IF NOW ISN'T THE MOST INFLUENTIAL TIME EVER, WHEN IS?

The Precipice: Existential Risk and the Future of Humanity (Bloomsbury, 2020)
by Toby Ord
reviewed by Kritika Maheshwari

According to Toby Ord's *The Precipice,* humanity is living through unprecedented times. There is a good chance that you've heard these exact words being (over) used in recent months. Many think that they perfectly describe our current predicament of grappling with this century's first global pandemic. But for Ord, these words echo the long story of humanity's transition into times of peril as we continue to gain more and more power over shaping the future of our entire existence. And the reality of the ongoing pandemic is just one example of that.

To understand why, consider the following facts. By now, the Covid-19 pandemic has reached all inhabited parts of the world, and the death toll has escalated to over a million and counting. Our societies are struggling to cope with the immediate consequences of this crisis. Lockdowns have brought personal and social life to a standstill, health care systems are crumbling under pressure, and tensions about potential global economic crises are on the rise. Simultaneously, we are also faced with the local and global impacts of climate change, exposed to the constant threat of totalitarian regimes taking over or a nuclear conflict breaking out, and are also vulnerable to several known and unknown risks posed by rapid and radical technological developments. In theory, it's plausible that there are mechanisms by which the cumulative effects of these risks *could* have far-reaching consequences that are bringing humanity closer to what Ord refers to as "a crumbling ledge on the brink of a precipice".

Ord's book, *The Precipice,* is a timely reminder that such mechanisms do exist and, in fact, they are up and running. The long history of humanity, which is marked by transitions from agricultural to scientific to industrial revolutions, has now brought us to a precipitous moment in which we are faced with a number of existential risks – risks that threaten the destruction of humanity's long-term *potential* (more on this below) by, for example, causing the extinction of humanity or triggering a global collapse of civilization from which we cannot recover. Even if a global pandemic, by itself, may fail to extinguish humanity's long-term future, our failure to avoid or prevent actions that may aid and abet the spread of natural or human-engineered pandemics, combined with our failure to mitigate the above-mentioned risks can together bring about a permanent collapse of civilisation, if

not complete *extinction*. According to Ord, if humanity doesn't get its act together, all it will take is a little push for us to fall over the edge of the precipice. And if we fall, it's game over.

To fathom the impact of existential risks *if* they materialised, consider the following grim picture that Derek Parfit paints for us in his now classic book *Reasons and Persons* (1984). Imagine that a nuclear war breaks out and kills 99 percent of the world's population and leaves the remaining civilisation in a dark age that could last centuries. Contrast this scenario with a second one wherein a nuclear war kills a full 100 percent of the world's population. Intuitively, the second war seems far worse than the first – not just quantitatively but also qualitatively. Although both wars kill billions and billions of people and destroy our present, the second war wipes out humanity entirely and therefore destroys our future. Existential risks, if left ignored, hold out the promise of ensuring that humanity ends up in the second scenario.

***

The idea of humanity going extinct or simply getting locked in a bleak dystopian scenario from which it fails to recover is deeply frightening. Intuitively, it seems that it would be terrible if we allowed this to happen one way or the other, either knowingly or unknowingly. But *why* exactly would it be bad? Some philosophers think that extinction would be bad because it would cause people to suffer painful deaths and psychological harms. But imagine an extinction scenario in which humanity goes out of existence quietly and painlessly. Is there still something wrong about causing or allowing our extinction this way? We need to explore other reasons. Ord offers us a couple, ranging from appeals to the instrumental value of humanity, to our cosmic significance stemming from our unique capabilities and to what he calls humanity's "civilisational virtues". But Ord's chief reason is grounded in the value of humanity's future, or, in other words, humanity's long-term *potential*. As he puts it:

> Human extinction would foreclose our future. It would destroy our potential. It would eliminate all possibilities but one: a world bereft of human flourishing. Extinction would bring about this failed world and lock it in forever – there would be no coming back.

