had, for several years, been turning its considerable resources and talent towards the reduction of existential risk – the overwhelming priority of FHI.

There was no whiteboard panopticon in FHI's new home, and the new office rooms, stacked for months with unopened cardboard boxes, weren't much more spacious than the last ones. Senior staff were seen on-site less often than they used to be, and the hiring freeze – which briefly thawed every now and then on a negotiated basis – still constrained FHI to a painful degree. Tanya Singh, in particular, was working hours that would put a paperclip maximiser to shame. She left in June 2022, the institute's headcount having fallen to the same number as it was in 2017, when she had arrived. But the ascent of Sam Bankman-Fried, an EA crypto mogul who had suddenly become the world's youngest billionaire, must have felt like an exceptionally promising development.

For a single, delirious summer, it seemed that every other person at Trajan House, including FHI staff, was to-ing and fro-ing between Oxford and the Bahamas. There, Bankman-Fried and his allies were planning a barrage of longtermist philanthropy. 'The mood was pretty buoyant,' recalled Lewis Hammond, an AI researcher who had joined FHI in 2019. Ord, apparently, was one of the few who evinced reluctance to get involved. In October 2022, however, Bankman-Fried's empire disintegrated. Millions of people had been swindled and Bankman-Fried's deceit had tarred his allies: EA, FHI and the many individuals associated with them. One can imagine the disgruntlement in the Ryle room. By now, FHI's seminars and salons were tailing off. When the Stakhanovite Singh departed, the faculty had seconded a part-time administrator who covered a fraction of her hours. 'It felt like FHI was dying a slow death,' said Hammond. Or, as Singh had it: 'Death by a thousand paper cuts.'

That death became a painful one. On 9 January 2023, Sandberg posted to Twitter a document written by Bostrom:

'Apology for an Old Email'. The email was sent in 1996 to the Extropians listserv. In a conversation about offensive communication styles, a 23-year-old Bostrom wrote that he had always liked 'the uncompromisingly objective way of speaking'. The more counterintuitive and repugnant a formulation, he wrote, the more it appealed to him, assuming it was correct.

As an example, he wrote that 'Blacks are more stupid than whites', commenting that 'I like that sentence and think it is true'. This did not mean, Bostrom remarked, that he disliked black people and thought it was right that they be treated badly. 'It's just that based on what I have read, I think it is probable that black people have a lower average IQ than mankind in general.' Yet the sentence he'd posited, he wrote, the claim that black people are stupider than whites, would be taken by most people as 'I hate those bloody niggers!!!!'. By Bostrom's account, he apologised for the email within 24 hours of sending it.

Seventeen years later, a longtermist-turned-critic, Émile P. Torres, found the email. Bostrom, informed that Torres was likely to publish the email, published it himself, with an apology, and disseminated the document via Sandberg (Bostrom does not use X/Twitter). 'I completely repudiate this disgusting email,' Bostrom wrote. Explaining his views, he said that 'it is deeply unfair that unequal access to education, nutrients, and basic healthcare leads to inequality in social outcomes, including sometimes disparities in skills and cognitive capacity'. Whether there are genetic or epigenetic contributors to differences between groups in cognitive abilities, he wrote, was not his area of expertise. Torres, and other critics, took the email and apology as evidence for their case that longtermism was motivated by a hateful form of eugenics.

The university suspended Bostrom while it and the faculty investigated him. Some FHI staff, along with outsiders, disapproved of the apology. It was 'insensitive', an FHI alumnus told me, and the controversy as a whole – both the original email and the debate over the apology – had further damaged FHI's reputation. Bostrom, meanwhile, was forbidden to contact his

colleagues, or undertake anything FHI-related, for the duration of his suspension. By now, the existence of FHI was plainly fragile. Bostrom's staff wanted clarity from the university, but the university, I was told, refused to speak to them. This is because it was a matter for the director – whom the university had exiled. 'And contrary to popular conception,' says a former member of FHI staff, 'Nick is an extreme stickler for rules.'

The outcome of the investigation, though, was in Bostrom's favour. 'We do not consider you to be a racist or that you hold racist views,' a representative of the university told Bostrom in August 2023, 'and we consider that the apology you posted in January 2023 was sincere.' He returned to official duties.

