Against longtermism https://aeon.co/essays/why-longtermism-is-the-worlds-most-dangerous-secular-credo

It started as a fringe philosophical theory about humanity's future. It's now richly funded and increasingly dangerous

Scarecrows keep away migratory birds from the dangers of the tailing ponds created by the exploitation on the tar sands at Fort McMurray, Alberta, Canada. *Photo by Larry Towell/Magnum*

Émile P Torresis a PhD candidate in philosophy at Leibniz Universität Hannover in Germany. Their writing has appeared in *Philosophy Now*, *Nautilus*, Motherboard and the *Bulletin of the Atomic Scientists*, among others. They are the author of *The End: What Science and Religion Tell Us About the Apocalypse* (2016), *Morality, Foresight, and Human Flourishing: An Introduction to Existential Risks* (2017) and *Human Extinction: A History of the Science and Ethics of Annihilation* (forthcoming from Routledge).

There seems to be a growing recognition that humanity might be approaching the 'end times'. Dire predictions of catastrophe clutter the news. Social media videos of hellish wildfires, devastating floods and hospitals overflowing with COVID-19 patients dominate our timelines. Extinction Rebellion activists are shutting down cities in a desperate attempt to save the world. One survey even found that more than half of the people asked about humanity's future 'rated the risk of our way of life ending within the next 100 years at 50 per cent or greater.'

'Apocalypticism', or the belief that the end times are imminent, is of course nothing new: people have warned that the end is nigh for millennia, and in fact many New Testament scholars believe that Jesus himself expected the world to end during his own lifetime. But the situation today is fundamentally different than in the past. The 'eschatological' scenarios now being discussed are based not on the revelations of religious prophets, or secular metanarratives of human history (as in the case of Marxism), but on robust scientific conclusions defended by leading experts in fields such as climatology, ecology, epidemiology and so on.

We know, for example, that climate change poses a dire threat to civilisation. We know that biodiversity loss and the sixth mass extinction could precipitate sudden, irreversible, catastrophic shifts in the global ecosystem. A thermonuclear exchange could blot out the Sun for years or decades, bringing about the collapse of global agriculture. And whether or not SARS-CoV-2 came from a Wuhan laboratory or was cooked up in the kitchen of nature (the latter seems more probable right now), synthetic biology will soon enable bad actors to design pathogens far more lethal and contagious than anything Darwinian evolution could possibly invent. Some philosophers and scientists have also begun sounding the alarm about 'emerging threats' associated with machine superintelligence, molecular nanotechnology and stratospheric geoengineering, which look no less formidable.

Such considerations have led many scholars to acknowledge that, as Stephen Hawking wrote in *The Guardian* in 2016, 'we are at the most dangerous moment in the development of humanity.' Lord Martin Rees, for example, estimates that civilisation has a 50/50 chance of making it to 2100. Noam Chomsky argues that the risk of annihilation is currently 'unprecedented in the history of *Homo sapiens*'. And Max Tegmark contends that 'it's probably going to be within our lifetimes … that we're either going to self-destruct or get our act together.' Consistent with these dismal declarations, the *Bulletin of the Atomic Scientists* in 2020 set its iconic Doomsday

Clock to a mere 100 seconds before midnight (or doom), the closest it's been since the clock was created in 1947, and more than 11,000 scientists from around the world signed an article in 2020 stating 'clearly and unequivocally that planet Earth is facing a climate emergency', and without 'an immense increase of scale in endeavours to conserve our biosphere [we risk] untold suffering due to the climate crisis.' As the young climate activist Xiye Bastida summed up this existential mood in a *Teen Vogue* interview in 2019, the aim is to 'make sure that we're not the last generation', because this now appears to be a very real possibility.

Given the unprecedented dangers facing humanity today, one might expect philosophers to have spilled a considerable amount of ink on the ethical implications of our extinction, or related scenarios such as the permanent collapse of civilisation. How morally bad (or good) would our disappearance be, and for what reasons? Would it be wrong to prevent future generations from coming into existence? Does the value of past sacrifices, struggles and strivings depend on humanity continuing to exist for as long as Earth, or the Universe more generally, remains habitable?

