‘Eugenics on steroids’: the toxic and contested legacy of Oxford’s Future of Humanity Institute

Publication Date: 2024-04-28

Author: Andrew Anthony

Section: Technology

Tags: Technology, The Observer, Academics, Artificial intelligence (AI), Philosophy, Philosophy books, University of Oxford, features

Article URL: https://www.theguardian.com/technology/2024/apr/28/nick-bostrom-controversial-future-of-humanity-institute-closure-longtermism-affective-altruism

Two weeks ago it was quietly announced that the Future of Humanity Institute, the renowned multidisciplinary research centre in Oxford, no longer had a future. It shut down without warning on 16 April. Initially there was just a brief statement on its website stating it had closed and that its research may continue elsewhere within and outside the university. The institute, which was dedicated to studying existential risks to humanity, was founded in 2005 by the Swedish-born philosopher Nick Bostrom and quickly made a name for itself beyond academic circles – particularly in Silicon Valley, where a number of tech billionaires sang its praises and provided financial support. Bostrom is perhaps best known for his bestselling 2014 book Superintelligence, which warned of the existential dangers of artificial intelligence, but he also gained widespread recognition for his 2003 academic paper “Are You Living in a Computer Simulation?”. The paper argued that over time humans were likely to develop the ability to make simulations that were indistinguishable from reality, and if this was the case, it was possible that it had already happened and that we are the simulations. I interviewed Bostrom more than a decade ago, and he possessed one of those elusive, rather abstract personalities that perhaps lend credence to the simulation theory. With his pale complexion and reputation for working through the night, he looked like the kind of guy who didn’t get out much. The institute seems to have recognised this social shortcoming in its final report, a long epitaph written by FHI research fellow Anders Sandberg, which stated: “We did not invest enough in university politics and sociality to form a long-term stable relationship with our faculty… When epistemic and communicative practices diverge too much, misunderstandings proliferate.” Like Sandberg, Bostrom has advocated transhumanism, the belief in using advanced technologies to enhance longevity and cognition, and is said to have signed up for cryogenic preservation. Although proudly provocative on the page, he was wary and defensive in person, as if he were privy to an earth-shattering truth that required vigilant protection. His office, located in a medieval backstreet, was a typically cramped Oxford affair, and it would have been easy to dismiss the institute as a whimsical undertaking, an eccentric, if laudable, field of study for those, like Bostrom, with a penchant for science fiction. But even a decade ago, when I paid my visit, the FHI was already on its way to becoming the billionaire tech bros’ favourite research group. In 2018 it received £13.3m from the Open Philanthropy Project, a non-profit organisation backed by Facebook co-founder Dustin Moskovitz. And Elon Musk has also been a benefactor. Bostrom’s warnings on AI were taken seriously by big tech. But as competition has heated up in the race to create a general artificial intelligence, ethics have tended to take a back-seat. Among the other ideas and movements that have emerged from the FHI are

longtermism – the notion that humanity should prioritise the needs of the distant future because it theoretically contains hugely more lives than the present – and effective altruism (EA), a utilitarian approach to maximising global good. These philosophies, which have intermarried, inspired something of a cult-like following, which may have alienated many in the wider philosophy community in Oxford, and indeed among the university's administrators. According to the FHI itself, its closure was a result of growing administrative tensions with Oxford's faculty of philosophy. "Starting in 2020, the Faculty imposed a freeze on fundraising and hiring. In late 2023, the Faculty of Philosophy decided that the contracts of the remaining FHI staff would not be renewed," the final report stated. But both Bostrom and the institute, which brought together philosophers, computer scientists, mathematicians and economists, have been subject to a number of controversies in recent years. Fifteen months ago Bostrom was forced to issue an apology for comments he'd made in a group email back in 1996, when he was a 23-year-old postgraduate student at the London School of Economics. In the retrieved message Bostrom used the N-word and argued that white people were more intelligent than black people. The apology did little to placate Bostrom's critics, not least because he conspicuously failed to withdraw his central contention regarding race and intelligence, and seemed to make a partial defence of eugenics. Although, after an investigation, Oxford University did accept that Bostrom was not a racist, the whole episode left a stain on the institute's reputation at a time when issues of anti-racism and decolonisation have become critically important to many university departments. It was Émile Torres, a former adherent of longtermism who has become its most outspoken critic, who unearthed the 1996 email. Torres says that it's their understanding that it "was the last straw for the Oxford philosophy department". Torres has come to believe that the work of the FHI and its offshoots amounts to what they call a "noxious ideology" and "eugenics on steroids". They refuse to see Bostrom's 1996 comments as poorly worded juvenilia, but indicative of a brutal utilitarian view of humanity. Torres notes that six years after the email thread, Bostrom wrote a paper on existential risk that helped launch the longtermist movement, in which he discusses "dysgenic pressures" – dysgenic is the opposite of eugenic. Bostrom wrote: "Currently it seems that there is a negative correlation in some places between intellectual achievement and fertility. If such selection were to operate over a long period of time, we might evolve into a less brainy but more fertile species, homo philoprogenitus ('lover of many offspring')." Bostrom now says that he doesn't have any particular interest in the race question, and he's happy to leave it to others with "more relevant knowledge". But the 28-year-old email is not the only issue that Oxford has had to consider. As Torres says, the effective altruism/longtermist movement has "suffered a number of scandals since late 2022". Just a month before Bostrom's incendiary comments came to light, the cryptocurrency entrepreneur Sam Bankman-Fried was extradited from the Bahamas to face charges in the US relating to a multibillion-dollar fraud. Bankman-Fried was a vocal and financial supporter of effective altruism and a close friend of William MacAskill, an academic who has strong links to the FHI and who set up the Centre for Effective Altruism, where Bankman-Fried worked briefly. It was MacAskill who was said to have persuaded Bankman-Fried a decade ago to seek to earn as much money as possible so that he could give it away. The entrepreneur seemed to follow the first part of that injunction, but then went on to spend $300m in fraudulently earned money on Bahamian real estate. His downfall and subsequent 25-year prison sentence have done little for the moral arguments put forward by the FHI and its associate groups. If that wasn't enough, the coup last November that briefly dislodged Sam Altman as the CEO of Open AI, the company behind ChatGPT, was attributed to company board members who were supporters of EA. Altman's speedy return was seen as a defeat for the EA community, and, says Torres, "has seriously undermined the influence of EA/longtermism within Silicon Valley". All of this, of course, seems a long way from the not insubstantial matter of preserving humanity, which is the cause for which the FHI was ostensibly set up. No doubt that noble endeavour will find other academic avenues to explore, but perhaps without the cultish ideological framework that left the institute with a bright future behind it.