Throughout this difficult later era for FHI, senior staff hunted for a solution. They discussed the idea of Bostrom retaining directorship of the institute's research, but handing over management to a CEO. They mooted becoming part of one of the university's constituent colleges, or spinning out of the university altogether. They were even in discussions with the Faculty of Physics over an interdepartmental transfer. Yet none of these plans became reality. Staff valued their association with Oxford, making them reluctant to leave the university. As for the interdepartmental transfer, FHI staff suspected that the plan was sabotaged by the Faculty of Philosophy.

'We were in this weird limbo for ages,' Lewis Hammond said. But one day in March 2024, while in Trajan House, Hammond was interrupted at his desk by a university IT worker. 'FHI is closing down,' he was told. 'You're going to need to give that computer back.' The university was finally swinging the axe. On 16 April 2024, FHI was shuttered.*

* I noted at the beginning of the chapter that FHI's detractors had mocked the organisation's lack of foresight into its own demise. Yet the institute had, in fact, foreseen its doom. An FHI veteran disclosed that he had, in 2019, been party to a brainstorming session regarding existential risks to FHI as

The reason given by Oxford, Bostrom reported, was that the university did not have the operational bandwidth to manage FHI. He told me that he asked whether the email incident had anything to do with the decision and that the university said it had not. Oxford has not contradicted this account.

'I wasn't even surprised that I was hearing it from the IT guy,' said Hammond. 'It felt very emblematic of how things were being communicated and organised at that point.' He and his colleagues left the building, and the university staff forbade the other residents of Trajan House from using the empty room. Bostrom, meanwhile, was abroad, working on a new book in an Alpine chalet. FHI's final demise, he later told me, came as a relief. 'I'd imagined the faculty would, as it were, let us bleed out,' he said. After receiving the news, Bostrom returned to work.

Back in Oxford, Toby Ord organised a pub trip that staff now refer to as FHI's 'wake'. The remaining staff, and some friends of the institute, met at the Holly Bush to mark the end of FHI's 19-year-life. Ord delivered the eulogy. FHI's death had been 'over-determined', he said, but it had far outlasted the three-year period for which it had been initially funded. The institute had achieved a lot, and its offspring were alive and well. Maybe there wouldn't be anything that would replace the particular kind of organisation that FHI was, Ord concluded, but maybe that was okay. After the wake, he updated FHI's website with the statement announcing its closure.

* * *

an organisation. 'Basically, it was a big Google doc in which everyone dumped ideas, which could then be up- or down-voted. Many people suggested some version of "administrative headwinds" (including more specific things like the hiring freeze). Amusingly, someone also raised the risk that "too much of the research ends up on private Google docs and so doesn't have much wider impact". This received 10 upvotes (the most any single suggestion received was 15) and appears to have been the fate of the document in question.'

After a funeral come the obituaries. Sandberg wrote FHI's official retrospective;[27] Moorhouse wrote a valediction of his own.[28] Fans turned one of Bostrom's poems[29], one that portrayed super-heroic FHI staff blending in among tweedy dons, into a rousing club anthem.[30] (FHI was a 'misfit prodigy', Bostrom wrote, contending with 'thousand meetings, thousand emails, thousand rules'.) The *Guardian*, making the most of Musk's single donation to FHI, took the opportunity to invoke its favourite folk devil – 'Oxford shuts down institute run by Elon Musk-backed philosopher,' went the headline – and retold the stories of EA's various scandals.[31]

As Ord had pointed out at the wake, FHI was survived by many offspring. Who are they? To the extent that FHI is their parent – it might otherwise be called their midwife – there are rationalism, effective altruism and longtermism. The latter two movements, which overlap enormously, have been responsible for a colossal amount of work and funding being directed towards pandemic prevention, technical research into AI safety and other FHI-style endeavours. More specifically, there are the projects and organisations, such as the Centre for the Governance of AI (GovAI), that span directly out of FHI. GovAI helps shape the way society is managing the development of artificial intelligence, as do the dozens of FHI alumni who now work at the leading commercial labs, in think tanks and in related government agencies. Similarly, what was once an apparently wacky FHI idea – the theory that we should explore the moral patienthood of forthcoming AIs that might be able to feel as well as think – is now being taken more seriously. Bostrom had encouraged his staff to start work on this nascent field.