Yet this is not the case: the topic of our extinction has received little sustained attention from philosophers until recently, and even now remains at the fringe of philosophical discussion and debate. On the whole, they have been preoccupied with other matters. However, there is one notable exception to this rule: over the past two decades, a small group of theorists mostly based in Oxford have been busy working out the details of a new moral worldview called longtermism, which emphasizes how our actions affect the very long-term future of the universe – thousands, millions, billions, and even trillions of years from now. This has roots in the work of Nick Bostrom, who founded the grandiosely named Future of Humanity Institute (FHI) in 2005, and Nick Beckstead, a research associate at FHI and a programme officer at Open Philanthropy. It has been defended most publicly by the FHI philosopher Toby Ord, author of *The Precipice: Existential Risk and the Future of Humanity* (2020). Longtermism is the primary research focus of both the Global Priorities Institute (GPI), an FHI-linked organisation directed by Hilary Greaves, and the Forethought Foundation, run by William MacAskill, who also holds positions at FHI and GPI. Adding to the tangle of titles, names, institutes and acronyms, longtermism is one of the main 'cause areas' of the so-called effective altruism (EA) movement, which was introduced by Ord in around 2011 and now boasts of having a mind-boggling $46 billion in committed funding.

It is difficult to overstate how influential longtermism has become. Karl Marx in 1845 declared that the point of philosophy isn't merely to interpret the world but change it, and this is exactly what longtermists have been doing, with extraordinary success. Consider that Elon Musk, who has cited and endorsed Bostrom's work, has donated $1.5 million dollars to FHI through its sister organisation, the even more grandiosely named Future of Life Institute (FLI). This was cofounded by the multimillionaire tech entrepreneur Jaan Tallinn, who, as I recently noted, doesn't believe that climate change poses an 'existential risk' to humanity because of his adherence to the longtermist ideology.

Meanwhile, the billionaire libertarian and Donald Trump supporter Peter Thiel, who once gave the keynote address at an EA conference, has donated large sums of money to the Machine Intelligence Research Institute, whose mission to save humanity from superintelligent machines is deeply intertwined with longtermist values. Other organisations such as GPI and the Forethought Foundation are funding essay contests and scholarships in an effort to draw young people into the community, while it's an open secret that the Washington, DC-based Center for Security and Emerging Technologies (CSET) aims to place longtermists within high-level US government positions to shape national policy. In fact, CSET was established by Jason Matheny, a former research assistant at FHI who's now the deputy assistant to US President Joe Biden for technology and national security. Ord himself has, astonishingly for a philosopher, 'advised the World Health Organization, the World Bank, the World Economic Forum, the US National Intelligence Council, the UK Prime Minister's Office, Cabinet Office, and Government Office

for Science', and he recently contributed to a report from the Secretary-General of the United Nations that specifically mentions 'long-termism'.

The point is that longtermism might be one of the most influential ideologies that few people outside of elite universities and Silicon Valley have ever heard about. I believe this needs to change because, as a former longtermist who published an entire book four years ago in defence of the general idea, I have come to see this worldview as quite possibly the most dangerous secular belief system in the world today. But to understand the nature of the beast, we need to first dissect it, examining its anatomical features and physiological functions.

The initial thing to notice is that longtermism, as proposed by Bostrom and Beckstead, is not equivalent to 'caring about the long term' or 'valuing the wellbeing of future generations'. It goes way beyond this. At its core is a simple – albeit flawed, in my opinion – analogy between individual persons and humanity as a whole. To illustrate the idea, consider the case of Frank Ramsey, a scholar at the University of Cambridge widely considered by his peers as among his generation's most exceptional minds. 'There was something of Newton about him,' the belletrist Lytton Strachey once said. G E Moore wrote of Ramsey's 'very exceptional brilliance'. And John Maynard Keynes described a paper of Ramsey's as 'one of the most remarkable contributions to mathematical economics ever made'.