But what actually *is* humanity's potential? For Ord, our potential lies in what humanity can achieve through the combined actions of each and every human, and perhaps even those trans-humans who may one day replace us. The scope of this potential knows no bounds. Ord devotes the final chapter of his book to spelling out why this is so. Think of the trillions of human lives that can still be lived, the vast amount of knowledge we are yet to produce, the technological heights we are yet to reach, the scientific breakthroughs that are yet to happen, and the many intergalactic space missions we are yet to undertake. Accordingly, what makes the materialisation of existential risks *uniquely* morally significant is the permanent loss of the vast range of possible futures that remain open to us.

WHAT MAKES THE MATERIALISATION OF EXISTENTIAL RISKS UNIQUELY MORALLY SIGNIFICANT IS THE PERMANENT LOSS OF THE VAST RANGE OF POSSIBLE FUTURES THAT REMAIN OPEN TO US

If we follow Ord's understanding of humanity's long-term potential as being defined by the set of *all* possible futures that remain open to us, then, presumably, some of the possible futures in this set also include ones that involve many years of mass suffering and pain, genocides, wars, and all other kinds of imaginable futures which seem objectively bad. So if it is the loss of humanity's *expansive* potential, constituted by *both* prosperous futures as well as dystopian ones, that grounds the unique badness of humanity's extinction, then we end up with the odd claim that extinction is bad not only because of loss of futures with astonishing value, but also ones with astonishing disvalue. However, we can easily fix this in at least two ways:

1) By restricting our definition of humanity's potential to include only all *good* possible

futures, however we choose to understand what "good" futures are.
2) By reframing Ord's claim as follows: that the badness of extinction risk materialising is grounded in the loss of humanity's *particular kind* of long-term future, one that involves fulfilling our *positive* potential.

But why should we think in terms of humanity's *potential* in the first place? Ord asks us to think of an analogy with a human's life, an idea that is similar to what George Kavka proposes in his 1978 essay "The Futurity Problem", in which he highlights the parallels between the narrative structure of our species' history and that of an individual life. The structure of an ordinary human life constitutes different stages, including childhood, youth, adolescence and finally old age. Our childhood and youth mark the beginning of our life during which we learn and grow; and then we slowly progress towards adulthood during which we fulfil our potential and accomplish various personal goals before finally reaching the resting stage of old age.

According to Ord, humanity's life story has a similar structure. From the perspective of a geological timescale, Ord claims that humanity is still very much in its infancy. As he explains:

> On the timescale of an individual human life, our 200,000-year history seems almost incomprehensibly long. But on a geological timescale, it is short, and vanishingly so on the timescale of the universe as a whole. Our cosmos has a 14-billion-year history and even that is short on the grandest scales. Trillions of years lie ahead of us.

Just as many seem to find the premature death of an individual horrific, cutting short humanity's future before it can achieve its long-term potential strikes Ord, and many others, as comparably bad, or perhaps even worse.

Drawing a comparison between humanity's life narrative and the narrative of an ordinary human life might seem appealing at first, but not everyone finds this convincing. Philosophers like James Lenman doubt whether such an analogy holds any water. Lenman's position on the matter is roughly as follows:

1) He doubts whether there is some kind of grand philosophical vision of human history that is goal-oriented in a way that it would be tragic if we fail to attain this goal
2) He thinks that the tragedy of an individual's premature death has no obvious analogue in the career of our species as a whole. For if humanity exists for another million year, then *today* would seem to be humanity's childhood. But if humanity ceases to exist *tomorrow*, then today would seem to be humanity's old age.

Ord does not give us an argument for why exactly humanity's life-story mirrors that of a human life, nor does he address Lenman's claims directly, but I imagine that he could respond as follows. To Lenman's first point, Ord may reiterate that humanity's life-story *does* have a goal, and that goal is to achieve its full potential. To Lenman's second point, Ord may contend that because humanity has this goal, and because on a geological time-scale we are still very young, extinction of humanity tomorrow would be akin to the premature death of an infant whose life was once full of potential.

