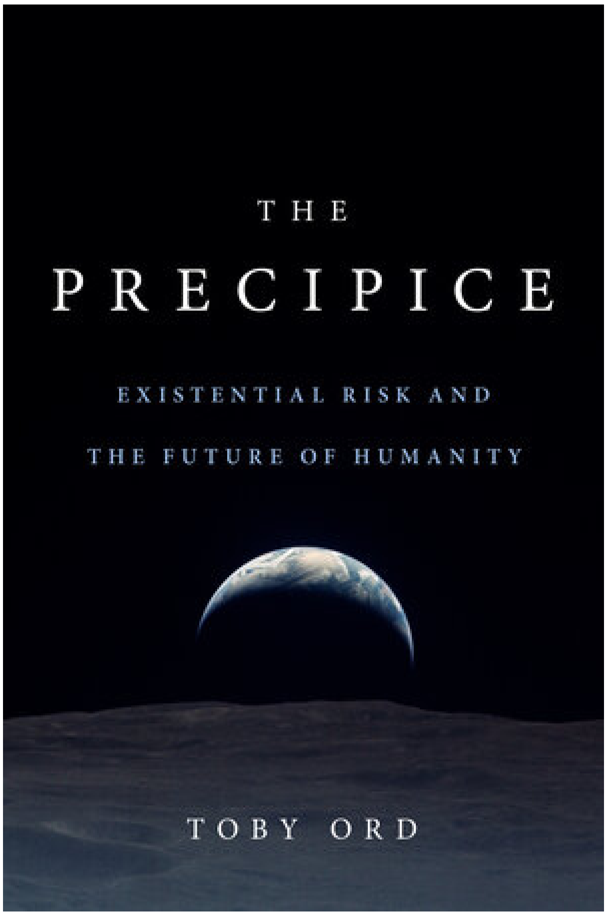
We Have the Power to Destroy Ourselves Without the Wisdom to Ensure That We Don't

29 min. read · View original

WE HAVE THE POWER TO DESTROY OURSELVES WITHOUT THE WISDOM TO ENSURE THAT WE DON'T



Lately, I've been asking myself questions about the future of humanity, not just about the next five years or even the next hundred years, but about everything humanity might be able to achieve in the time to come.

The past of humanity is about 200,000 years. That's how long *Homo sapiens* have been around according to our current best guess (it might be a little bit longer). Maybe we should even include some of our other hominid ancestors and think about humanity somewhat more broadly. If we play our cards right, we could live hundreds of thousands of years more. In fact, there's not much stopping us living millions of years. The typical species lives about a million years. Our 200,000 years so far would put us about in our adolescence, just old enough to be getting ourselves in trouble, but not wise enough to have thought through how we should act.

But a million years isn't an upper bound for how long we could live. The horseshoe crab, for example, has lived for 450 million years so far. The Earth should remain habitable for at least that long. So, if we can survive as long as the horseshoe crab, we could have a future stretching millions of centuries from now. That's millions of centuries of human progress, human achievement, and human flourishing. And if we could learn over that time how to reach out a little bit further into the cosmos to get to the planets around other stars, then we could have longer yet. If we went seven light-years at a time just making jumps of that distance, we could reach almost every star in the galaxy by continually spreading out from the new location. There are already plans in progress to send spacecraft these types of distances. If we could do that, the whole galaxy would open up to us.

Some of these stars will still be burning trillions of years from now. If we can play our cards right, we could survive for not just hundreds of millions of years, but trillions of years into the future, exploring billions of worlds in the heavens above us. This is a vast potential, the upper bound to what we might be able to achieve over that time, however you think about achievement. Whether you think of it as the well-being in every life that's lived, or as the greatest knowledge that we find, or the greatest works of art that we create, the most just societies that we create—almost all of this should be in the future, because our future is potentially so much more vast than our past, and certainly much longer than our fleeting present.

We don't often take this seriously, contemplating how small a slice of humanity we're seeing at the moment, how transient the political affairs of the day are, and how maybe the most important thing about all of these affairs is what role, if any, they will play on shaping that entire future of humanity. And yet this whole future is at risk.

We've survived 200,000 years so far—2000 centuries. There've been risks during that time, most famously of asteroids hitting the Earth, as seems to have happened 65 million years ago when the dinosaurs were wiped out. That's one type of risk we've been exposed to that whole time. There are also comets, supervolcanoes, like the Toba eruption, and risks of natural pandemics. These things have been around for 2000 centuries. The good news is that in most cases, we know that those risks must be small, because if they were even 1% per century, it's vanishingly unlikely that we would have survived for 2000 centuries. We know that the natural background of risk must be fairly safe. It's the type of thing that lets species live for a million years on average.