But Ramsey's story isn't a happy one. On 19 January 1930, he died in a London hospital following a surgical procedure, the likely cause of death being a liver infection from swimming in the River Cam, which winds its way through Cambridge. Ramsey was only 26 years old.

One could argue that there are two distinct reasons this outcome was tragic. The first is the most obvious: it cut short Ramsey's life, depriving him of everything he could have experienced had he survived – the joys and happiness, the love and friendship: all that makes life worth living. In this sense, Ramsey's early demise was a personal tragedy. But, secondly, his death also robbed the world of an intellectual superstar apparently destined to make even more extraordinary contributions to human knowledge. 'The number of trails Ramsey laid was remarkable,' writes Sir Partha Dasgupta. But how many more trails might he have blazed? 'The loss to your generation is agonising to think of,' Strachey lamented, 'what a light has gone out' – which leaves one wondering how Western intellectual history might have been different if Ramsey hadn't died so young. From this perspective, one could argue that, although the personal tragedy of Ramsey's death was truly terrible, the immensity of his potential to have changed the world for the better makes the second tragedy even worse. In other words, the badness of his death stems mostly, perhaps overwhelmingly, from his unfulfilled potential rather than the direct, personal harms that he experienced. Or so the argument goes.

Longtermists would map these claims and conclusions on to humanity itself, as if humanity is an individual with its very own 'potential' to squander or fulfil, ruin or realise, over the course of 'its lifetime'. So, on the one hand, a catastrophe that reduces the human population to zero would be tragic because of all the suffering it would inflict upon those alive at the time. Imagine the horror of starving to death in subfreezing temperatures, under pitch-black skies at noon, for years or decades after a thermonuclear war. This is the first tragedy, a personal tragedy for those directly affected. But there is, longtermists would argue, a second tragedy that is astronomically worse than the first, arising from the fact that our extinction would permanently foreclose what could be an extremely long and prosperous future over the next, say, ~10^{100} years (at which point the 'heat death' will make life impossible). In doing this, it would irreversibly destroy the 'vast and glorious' longterm potential of humanity, in Ord's almost religious language – a 'potential' so huge, given the size of the Universe and the time left before reaching thermodynamic equilibrium, that the first tragedy would utterly pale in comparison.

This immediately suggests another parallel between individuals and humanity: death isn't the only way that someone's potential could be left unfulfilled. Imagine that Ramsey hadn't died young but, instead of studying, writing and publishing scholarly papers, he'd spent his days in the local bar playing pool and drinking. Same outcome, different failure mode. Applying this to humanity, longtermists would argue that there are failure modes that could leave our potential unfulfilled without us dying out, which I will return to below.

On this view, a climate catastrophe will be a small blip – like a 90-year-old who stubbed his toe when he was two

To summarise these ideas so far, humanity has a 'potential' of its own, one that transcends the potentials of each individual person, and failing to realise this potential would be extremely bad – indeed, as we will see, a moral catastrophe of literally cosmic proportions. This is the central dogma of longtermism: nothing matters more, ethically speaking, than fulfilling our potential as a species of 'Earth-originating intelligent life'. It matters so much that longtermists have even coined the scary-sounding term 'existential risk' for any possibility of our potential being destroyed, and 'existential catastrophe' for any event that actually destroys this potential.

Why do I think this ideology is so dangerous? The short answer is that elevating the fulfilment of humanity's supposed potential above all else could nontrivially increase the probability that actual people – those alive today and in the near future – suffer extreme harms, even death. Consider that, as I noted elsewhere, the longtermist ideology inclines its adherents to take an insouciant attitude towards climate change. Why? Because even if climate change causes island nations to disappear, triggers mass migrations and kills millions of people, it probably isn't going to compromise our longterm potential over the coming trillions of years. If one takes a cosmic view of the situation, even a climate catastrophe that cuts the human population by 75 per cent for the next two millennia will, in the grand scheme of things, be nothing more than a small blip – the equivalent of a 90-year-old man having stubbed his toe when he was two.