But even if we manage to deflect the objection that the analogy doesn't work, one may still question how watertight this goal-oriented talk is. Thinking of the value of mitigating existential risks from the perspective of achieving humanity's goal (reaching its *full* potential) seems to be motivated by the idea that humanity has an "end-stage" such that there will be a time when that end-stage will be reached. But unless Ord is assuming that humanity's extinction is imminent, as many moral philosophers do, and that it's an end-stage we cannot avoid, how else are we to make sense of what happens once we reach this point? Can we *then* say that humanity's extinction might perhaps not be morally bad, even though it might be regrettable? More importantly, how and when exactly do we establish that humanity has indeed fulfilled this potential? Or how do we establish that we haven't already done so?

***

We needn't settle on any definitive answers for now, as the idea that we must protect humanity's long-term future from existential risks can still get off the ground without establishing whether or not

humanity's story has a narrative structure of the sort that Ord posits. What I find more intriguing is a different suggestion that Ord offers us for making sense of humanity's potential: thinking of humanity as a group agent in itself. Imagine what humanity's future could and should be like, and what decisions it would make, were it sufficiently rational and wise. Just as individual agents have potentials that they can choose to realise as they progress further in life, we can think of humanity as a group agent with certain character traits and dispositions that can realise its potential as it progresses further in time. With this line of thought, Ord sets the tone of the book's discussion from the viewpoint of "humanity's morality", rather than morality from an individual's perspective. As is well known, ethics is most commonly discussed from an individual's viewpoint: what should *I* do? And only sometimes do we consider questions from a group's viewpoint: what should a *group* do? But Ord suggests that we take a step further, and think about what matters most from the perspective of *humanity*: what is in its best interest, and so on.

THE IDEA OF CONCEIVING HUMANITY AS A GROUP AGENT IS AN INTERESTING ONE, BUT ALSO ONE THAT WARRANTS CAUTION

The idea of conceiving humanity as a group agent is an interesting one, but also one that warrants caution. It is not uncommon to think of large collectives like nations, organizations, or corporations as group agents who can be subject to demands of morality. However, it is not clear whether, and how, we can extend our ideas of what conditions suffice for group agency to humanity as well. Humanity – in an abstract and intuitive sense of the word – is a fragmented and disorganized body, spanning across times and distances that makes it hard to pin it down as *one* coherent agent. Moreover, in order to ascribe agency, humanity *at least* needs to exhibit a form of cohesion that resembles group behaviour or mimics decision-making capacities *as a group*, and not just a mere collection of individuals. We can say at least this much without necessarily committing to any particular account of group moral agency, for group-level rational decision-making procedure is commonly accepted as a hallmark of group agency across the board. And humanity as a whole, as of now, seems to lack one.

So perhaps what we should really be asking is whether and how humanity's *potential* to become a group agent can be fulfilled, or what kinds of steps or actions humanity can take to *become* a group agent. These questions remain unanswered for now, but to Ord's credit his discussion opens up a whole range of interesting ideas to consider. For instance, *if* we can ascribe group agency to humanity, then can it *itself* be held accountable if it fails to fulfil its own potential? Is it appropriate to speak of *humanity's* having certain beliefs, desires, or intentions? Would humanity be a different *kind* of group agent if it included new kinds of moral agents, say AI robots, that may live millions years from now? And if so, what would this difference in kind amount to?

These, then, are some of the key philosophical aspects of Ord's discussion of existential risks. Moving onto the risks themselves, what are some of the most pressing existential risks that threaten humanity's potential, and, more importantly, precisely *how* urgent are they and why?

***

In the second part of *The Precipice*, Ord examines the scientific underpinnings of different sources of existential risks, ranging from natural risks (such as those posed by asteroid collision and stellar explosions) to anthropogenic risks (such as climate change and nuclear wars), as well as novel future risks arising from radical technological developments in nanoscience and space exploration that we might face in centuries to come. Much of his discussion aims to clarify whether, why, and to what extent these and other sources of risks do in fact pose a real threat of humanity's *extinction*.