Had FHI never existed, and therefore failed to lay the early groundwork, the world would still have invented AI governance and would still have got round to considering the possibility of digital sentience. It would be even shorter of talent and expertise, though, and it would have been slower to wake up to the threats of rogue and misused AI. Bostrom and his team are

widely and deservedly credited not only with making an effective case for taking AI safety seriously, but for making the study of existential risk a serious academic business. Where FHI was, at the time of its inception, the world's only institute of its kind, it grew to have several cousins. It's worth noting that one of these organisations has survived the Faculty of Philosophy's bureaucracy – namely the Global Priorities Institute. GPI, as it's known, is another relatively young big-picture sort of institute; it's also the one that employs Will MacAskill, the philosopher who is seen, with Ord, as the co-founder of the Effective Altruism movement. GPI is less zany than FHI, and more academically legible.

An academic with links to the Faculty of Philosophy contrasted the fate of FHI with the continued existence of other such institutes. The academic told me: 'I'm not sure if we should find it amusing or terrifying, regarding humanity's prospects of surviving global threats, that this research SAS team was killed by paperwork.' There has also been a proposal to found an 'FHI of the West' – that is, the Western hemisphere – in Berkeley, California.[32] Sandberg told me that our current culture is too risk-averse and pessimistic, and that we might make better progress if we have more things like FHI in the world, rather than fewer. 'Historically, clusters of intense creative intellectual work have counted for a lot – the cafés of 1920s Vienna, Xerox PARC, Los Alamos, various Enlightenment era salons, those golden decades in Athens, and so on. I'm not claiming FHI was anywhere near that level, but when smart people come together and begin bouncing ideas off each other and have the wherewithal to make things, then breakthroughs and new visions bloom. And we need breakthroughs and new visions.'

As for the existing organisations, none of them can claim to be half as philosophically fecund as FHI was. It started early: in 2006, FHI staff helped birth the Rationalist community by co-founding the blog Overcoming Bias, from which arose the forum LessWrong.[33] This forum would become the battlefield

for different schools of AI safety, a community whose story I will tell in Chapter 13. Philosophy around AI, alien life and the far, far future is now replete with FHI ideas and coinages. 'Note how much of the literal terminology was coined on (one imagines) a whiteboard,' wrote Fin Moorhouse, who, apart from assisting Ord, conducted research into space governance and into institutions that might support the interests of future generations.[34] He will look back on his time at FHI 'with fondness', he told me. 'It was a crazy and interesting thing to be part of.' Lewis Hammond spoke of FHI in similar terms. 'Overall,' he said, 'my main feeling looking back on FHI is one of gratitude. It was both exciting and humbling to be surrounded by such thoughtful people and important ideas.'

Appropriately for an institute founded by a transhumanist, FHI is living many afterlives. These successes are part of a rich legacy for FHI, though its counterfactual impact, to use another FHI-ish phrase, is not straightforward to analyse. As Bostrom told me, 'it's like the comment about the French Revolution, right? Was it good or bad? It's too early to tell.' On one hand, FHI has, undeniably, led the way in developing our species' understanding of the risks we face and the destinies that might await us. On the other hand, it's possible that CEOs such as Sam Altman, he of the enthusiastic review of *Superintelligence*, might be less alive to the power of advanced general intelligence (AGI) had Bostrom illustrated it less vividly. One of the most vivid of those illustrations was that of the paperclip maximiser, which epitomised a style of forecast about AI that Seth Baum, of the GCRI think tank, believes to be outmoded. The current direction for AI, Baum tells me, seems to be towards a plurality of systems rather than a singular one that, in the style of the paperclip maximiser, breaks out and takes over the world.

'It feels a little harsh to call that a mistake,' said Baum of the forecast of highly agentic AI. At the point when Bostrom was dragging these ideas into the public sphere, Baum explains,

meaningful AI was yet to exist. 'And who knows? Maybe the technology will take a turn in a different direction, and now their ideas will seem correct again.' Baum's view was that FHI's years of academic graft compared favourably to the relative dilettantism of commentators such as Stephen Hawking, Jared Diamond and Steven Pinker; each has written on societal collapse and humanity's destiny without acquiring deep expertise in those topics. 'To FHI's credit, and to Nick Bostrom's credit, they have backed up their public prominence with significant research.'