Bostrom's argument is that 'a non-existential disaster causing the breakdown of global civilisation is, from the perspective of humanity as a whole, a potentially recoverable setback.' It might be 'a giant massacre for man', he adds, but so long as humanity bounces back to fulfil its potential, it will ultimately register as little more than 'a small misstep for mankind'. Elsewhere, he writes that the worst natural disasters and devastating atrocities in history become almost imperceptible trivialities when seen from this grand perspective. Referring to the two world wars, AIDS and the Chernobyl nuclear accident, he declares that 'tragic as such events are to the people immediately affected, in the big picture of things … even the worst of these catastrophes are mere ripples on the surface of the great sea of life.'

This way of seeing the world, of assessing the badness of AIDS and the Holocaust, implies that future disasters of the same (non-existential) scope and intensity should also be categorised as 'mere ripples'. If they don't pose a direct existential risk, then we ought not to worry much about them, however tragic they might be to individuals. As Bostrom wrote in 2003, 'priority number one, two, three and four should … be to reduce existential risk.' He reiterated this several years later in arguing that we mustn't 'fritter … away' our finite resources on 'feel-good projects of suboptimal efficacy' such as alleviating global poverty and reducing animal suffering, since neither threatens our longterm potential, and our longterm potential is what really matters.

Ord echoes these views in arguing that, of all the problems facing humanity, our 'first great task … is to reach a place of safety – a place where existential risk' – as he defines it – 'is low and stays low', which he dubs 'existential security'. More than anything else, what matters is doing everything necessary to 'preserve' and 'protect' our potential by 'extracting ourselves from immediate danger' and devising robust 'safeguards that will defend humanity from dangers over the longterm future, so that it becomes impossible to fail.' Although Ord gives a nod to climate change, he also claims – based on a dubious methodology – that the chance of climate change causing

an existential catastrophe is only ~1 in 1,000, which is a whole two orders of magnitude lower than the probability of superintelligent machines destroying humanity this century, according to Ord.

What's really notable here is that the central concern isn't the effect of the climate catastrophe on actual people around the world (remember, in the grand scheme, this would be, in Bostrom's words, a 'small misstep for mankind') but the slim possibility that, as Ord puts it in *The Precipice*, this catastrophe 'poses a risk of an unrecoverable collapse of civilisation or even the complete extinction of humanity'. Again, the harms caused to actual people (especially those in the Global South) might be significant in absolute terms, but when compared to the 'vastness' and 'glory' of our longterm potential in the cosmos, they hardly even register.

Yet the implications of longtermism are far more worrisome. If our top four priorities are to avoid an existential catastrophe – ie, to fulfil 'our potential' – then what's not on the table for making this happen? Consider Thomas Nagel's comment about how the notion of what we might call the 'greater good' has been used to 'justify' certain atrocities (eg, during war). If the ends 'justify' the means, he argues, and the ends are thought to be sufficiently large (eg, national security), then this 'can be brought to bear to ease the consciences of those responsible for a certain number of charred babies'. Now imagine what might be 'justified' if the 'greater good' isn't national security but the cosmic potential of Earth-originating intelligent life over the coming trillions of years? During the Second World War, 40 million civilians perished, but compare this number to the 10^{54} or more people (in Bostrom's estimate) who could come to exist if we can avoid an existential catastrophe. What shouldn't we do to 'protect' and 'preserve' this potential? To ensure that these unborn people come to exist? What means can't be 'justified' by this cosmically significant moral end?