According to Ord, the total natural risks that arise from the threat of collision with asteroids that are bigger than ten kilometres in size, or similarly sized comets, as well as super-volcanic and stellar collisions, taken

altogether, are estimated to carry merely a 1 in 10,000 chance of extinction in next 100 years. These estimates are based on studying fossil records of how long species similar to us survived, and thus extrapolating to the total extinction risk they faced. Ord deems the extinction risk from natural causes to be significant, yet fairly low in comparison to those risks posed by anthropogenic causes. But before turning to those, I want to briefly turn our attention to the following question: how can we tell if a natural phenomenon poses a *real* risk of our extinction or permanent collapse?

HOW CAN WE TELL IF A NATURAL PHENOMENON POSES A REAL RISK OF OUR EXTINCTION OR PERMANENT COLLAPSE?

To illustrate what I mean, consider the following. It is clear why we would think that collision with an asteroid that is bigger than ten kilometres in size poses a real risk of extinction. The answer is simple: because in the last extinction event, as Ord highlights, "*all* land-based vertebrates weighing more than five kilograms were killed". This counts as clear evidence that collisions with large asteroids *do* in fact pose a real risk of extinction: they have led to extinction in the past and might do so again in the future. But then consider what Ord has to say about the risk posed by the natural phenomenon of magnetic field shifting:

> "[T]he Earth's entire magnetic field can shift dramatically, and sometimes reverses its direction entirely. These shifts leave us more exposed to cosmic rays during the time it takes to reorient. However, this happens often enough that we can tell it isn't an extinction risk (it has happened about 20 times in the 5 million years since humans and chimpanzees diverged). And since the only well-studied effects appear to be somewhat increased cancer rates, it is not a risk of civilization collapse either.

One may wonder whether an event happening 20 times in 5 million years is infrequent enough to disregard the claim that it poses an extinction risk. But the frequency of a natural event cannot be the *only* factor that determines whether it poses an extinction risk. And simply referring to studies that trace the impact of this natural event on increased cancer rates does not negate the possibility that such events may still pose a risk of extinction, or risk of some other kind.

My motivation behind raising these points is not to offer a definitive answer, but merely to point to the idea that it remains unclear how we should filter extinction-triggering natural risks from the others. This is important for our aims of defining the boundaries of our natural existential risk landscape. The boundaries of our anthropogenic risk landscape, by contrast, are clearly very wide and stretchable. They seem to keep extending as humanity keeps its engine of technological innovation running. Amongst different sources of anthropogenic risks, Ord estimates the risk from engineered pandemics and what he calls "unaligned" artificial intelligence to be the most pressing risk of our times, estimating them to pose respectively at least a 1 in 30 and a 1 in 10 chance of destroying humanity's potential this century.

Surprisingly, risks independently posed by climate change, natural pandemics, and nuclear wars that are commonly touted as harbingers of humanity's ultimate end – in both fiction and non-fiction – turn out to be ones that are unlikely to push humanity to the edge of extinction, according to Ord's estimations. Consider what he says about risks from nuclear wars: If a full-scale nuclear war did break out, what would happen? Would it really threaten extinction or the permanent collapse of civilization? According to Ord, the threat to humanity would come from the global effects of nuclear war, rather than local effects. One such global effect involves covering of entire surface of the Earth with radioactive dust from nuclear weapons. However, in practice, Ord points out that this scenario would require ten times as many weapons as there are in humanity's possession as of now. We can, he assures us, dismiss the claim that humanity has enough nuclear weapons to destroy itself as hyperbolic and untrue.