Not everyone agrees that FHI's work was of good quality. Carla Cremer was once an FHI research scholar and is now a softly spoken apostate. 'Research out of FHI might been impactful,' she told me, 'but I think a good part of it was of low quality and maybe even harmful for understanding and acting on global risk.' Cremer told me that one ought not to think of existential risks as discrete, separating them into buckets such as AI, biological threats, climate change and so on. 'Instead,' she said, 'we should think about a level of X-risk* in the system that is heightened by the various actors, institutions, and dynamics that produce this kind of risk.'

To borrow the terminology of the philosopher Daniel Schmactenberger, these actors, institutions and dynamics together comprise 'production functions'. If we apply this way of thinking to catastrophic risks, we might consider zoonotic pandemics – those that originally arise in the animals, such as the billions imprisoned in factory farms. To contain the damage of such pandemics, the X-risk community champions such innovations as next-generation personal protective equipment. But if we look upstream, we will see that society has made zoonotic pandemics much more frequent and severe. The rewards of factory farming are concentrated, making the practice attractive, but the risks are dispersed across humanity. As a result,

* A common abbreviation for existential risk.

factory farms have caused many outbreaks of disease, but we have ignored these signals.

Responding to that example, Cremer said that such situations are analogous to 'a human brain overriding bodily signals of stress'. In the same way, she said, Bostrom's autocratic leadership overrode the signals that things were amiss at FHI. His stubbornness, she said, 'is a quality. It is a quality that you might want in a philosopher, but you might not want it as the head of an institute that needs to navigate a risky landscape that is deeply political.' (The usual FHI response to this sort of criticism is that looking upstream is valid, but unlikely to yield immediate results, especially when conducted by a small institution.)

Oxford's 'administrative headwinds', having blown down FHI, have also dispersed the band of researchers that made FHI famous. Shattered is the carapace, scattered are the jewels – though they have largely ended up in the same field. Stuart Armstrong is the co-founder, and chief mathematician, of a start-up that is trying to develop fundamentally safe artificial intelligence. Eric Drexler continues to work on AI governance and strategy. Anders Sandberg, at the time of writing, was soon to publish *Grand Futures*, a reputedly gigantic tome in which he will map the physical limits of what advanced civilisations can achieve,* and has joined the newly-founded Mimir Center for Long Term Futures Research. Mimir the Norse god was beheaded in wartime before being relied on (while still beheaded) by Odin for his wise counsel. Mimir the independent research institute is based in Stockholm, and looks like it will afford Sandberg a similarly eclectic research brief to the one in which he revelled at FHI. (I'm reminded of the Galápagos finches, of which each species has developed a beak that works perfectly for the food sources on their respective islands, but not

* If you think this book will be about collecting energy from the sun by encasing it in a Dyson sphere, you are not thinking big enough.

elsewhere. Sandberg, improbably, has managed to change islands without having to change his diet.)

At least some of the old band – including Toby Ord and Fin Moorhouse – have, as of 2025, got back together at a nonprofit called the Forethought Foundation. Forethought is a nonprofit whose research focuses on the world's transition to superintelligent AI systems. Forethought, which also employs Will MacAskill, is based in Oxford, but is not affiliated to the university.

Speaking of which: within four months of closing FHI, Oxford announced the establishment of the Human-Centered AI Lab, promising 'a vibrant community for big-picture thinking about a future of AI that enhances human flourishing.' The description made the lab, which would be housed within the Faculty of Philosophy, sound much like the institute we can assume it was designed to replace, albeit with two small exceptions: an entirely new set of staff, and much more obeisance paid to classical philosophy.

When I spoke to Bostrom in 2024, he was midway through the publicity campaign for his own new book, *Deep Utopia*.[35] In the book, Bostrom considers a world in which the development of super-intelligent AI has gone well. Some observers, he told me, have assumed that the penning of a more optimistic follow-up to *Superintelligence* suggests that Bostrom feels a greater bullishness about humanity's prospect of surviving and thriving. Alas. 'We can see the thing with more clarity now,' said Bostrom, 'but there has been no fundamental shift in my thinking.' When he wrote *Superintelligence*, he said, there seemed an urgent need to explore the risks of advanced AI and to catalyse work that might address those risks. 'There seemed less urgency to develop a very granular picture of what the upside could be. And now it seems like time to maybe fill in that other part of the map a bit more.'