Bostrom himself argued that we should seriously consider establishing a global, invasive surveillance system that monitors every person on the planet in realtime, to amplify the 'capacities for preventive policing' (eg, to prevent omnicidal terrorist attacks that could devastate civilisation). Elsewhere, he's written that states should use preemptive violence/war to avoid existential catastrophes, and argued that saving billions of actual people is the moral equivalent of reducing existential risk by utterly minuscule amounts. In his words, even if there is 'a mere 1 per cent chance' of 10^{54} people existing in the future, then 'the expected value of reducing existential risk by a mere *one billionth of one billionth of one percentage point* is worth 100 billion times as much as a billion human lives.' Such fanaticism – a word that some longtermists embrace – has led a growing number of critics to worry about what might happen if political leaders in the real world were to take Bostrom's view seriously. To quote the mathematical statistician Olle Häggström, who – perplexingly – tends otherwise to speak favourably of longtermism:

I feel extremely uneasy about the prospect that [the calculations above] might become recognised among politicians and decision-makers as a guide to policy worth taking literally. It is simply too reminiscent of the old saying 'If you want to make an omelette, you must be willing to break a few eggs,' which has typically been used to explain that a bit of genocide or so might be a good thing, if it can contribute to the goal of creating a future utopia. Imagine a situation where the head of the CIA explains to the US president that they have credible evidence that somewhere in Germany, there is a lunatic who is working on a doomsday weapon and intends to use it to wipe out humanity, and that this lunatic has a one-in-a-million chance of succeeding. They have no further information on the identity or whereabouts of this lunatic. If the president has taken Bostrom's argument to heart, and if he knows how to do the arithmetic, he may conclude that it is worthwhile conducting a full-scale nuclear assault on Germany to kill every single person within its borders.

Here, then, are a few reasons I find longtermism to be profoundly dangerous. Yet there are additional, fundamental problems with this worldview that no one, to my knowledge, has previously noted in writing. For

example, there's a good case to make that the underlying commitments of longtermism are a major reason why humanity faces so many unprecedented risks to its survival in the first place. Longtermism might, in other words, be incompatible with the attainment of 'existential security', meaning that the only way to genuinely reduce the probability of extinction or collapse in the future might be to abandon the longtermist ideology entirely.

To Bostrom and Ord, failing to become posthuman would prevent us from realising our vast, glorious potential

To understand the argument, let's first unpack what longtermists mean by our 'longterm potential', an expression that I have so far used without defining. We can analyse this concept into three main components: transhumanism, space expansionism, and a moral view closely associated with what philosophers call 'total utilitarianism'.

The first refers to the idea that we should use advanced technologies to reengineer our bodies and brains to create a 'superior' race of radically enhanced posthumans (which, confusingly, longtermists place within the category of 'humanity'). Although Bostrom is perhaps the most prominent transhumanist today, longtermists have shied away from using the term 'transhumanism', probably because of its negative associations. Susan Levin, for example, points out that contemporary transhumanism has its roots in the Anglo-American eugenics movement, and transhumanists such as Julian Savulescu, who co-edited the book *Human Enhancement* (2009) with Bostrom, have literally argued for the consumption of 'morality-boosting' chemicals such as oxytocin to avoid an existential catastrophe (which he calls 'ultimate harm'). As Savulescu writes with a colleague, 'it is a matter of such urgency to improve humanity morally … that we should seek *whatever* means there are to effect this.' Such claims are not only controversial but for many quite disturbing, and hence longtermists have attempted to distance themselves from such ideas, while nonetheless championing the ideology.

Transhumanism claims that there are various 'posthuman modes of being' that are far better than our current human mode. We could, for instance, genetically alter ourselves to gain perfect control over our emotions, or access the internet via neural implants, or maybe even upload our minds to computer hardware to achieve 'digital immortality'. As Ord urges in *The Precipice*, think of how awesome it would be to perceive the world via echolocation, like bats and dolphins, or magnetoreception, like red foxes and homing pigeons. 'Such uncharted experiences,' Ord writes, 'exist in minds much less sophisticated than our own. What experiences, possibly of immense value, could be accessible, then, to minds much greater?' Bostrom's most fantastical exploration of these possibilities comes from his evocative 'Letter from Utopia' (2008), which depicts a techno-Utopian world full of superintelligent posthumans awash in so much 'pleasure' that, as the letter's fictional posthuman writes, 'we sprinkle it in our tea.'