Similarly, he contends that while climate change risks posed by extreme levels of warming are considered serious enough to cause "a global calamity of unprecedented scale", it does not pose a direct existential

risk to humanity. Notwithstanding the fact that other risks associated with climate change, such as risk of mass migration, environmental damage or mass starvation are extremely worrisome and problematic, Ord feels that the mechanisms through which climate change can cause our direct extinction or irrevocable damage, such as heat stress or runaway greenhouse effects, are unlikely to actualize in reality. Nevertheless, this is not to say that these risks are not important. They play a very significant role as risk factors, that is, in increasing the overall risk of global catastrophes, and leaving us more vulnerable to existential risks in the future.

***

The more urgent and plausible existential risk (1 in 10 this century), according to Ord, comes from expected growth in the development of artificial general intelligence (AGI). Ord notes that since its early days, our goals of developing AI abilities have become grander: from merely recognizing cats to developing agents with an artificial general intelligence that could potentially surpass our own abilities in almost every domain. Technical advances in deep learning, such as improved datasets, computing power, design and training of neural networks, have allowed AI researchers to achieve the former goal: neural networks can not only recognize cats and differentiate between different breeds, they can also outperform humans at games likes chess and Go.

The latter goal of creating an artificial general intelligence that could potentially surpass our own remains a mere possibility. This possibility, however, is one that warrants our serious attention. Ord offers a number of reasons for this. For instance, sufficiently intelligent AGI systems could reach a stage where its behaviour is unpredictable and uncontrollable; it could

Will our future look like this?

Or will it look like this?

gain control of the Internet or escalate its power to the extent of improving its own intelligence levels far beyond our own; it could even acquire an instrumental goal of survival and succeed in blocking any of our attempts to shut it down. Although no current AI system can improve its own intelligence, develop new weapons technology, take over control or cripple the rest of humanity, if any of this were to become a reality, one immediate threat would be that humanity will lose all control of its own future.

It's not clear exactly *how* an AGI system might seize control, but what is clear to Ord is that *our* potential loss of control over humanity's long-term future poses a real existential risk in the following way: "In the act of creation, we would cede our status as the most intelligent entities on Earth. So without a very good plan to keep control, we should also expect to cede our status as the most powerful species, and the one that controls its own destiny". In the scenario where we are succeeded by AGI, our future could potentially be at the mercy of a system that may or may not align its moral values with ours, that may allow for a decent outcome if we are lucky, or may leave us locked in a dystopian future forever.

Ord admits that the case for existential risks resulting in our extinction or some existential catastrophe from AGI is rather speculative. But he also thinks that it's reasonable that a speculative case for risks involving losing control over humanity's future should raise our threshold of concern more than a robust case for extremely low probability existential risks that we face from, say, asteroid collisions. But one may ask: how reasonable is this concern?

In a recent review of Brian Smith's *The Promise of Artificial Intelligence*, Tim Crane discusses how despite

AI's recent successes there is still very little reason to believe that it is ever likely to create genuine thinking machines – let alone machines that can one day get to the point of taking over control of our humanity. According to Crane, grand claims about the possibility of AGI – and everything that they might be able to achieve – ignore some significant differences between relatively well-defined domains where AI has historically succeeded and poorly-defined domains such as "general intelligence". For instance, the former domains are characterised by goal-oriented tasks. Consider how machines are trained to play a game of Go or recognise speech. Both tasks come with their own set of clear, pre-defined rules and a goal that the machine tries to achieve. In the case of Go, the goal is to win the game and in the case of speech recognition, the goal is to recognise spoken words.

It is not, however, remotely clear what the goal or target of "general intelligence" might be. How can we characterise, in abstract terms, the problem(s) that general intelligence tries to solve? Consider, for instance, the activity of conversing with someone. This could have a number of goals, from wanting to ask for information, to asking for help, to expressing one's emotion, to ask for directions, and so on. But what is the overall goal of a conversation? According to Crane, there is no single goal. Conversation, as an activity, simply lacks the feature of having an easily expressible goal to which we direct our intelligence. The difficulty in defining a task or goal-oriented domain of general intelligence appears to undermine the plausibility of the idea that current successes in AI imply that AGI is a real possibility in the future.