Humanity is not a typical species. One of the things that most worries me is the way in which our technology might put us at risk. If we look back at the history of humanity these 2000 centuries, we see this initially gradual accumulation of knowledge and power. If you think back to the earliest humans, they weren't that remarkable compared to the other species around them. An individual human is not that remarkable on the Savanna compared to a cheetah, or lion, or gazelle, but what set us apart was our ability to work together, to cooperate with other humans to form something greater than ourselves. It was teamwork, the ability to work together with those of us in the same tribe that let us expand to dozens of humans working together in cooperation. But much more important than that was our ability to cooperate across time, across the generations. By making small innovations and passing them on to our children, we were able to set a chain in motion wherein generations of people worked across time, slowly building up these innovations and technologies and accumulating power.

Often, when we look back at our history, we just look to the very first writing. We think about the first names that have been written down, the first legends, and the first histories. So, when we think of amazing things that humanity has done, we're restricting ourselves to the last 5,000 years of written language. But for 195,000 years before that, we were doing other amazing things. We had the first humans to enter each new area of the world, coming out of Africa and finding the new animals and plants of each region, working out how the ecosystems fit together, naming all of these things, understanding which plants could be used for early medicines, which animals to avoid, how to hunt them, how to evade them. There were the first humans who set foot in Australia, the strange new world with very different species that had been cut off for millions of years. There are amazing things that we've done, but most amazing was this accumulation of innovation.

There have been around about a hundred billion humans who've ever existed, and 7 billion now. These hundred billion lives have put together these innovations and accumulated information and power about the natural world and how we can influence it. This escalating power has gone through major transitions and has started increasing the rate of progress, first with the agricultural revolution, when we developed farming. That led to permanent large settlements, where instead of tens of thousands of people in cooperation across the generations, we had millions at any one time and billions across this wider network of humans. That was one major acceleration. Another was the scientific revolution, when we developed the ability to understand the world around us, to break free of dogma about that world, to properly test and discard bad explanations, wherein we could use some of this information we gained to accelerate our knowledge of the world around us and shape it towards human ends.

The Industrial Revolution of course accelerated that again when we found ourselves with access to a tiny portion of the sunlight that has shown down upon the Earth over millions of years. We use that with our innovations to lead to the modern era of sustained growth. Our lifespans doubled over that industrial era across the whole world. The poorest countries now live longer than the richest countries did back then. We've had this huge acceleration of progress and power over the world. That has led to a new transition that is probably even more important than any that have come before.

In 1945, we developed nuclear weapons. At that point, humanity's increasing power finally reached a level where we posed a risk to ourselves, not just to individual humans, but to humanity—a risk and a chance that everything that we've ever built before, over hundreds of billions of lives would come to naught. All of our achievements over these vast eons that we could live to see as a species could be severed.

We're not 100% sure whether nuclear war could destroy us, but that was the first time it became plausible that it could. Since then, technological progress has only continued and escalated. We've seen a massive amount of carbon emissions since that time, leading to the current crisis of global warming. And that's something, which, again, might be an end to humanity once more. We're not completely sure. It might be that even the worst cases of global warming turn out to be something that we can survive. But we've reached a situation where there's a very real possibility that we couldn't.

With these changes, we've entered this new period, which I call "the precipice," because I sometimes think of humanity's long history as a journey through the wilderness with occasional times of hardship and moments of sudden progress and heady views. Since 1945, we've been coming through a pass in the mountains and finding ourselves inching our way across a narrow ledge on a cliffside, at the brink of a deep precipice. In the distance, we can see more fertile and beautiful lands that we might still be able to reach, but right now we're exposed to risks. We don't know exactly how large those risks are. We can't say with precision because we've never done this before. It's not an experiment that we can run a thousand times and lose a hundred Earths in order to work out these probabilities. This is a case where we have to see the threat even without access to those rigorous numbers. But it still seems, and there's a good case for this, that this is the most dangerous time yet. This will be a pivotal moment in our journey.

If we survive, our distant descendants will look back at this time as a uniquely important time. They would see that, say, the Industrial Revolution was important, but not as important as the precipice, because this is the moment where that entire future was at risk. If we didn't have the Industrial Revolution when we did, we probably would have found access to these fossil fuels at a later point. A lot of breakthroughs that we have are a matter of *when* we make the discovery rather than *if* we make the discovery, such that what's at stake is something about speeding up the trajectory of humanity or slowing it down, as opposed to whether we flourish or whether we're forever gone. This moment is different, and if the chances are as high as I think they might be, we can't survive many centuries with risk like this.

We have to work out how to lower these levels of risk, face these technologies that pose some of these threats and manage them successfully, and we have to make commitments to govern ourselves in the future to keep the risk down to a sustainable level. We get it low, and we keep it low. That's one way the precipice could end. The other way of course is if we let the risk stay at the current levels or higher, let it keep escalating, in which case there's only so many more rolls of the dice that we could survive. It is at that level of a roll of the dice. Last century, humanity faced something like a 1% chance of ending its story in the nuclear age. This century, the risk is even higher. And I put it at about one in six, so a die roll or Russian roulette.