The connection with longtermism is that, according to Bostrom and Ord, failing to become posthuman would seemingly prevent us from realising our vast and glorious potential, which would be existentially catastrophic. As Bostrom put it in 2012, 'the permanent foreclosure of any possibility of this kind of transformative change of human biological nature may itself constitute an existential catastrophe.' Similarly, Ord asserts that 'forever preserving humanity as it is now may also squander our legacy, relinquishing the greater part of our potential.'

The second component of our potential – space expansionism – refers to the idea that we must colonise as much of our future light cone as possible: that is, the region of spacetime that is theoretically accessible to us. According to longtermists, our future light cone contains a huge quantity of exploitable resources, which they refer to as our 'cosmic endowment' of negentropy (or reverse entropy). The Milky Way alone, Ord writes, is '150,000 light years across, encompassing more than 100 billion stars, most with their own planets.' Attaining humanity's longterm potential, he continues, 'requires only that [we] eventually travel to a nearby star and establish enough of a foothold to create a new flourishing society from which we could venture further.' By spreading 'just six light years

at a time', our posthuman descendants could make 'almost all the stars of our galaxy … reachable' since 'each star system, including our own, would need to settle just the few nearest stars [for] the entire galaxy [to] eventually fill with life.' The process could be exponential, resulting in ever-more 'flourishing' societies with each additional second our descendants hop from star to star.

But why exactly would we want to do this? What's so important about flooding the Universe with new posthuman civilisations? This leads to the third component: total utilitarianism, which I will refer to as 'utilitarianism' for short. Although some longtermists insist that they aren't utilitarians, we should right away note that this is mostly a smoke-and-mirrors act to deflect criticisms that longtermism – and, more generally, the effective altruism (EA) movement from which it emerged – is nothing more than utilitarianism repackaged. The fact is that the EA movement is deeply utilitarian, at least in practice, and indeed, before it decided upon a name, the movement's early members, including Ord, seriously considered calling it the 'effective utilitarian community'.

This being said, utilitarianism is an ethical theory that specifies our sole moral obligation as being to maximise the total amount of 'intrinsic value' in the world, as tallied up from a disembodied, impartial, cosmic vantage point called 'the point of view of the Universe'. From this view, it doesn't matter how value – which utilitarian hedonists equate with pleasure – is distributed among people across space and time. All that matters is the total net sum. For example, imagine that there are 1 trillion people who have lives of value '1', meaning that they are just barely worth living. This gives a total value of 1 trillion. Now consider an alternative universe in which 1 billion people have lives with a value of '999', meaning that their lives are extremely good. This gives a total value of 999 billion. Since 999 billion is less than 1 trillion, the first world full of lives hardly worth living would be morally better than the second world, and hence, if a utilitarian were forced to choose between these, she would pick the former. (This is called the 'repugnant conclusion', which longtermists such as Ord, MacAskill and Greaves recently argued shouldn't be taken very seriously. For them, the first world really might be better!)

Beckstead argued that we should prioritise the lives of people in rich countries over those in poor countries

The underlying reasoning here is based on the idea that people – you and I – are nothing more than means to an end. We don't matter in ourselves; we have no inherent value of our own. Instead, people are understood as the 'containers' of value, and hence we matter only insofar as we 'contain' value, and therefore contribute to the overall net amount of value in the Universe between the Big Bang and the heat death. Since utilitarianism tells us to maximise value, it follows that the more people (value containers) who exist with net-positive amounts of value (pleasure), the better the Universe will become, morally speaking. In a phrase: people exist for the sake of maximising value, rather than value existing for the sake of benefitting people.

This is why longtermists are obsessed with calculating how many people could exist in the future if we were to colonise space and create vast computer simulations around stars in which unfathomably huge numbers of people live net-positive lives in virtual-reality environments. I already mentioned Bostrom's estimate of 10^{54} future people, which includes many of these 'digital people', but in his bestseller *Superintelligence* (2014) he puts the number even higher at 10^{58} people, nearly all of whom would 'live rich and happy lives while interacting with one another in virtual environments'. Greaves and MacAskill are similarly excited about this possibility, estimating that some 10^{45} conscious beings in computer simulations could exist within the Milky Way alone.