HOW CAN WE CHARACTERISE, IN ABSTRACT TERMS, THE PROBLEMS THAT GENERAL INTELLIGENCE TRIES TO SOLVE?

Crane's scepticism towards the possibility of genuine AGI may be warranted, but let's assume for the sake of argument that there *is* a real possibility that humanity will one day succeed in creating AGI and consequently cede its status as the most intelligent entity. Even then, Ord's worry that humanity may lose control of its future seems to be in tension with his otherwise very tolerant and inclusive view of expanding our moral community and welcoming "new kinds of moral agents in the future" within the category of humanity, if this is what it takes to further our aims of achieving our full potential. If we entertain the possibility that humanity as we know it now could be replaced by some new moral agents in the form of, say, powerful, super-intelligent AGI systems, who may potentially be far more capable of and motivated towards achieving the best possible future for humanity, then should we be willing to at least entertain this as a good possibility for furthering our interest? I'm not sure how much credence Ord would assign to this possibility, but it certainly raises a number of interesting questions. For instance: is our philosophical notion of humanity malleable enough to subsume AGIs as our descendants or flag-bearers or co-partners in the endeavour of reaching our full potential? Needless to say, the kinds of worries that Ord raises about AGI locking our future in a permanent Orwell-style enforced totalitarian regime (or worse) may easily tip the balance against it.

***

Taking stock of these major existential risks, Ord offers his own best estimates of the *total* existential risks befalling this century. Combining his best estimates of each existential risk (and assuming some positive correlation between them) that reflects his overall impression of the risk landscape in light of scientific evidence, whilst also factoring in uncertainty and imprecision about these numbers, Ord places the overall existential risk in the next 100 years to be: *1 in 6*.

To get an intuitive sense of how low or high this probability is, compare it with Ord's estimate of extinction or unrecoverable collapse of civilisation befalling humanity in the twentieth century: 1 in 100. By comparison, a 1 in 6 chance seems to be a very high level of risk, as Ord acknowledges: "[T]his is not a small statistical probability that we must diligently bear in mind, like the chance of dying in a car crash, but something that could readily occur, like the roll of a die, or Russian roulette".

Ord's comparison of the total existential risk befalling us in the next 100 years to playing Russian Roulette with humanity's future is an interesting one. It is certainly eye-catching and makes for good newspaper headlines. On the one hand, it could be seen as a useful heuristic to convey the urgency of prioritizing existential risk, but on the other hand it also seems one that easily lends itself to misinterpretation, for it may lead people to under- or over-exaggerate the chance of existential risk this century.

For instance, in the philosophy of risk (a discipline that has been growing since the 1970s), it is commonplace to distinguish between three conceptions of probabilities: 1) fact-relative, 2) belief-relative and 3) evidence-relative. A fact-relative conception considers the risk to be facts about the world and is objective insofar it is independent of an individual's belief or evidence. In this sense, a 1 in 6 chance of dying in a game of Russian Roulette is a fact-relative risk. Belief- and evidence-relative conceptions, by contrast, consider the risk to be a measure of the strength of an agent's belief or the weight of an agent's evidence. Ord's estimate of total existential risk falls in this latter category.

Accordingly, the comparison between Ord's evidence-relative probability of going extinct this century with fact-relative probability of dying in a game of Russian Roulette may give the wrong impression that the comparison is between two sets of probabilities that are both objective, precise and accurate – something Ord himself warns us against: "[D]on't take these numbers to be completely objective". Whether or not we can even draw any sensible or meaningful comparison between these two conceptions of risk is a difficult question in itself. But for now, it suffices to say that readers should treat the comparison of total existential risk with Russian Roulette to be a metaphorical and not a literal one.