Bostrom was characteristically serious, though he occasionally wove into his speech a droll turn of phrase. He was frank in his assessment of the 'intolerable' situation with the faculty, but

gentle in his comments on the individuals who worked for it. He was pleased with FHI's development and dissemination of concepts relating to existential risk, and he said of James Martin, whom he'd sat next to at dinner all those years ago, that 'I think we lived up to the spirit of what he had in mind'. Bostrom also expressed his sense of privilege at having worked in such an intellectually lively environment with such talented people.

It was in this conversation that he made his comment about it being too early to tell, as with the French Revolution, whether FHI was a success. Elaborating, he said: 'It's very difficult to know how, ultimately, this shakes out with reactions and counter-reactions and counter-reactions to the counter-reactions and stuff. But if one thinks that it's better for humanity to have some ability to reflect and think and deliberate about these things, then I think we certainly helped accelerate that.'

In that respect, Bostrom's leadership of FHI indisputably changed the world. How did it change him? I expected Bostrom to view this question as a trivial one, but he found in it a kernel of a philosophical question. 'It would be kind of interesting if you could somehow enter a time machine and meet your earlier self. How would they strike you? Because, in reality, things change gradually.' He picked out his 'intensified sense of humility in relation to the big picture – just how profoundly we are in the dark regarding these things. It's not as if we know nothing – we know a whole bunch of things – but it might well be that the things we don't yet know are such that they would radically change the bottom line, as it were, of the overall picture. We are small, and kind of limited in insight and abilities.' He would have felt the same when he founded FHI, he said. The change is that he is now 'inhabiting that sense more actively, perhaps'.

I wondered what awaited him on the other side of the publicity campaign. 'Well,' he said, beginning to smile, 'I'm now a free man, which I very much enjoy. So I'm not in a hurry to put on the chains of bondage again. I can't imagine going back in time. I feel I've had my fill of academia.'

Don't expect to see Bostrom on the golf course. There is no prospect that he will leave his field of study. But *Deep Utopia* might well be the last of his books; they take years to write and Bostrom views shorter projects as more appropriate for our current point in history. AGI is coming, and time is short.

CHAPTER 8: THE MISFIT PRODIGY

1. 'Future of Humanity Institute', futureofhumanityinstitute.org, 17 April 2024. (https:/www.futureofhumanityinstitute.org/)
2. Khatchadourian, R. 'The Doomsday Invention', *New Yorker*, 16 November 2015. (https://www.newyorker.com/magazine/2015/11/23/doomsday-invention-artificial-intelligence-nick-bostrom)
3. Bostrom, N. 'Nick Bostrom – Curriculum Vitae', accessed 15 February 2025. (https://nickbostrom.com/cv.pdf)
4. Ibid.
5. Bostrom, N. 'Existential Risks: Analyzing Human Extinction Scenarios and Related Hazards', *Journal of Evolution and Technology*, 9 (2002). (https://nickbostrom.com/existential/risks.pdf)