That is what our 'vast and glorious' potential consists of: massive numbers of technologically enhanced digital posthumans inside huge computer simulations spread throughout our future light cone. It is for this goal that, in Häggström's scenario, a longtermist politician would annihilate Germany. It is for this goal that we must not 'fritter … away' our resources on such things as solving global poverty. It is for this goal that we should consider implementing a global surveillance system, keep pre-emptive war on the table, and focus more on superintelligent

machines than saving people in the Global South from the devastating effects of climate change (mostly caused by the Global North). In fact, Beckstead has even argued that, for the sake of attaining this goal, we should actually prioritise the lives of people in rich countries over those in poor countries, since influencing the long-term future is of 'overwhelming importance', and the former are more likely to influence the long-term future than the latter. To quote a passage from Beckstead's 2013 PhD dissertation, which Ord enthusiastically praises as one of the most important contributions to the longtermist literature:

Saving lives in poor countries may have significantly smaller ripple effects than saving and improving lives in rich countries. Why? Richer countries have substantially more innovation, and their workers are much more economically productive. [Consequently,] it now seems more plausible to me that saving a life in a rich country is substantially more important than saving a life in a poor country, other things being equal.

This is just the tip of the iceberg. Consider the implications of this conception of 'our potential' for the development of technology and creation of new risks. Since realising our potential is the ultimate moral goal for humanity, and since our descendants cannot become posthuman, colonise space and create ~10^{58} people in computer simulations without technologies far more advanced than those around today, failing to develop more technology would itself constitute an existential catastrophe – a failure mode (comparable to Ramsey neglecting his talents by spending his days playing pool and drinking) that Bostrom calls 'plateauing'. Indeed, Bostrom places this idea front-and-centre in his canonical definition of 'existential risk', which denotes any future event that would prevent humanity from reaching and/or sustaining a state of 'technological maturity', meaning 'the attainment of capabilities affording a level of economic productivity and control over nature close to the maximum that could feasibly be achieved.' Technological maturity is the linchpin here because controlling nature and increasing economic productivity to the absolute physical limits are ostensibly necessary for creating the maximum quantity of 'value' within our future light cone.

But reflect for a moment on how humanity got itself into the current climatic and ecological crisis. Behind the extraction and burning of fossil fuels, decimation of ecosystems and extermination of species has been the notion that nature is something to be controlled, subjugated, exploited, vanquished, plundered, transformed, reconfigured and manipulated. As the technology theorist Langdon Winner writes in *Autonomous Technology* (1977), since the time of Francis Bacon our view of technology has been 'inextricably bound to a single conception of the manner in which power is used – the style of absolute mastery, the despotic, one-way control of the master over the slave.' He adds:

There are seldom any reservations about man's rightful role in conquering, vanquishing, and subjugating everything natural. This is his power and his glory. What would in other situations seem [to be] rather tawdry and despicable intentions are here the most honourable of virtues. Nature is the universal prey, to manipulate as humans see fit.

This is precisely what we find in Bostrom's account of existential risks and its associated normative futurology: nature, the entire Universe, our 'cosmic endowment' is there for the plundering, to be manipulated, transformed and converted into 'value-structures, such as sentient beings living worthwhile lives' in vast computer simulations, quoting Bostrom's essay 'Astronomical Waste' (2003). Yet this Baconian, capitalist view is one of the most fundamental root causes of the unprecedented environmental crisis that now threatens to destroy large regions of the biosphere, Indigenous communities around the world, and perhaps even Western technological civilisation itself. While other longtermists have not been as explicit as Bostrom, there is a clear tendency to see the natural world the way utilitarianism sees people: as means to some abstract, impersonal end, and nothing more. MacAskill and a colleague, for example, write that the EA movement, and by implication longtermism, is

'tentatively *welfarist* in that its tentative aim in doing good concerns promoting wellbeing only and not, say, protecting biodiversity or conserving natural beauty for their own sakes.'