We can take the metaphor to suggest that humanity's each and every action (or lack thereof) that increases the total risk is akin to someone pulling the trigger on us and initiating a causal chain that will hit its target one in six times. This follows from our standard way of thinking about Russian Roulette cases. With that said, one might wonder how exactly this comparison is supposed to help us in thinking about mitigating, reducing or eliminating different existential risks from the risk landscape? Suppose humanity decides to reduce the risk of extinction from developments in unaligned AI by pulling the plug on the project altogether. Would this count as taking out one, two or three bullets out of our Existential Russian Roulette? Or is it more? How can we tell?

***

In the last part of his book, Ord moves towards suggesting how humanity should move forward. He argues that we have a responsibility for ensuring that the bullet doesn't fire. Currently, we spend less than a thousandth of a percent of gross world product (GWP) on targeted existential risk interventions. Moreover, as has become clear from the current (mis)handling of the pandemic, we also lack efficient international coordination, as well as health and political institutions that are capable of tackling not just short-term risks, but also longer-term existential risks. The challenge for us, then, is to prioritise the most important existential risks within the global political agenda in order to develop adequate strategies for protecting and preserving our long-term future. As most existential risks that Ord describes arise from human activity, it is entirely within our control how we choose to govern, regulate, and control our activities to prevent these risks from materialising or worsening. As Ord summarises the situation, the choices we are going to make will determine "whether we live or die; fulfil our potential or squander our chance at greatness".

CURRENTLY, WE SPEND LESS THAN A THOUSANDTH OF A PERCENT OF GROSS WORLD PRODUCT ON TARGETED EXISTENTIAL RISK INTERVENTIONS

So how, then, can we protect our potential? Ord offers us the following suggestions for safeguarding humanity. First, we need to reach a stage of *existential security* by bringing down existential risks to a sustainable level. According to Ord, "existential security is about reducing

the total existential risk by as many percentage points as possible". So if we manage to reduce the risk of future natural and engineered pandemics substantially as well as other risks that are within our control (although it remains unclear exactly how did could be done), we can soon be "past the stage of standing on a precipice, free to contemplate the range of futures that lie open before us".

IF WE STRIVE TO ACHIEVE EXISTENTIAL SECURITY, WE CAN IMPROVE THE EXPECTED QUALITY OF LIFE OF THOSE WHO WOULD COME AFTER US

This brings us to the second stage of Ord's strategy, namely, *long reflection*. This stage is marked by humanity having reached the stage of existential security, and finally setting itself the task of finding the "final answer to the question of which is the best kind of future for humanity". Ord imagines this idealistic (and rather romantic) period of long reflection to be one in which discussions about which path humanity should set itself onto will occupy the intellectual and public forums. Ultimately, the aim of this phase would be to steer ourselves in the direction of achieving humanity's *full potential*, the final stage in Ord's strategy.

From the point of view of our current world, one that is heavily divided and marked by sharp social, political, and cultural disagreements and conflicts, Ord's stage of long reflection sounds like an unimaginable utopia, and raises questions like: How are we to strike a balance between pursuing this task of protecting our long-term potential and the task of protecting humanity from living in inhospitable conditions characterised by hunger, mass migration and shortage of resources? How much of our political, economic, and personal attention should go into mitigating the proposed 1 in 10 chance of existential catastrophe posed by AGI takeovers in the next hundred years rather than making sure humanity thrives in bearable, decent living conditions by mitigating risks from starvation, poverty, and climate change in the next fifty years?

Perhaps Ord would say that we can pursue these tasks together. If we strive to achieve existential security, we can improve the expected quality of life of those who would come after us. We currently have the advantage of being at a very early stage of thinking about the long-term future of humanity. We also have the power to ensure that humanity can achieve a desirable future. And it is now up to us whether or not we choose to exercise this power correctly, and, most importantly, at the right time. As Ord notes, "through our choices we can pull back from the precipice and, in time, create a future of astonishing value – with richness of which we can barely dream, made possible by innovations we are yet to conceive".

Kritika Maheshwari is a PhD candidate at the Department of Ethics, Social and Political Philosophy at University of Groningen, The Netherlands. Her research interests include normative and meta-ethics, philosophy of risk and uncertainty, metaphysics and epistemology.