6. Bostrom, N. 'Are You Living in a Computer Simulation?', *Philosophical Quarterly*, 53 (211) (2003), pp. 243–255. (https://simulation-argument.com/simulation.pdf)
7. Ibid.
8. Ibid.
9. Williamson, M. 'Dr James Martin: Technology Guru and Philanthropist Who Predicted the Internet', *Independent*, 28 June 2013. (https://www.independent.co.uk/news/obituaries/dr-james-martin-technology-guru-and-philanthropist-who-predicted-the-rise-of-the-internet-8679451.html)
10. 'Dr James Martin – Our Founder', Oxford Martin School, accessed 15 February 2025. (https://www.oxfordmartin.ox.ac.uk/about/founder)
11. Edmonds, D. *Parfit: A Philosopher and His Mission to Save Morality*, Princeton, Princeton University Press, 2023.
12. Bostrom, N. and Sandberg, A. 'Whole Brain Emulation: A Roadmap', Future of Humanity Institute, Oxford University, 2008. (https://www.fhi.ox.ac.uk/brain-emulation-roadmap-report.pdf)
13. Bostrom, N. and Yudkowsky, E. 'The Ethics of Artificial Intelligence', in Ramsey, W. and Frankish, K. (eds). *Cambridge Handbook of Artificial Intelligence*, Cambridge University Press, 2011. (https://nickbostrom.com/ethics/artificial-intelligence.pdf)
14. Macaskill, W. 'Getting Inspired by Cost-Effective Giving', The Life You Can Save, 20 May 2013. (https://www.thelifeyoucansave.org/supporters-stories/getting-inspired-by-cost-effective-giving/)
15. Regis, E. 'The Incredible Shrinking Man', *Wired*, 1 October 2004. (https://www.wired.com/2004/10/drexler/)
16. Sandberg, A. 'Future of Humanity Institute 2005–2024: Final Report', University of Oxford, accessed 16 February 2025. (https://static1.squarespace.com/static/660e95991cf0293c2463bcc8/t/661a3fc3cecceb2b8ffce80d/1712996303164/FHI+Final+Report.pdf)
17. Sandberg, A. 'Andart II – Part of Anders' Exoself', aleph.se, 30 October 2021. (http://aleph.se/andart2/)
18. Bostrom, N. 'The Unfinished Fable of the Sparrows', OUP blog, 29 August 2014. (https://blog.oup.com/2014/08/unfinished-fable-sparrows-superintelligence/)
19. 'Elon Musk Funds Oxford and Cambridge University Research on Safe and Beneficial Artificial Intelligence', The Future of Humanity Institute, 1 July 2015. (https://www.fhi.ox.ac.uk/elon-musk-funds-oxford-and-cambridge-university-research-on-safe-and-beneficial-artificial-intelligence/)

20. Altman, S. 'Machine Intelligence, Part 1', blog.samaltman.com, 25 February 2015. (https://blog.samaltman.com/machine-intelligence-part-1)
21. 'Nick Bostrom Sets out Threats from Future Technologies at UN Meeting', Oxford Martin School, 12 October 2015. (https://www.oxfordmartin.ox.ac.uk/news/201510-bostrom-crbn-un)
22. Lin, S, et al. 'TruthfulQA: Measuring How Models Mimic Human Falsehoods', arXiv, 8 May 2022. (https://doi.org/10.48550/arXiv.2109.07958)
23. 'Home', AI Impacts, 19 December 2014. (https://aiimpacts.org/)
24. 'Future of Humanity Institute 2005–2024: Final Report', Effective Altruism Forum, 17 April 2024. (https://forum.effectivealtruism.org/posts/uK27pds7J36asqJPt/future-of-humanity-institute-2005-2024-final-report)
25. Lewis-Kraus, G. 'The Reluctant Prophet of Effective Altruism', *New Yorker*, 8 August 2022. (https://www.newyorker.com/magazine/2022/08/15/the-reluctant-prophet-of-effective-altruism)
26. MacAskill, W. 'Longtermism', Effective Altruism Forum, 25 July 2019. (https://forum.effectivealtruism.org/posts/qZyshHCNkjs3TvSem/longtermism)
27. Sandberg, A. 'Future of Humanity Institute 2005–024: Final Report', accessed 16 February 2025
28. Moorhouse, F. 'Goodbye to FHI', accessed 16 February 2025. (https://finmoorhouse.com/writing/fhi/)
29. Bostrom, N. 'Poetry', accessed 16 February 2025. (https://nickbostrom.com/poetry/poetry)
30. 'FHI at Oxford (Feat. Nick Bostrom)', YouTube, 2024. (https://www.youtube.com/watch?v=YLHS43Wo1As)
31. Robins-Early, N. 'Oxford Shuts down Institute Run by Elon Musk-Backed Philosopher', *Guardian*, 19 April 2024. (https://www.theguardian.com/technology/2024/apr/19/oxford-future-of-humanity-institute-closes)
32. 'Express Interest in an "FHI of the West"', LessWrong, 18 April 2024. (https://www.lesswrong.com/posts/ydheLNeWzgbco2FTb/express-interest-in-an-fhi-of-the-west)
33. Overcoming Bias, 5 January 2007. (https://web.archive.org/web/20070105153523/https:/www.overcomingbias.com/)
34. Hale, T. et al. 'Toward a Declaration on Future Generations', Blavatnik School of Government, University of Oxford, 12 January 2023. (http://www.bsg.ox.ac.uk/research/publications/toward-declaration-future-generations)

35. Bostrom, N. *Deep Utopia: Life and Meaning in a Solved World*, IdeaPress Publishing, 2024.