On this account, every problem arises from too little rather than too much technology

Just as worrisome is the longtermist demand that we must create ever-more powerful technologies, despite the agreed-upon fact that the overwhelming source of risk to human existence these days comes from these very technologies. In Ord's words, 'without serious efforts to protect humanity, there is strong reason to believe the risk will be higher this century, and increasing with each century that technological progress continues.' Similarly, in 2012 Bostrom acknowledges that

the great bulk of existential risk in the foreseeable future consists of anthropogenic existential risks – that is, arising from human activity. In particular, most of the biggest existential risks seem to be linked to potential future technological breakthroughs that may radically expand our ability to manipulate the external world or our own biology. As our powers expand, so will the scale of their potential consequences – intended and unintended, positive and negative.

On this view, there is only one way forward – more technological development – even if this is the most dangerous path into the future. But how much sense does this make? Surely if we want to maximise our chances of survival, we should oppose the development of dangerous new dual-use technologies. If more technology equals greater risk – as history clearly shows and technological projections affirm – then perhaps the only way to actually attain a state of 'existential security' is to slow down or completely halt further technological innovation.

But longtermists have an answer to this conundrum: the so-called 'value-neutrality thesis'. This states that technology is a morally neutral object, ie, 'just a tool'. The idea is most famously encapsulated in the NRA's slogan 'Guns don't kill people, people kill people,' which conveys the message that the consequences of technology, whether good or bad, beneficial or harmful, are entirely determined by the users, not the artefacts. As Bostrom put it in 2002, 'we should not *blame* civilisation or technology for imposing big existential risks,' adding that 'because of the way we have defined existential risks, a failure to develop technological civilisation would imply that we had fallen victims of an existential disaster.'

Ord similarly argues that 'the problem is not so much an excess of technology as a lack of wisdom,' before going on to quote Carl Sagan's book *Pale Blue Dot* (1994): 'Many of the dangers we face indeed arise from science and technology but, more fundamentally, because we have become powerful without becoming commensurately wise.' In other words, it is our fault for not being smarter, wiser and more ethical, a cluster of deficiencies that many longtermists believe, in a bit of twisted logic, could be rectified by technologically reengineering our cognitive systems and moral dispositions. Everything, on this account, is an engineering problem, and hence every problem arises from too little rather than too much technology.

We can now begin to see how longtermism might be self-defeating. Not only could its 'fanatical' emphasis on fulfilling our longterm potential lead people to, eg, neglect non-existential climate change, prioritise the rich over the poor and perhaps even 'justify' pre-emptive violence and atrocities for the 'greater cosmic good' but it also contains within it the very tendencies – Baconianism, capitalism and value-neutrality – that have driven humanity inches away from the precipice of destruction. Longtermism tells us to maximise economic productivity, our control over nature, our presence in the Universe, the number of (simulated) people who exist in the future, the total amount of impersonal 'value' and so on. But to maximise, we must develop increasingly powerful – and dangerous – technologies; failing to do this would itself be an existential catastrophe. Not to worry, though, because technology is not responsible for our worsening predicament, and hence the fact that most risks stem

directly from technology is no reason to stop creating more technology. Rather, the problem lies with us, which means only that we must create even more technology to transform ourselves into cognitively and morally enhanced posthumans.

This looks like a recipe for disaster. Creating a new race of 'wise and responsible' posthumans is implausible and, if advanced technologies continue to be developed at the current rate, a global-scale catastrophe is almost certainly a matter of when rather than if. Yes, we will need advanced technologies if we wish to escape Earth before it's sterilised by the Sun in a billion years or so. But the crucial fact that longtermists miss is that *technology is far more likely to cause our extinction before this distant future event than to save us from it*. If you, like me, value the continued survival and flourishing of humanity, you should care about the long term but reject the ideology of longtermism, which is not only dangerous and flawed but might be contributing to, and reinforcing, the risks that now threaten every person on the planet.

19 October 2021