We can rise to this challenge and make it through. I don't think that this is inevitable in any way, so I'm not trying to make a morality tale or fable about humanity's hubris—that we tried so hard to reach for the Sun, and that led to our inevitable downfall. I don't think that's right at all. We should reach for the Sun. Technology is probably essential to humanity achieving its potential. It's certainly essential to the types of stories that I just told about how long we might be able to last if only we could reach the stars. We're never going to do that without technology. Without technology or technological improvements from beyond where we are now, we probably wouldn't even reach 1% of what we're capable of. But technology is what's creating these risks, so it's a more complicated story than just whether technology is good or bad. It's essential, but it's something that we need to be more careful in how we govern and our attitude to it.

Our timescales these days are often set by the news cycle—a couple of weeks—or by an election cycle—three to five years. If we think about this analogy to a single human life, then a four-year election cycle is like the next four hours in our life. We're in the situation where visionary people are thinking three election cycles ahead, which corresponds to half a day. Humanity takes risks with its entire future, like an adolescent taking risks with the rest of their life, just thinking about the next few hours and being massively imprudent when it comes to these risks—very short-termist and impatient. We could think of all of these virtues and vices not just at an individual level, but at the level of humanity. I call these civilizational virtues and vices. To survive this time, humanity needs to be more prudent and more patient, and it needs to find wisdom. Carl Sagan put it that way. He suggested that humanity has become powerful more quickly than it's become wise. We've had this exponential improvement in our power, but our wisdom has grown only falteringly, if at all. We have the power to destroy ourselves without the wisdom to ensure that we don't. That's why our situation is so unsustainable.

I've been thinking about just how bright our future could be, how science knows almost no limits to what we could achieve, to the durations that we could last, to the portion of the cosmos that we could discover and explore, and to the heights of quality in each of our lives or the types of achievements we could make. This is something that probably is bounded. It probably isn't infinite, but it's so vast. We haven't yet dreamt of the bounds for a lot of these things. It's this vision of this wonderful and vast future that's at stake that inspires me to think more carefully about the risks we face now and the ways that we might imperil all of this with our actions. What things can only our generation or our children's generation do in order to protect this seed of humanity so that we can grow into something even more amazing, to protect our present and thereby protect our future?

People only started asking these questions about the survival of humanity and facing the possibility that we might not survive around about 1900. H.G. Wells gave a brilliant lecture about ways that humanity might fail, and he wrote some very stirring words on the matter. There were a few more thoughts about this over the next forty years, and then it came to a head with the development of nuclear weapons. A large number of the atomic scientists who developed these weapons went on to form the Bulletin of the Atomic Scientists. They started asking, "Have we created something that could end the human story?" This was picked up quickly by Bertrand Russell as well, who wrote a lot about these possibilities, including with Einstein. For about twenty years, it was those two leading the way in thinking about whether humanity might be coming to an end, what would that mean, and what should we do about it. Their suggestion was that we needed a global government in order to get through.

In the 1980s, this was picked up again. Jonathan Schell wrote a brilliant book called *The Fate of the Earth*. At that point, he was taken by the then theory that the ozone layer could be destroyed by nuclear weapons. Earlier, people had thought that radioactive dust falling back down to Earth and killing people through radiation might lead to the end of humanity. In the early eighties, it was thought that maybe ozone depletion would lead to a whole lot of UV radiation killing people and making it impossible for humanity to continue on. He wrote a very stirring book and came up with a lot of key insights about this. He was the first person, to my knowledge, to make this distinction that losing everyone is so much worse than losing *almost* everyone because it's not just this destruction of our present, but the destruction of our future.

He did a fantastic job of combining analytic philosophy and making precise observations, precise insights and distinctions, the kind of stuff I like on this central topic of nuclear war. He did that in combination with a form of continental philosophy, of writing stirringly and engagingly and trying to come to grips in a more artistic manner with what was at stake to stir readers' emotions, shake them, and force them to confront what their taxpayer dollars were funding—the threat to destroy the civilians of the opposing Soviet Union, and ultimately perhaps to destroy the entire future of humanity.

Then, in 1983, Carl Sagan had done some interesting work with some colleagues of his, where they looked at this possibility of a nuclear winter. This arose from some climate modeling they'd been doing for other planets when they had to come up with entire planetary climate models, very simple ones, much simpler than the things we have today for climate change. He noticed that there was a possibility that the soot from burning buildings from a nuclear war might be lofted not just into the normal heights of the atmosphere where clouds could form and rain it back down again making black rain where the soot falls out, but it might be able to be lofted so high that it gets into the stratosphere above the clouds. And then there's no easy mechanism for making it fall out of the atmosphere. We might be in a situation where this soot could stay for a decade, blocking out sunlight, chilling and darkening the Earth.

Sagan noticed that if that happened, it would cause a kind of winter over the whole world. It would reduce temperatures by a great deal in the center of continents. For example, in Iowa, the reductions would be tens of degrees colder, and then less in the coastal areas. On average, it could be more than five degrees of cooling of the world, and this could lead to early frosts, including summer frosts, greatly reducing the growing season for grains and staples, so they might not be able to last long enough between frosts to get a crop. This could lead to mass starvation, perhaps the collapse of civilization, regionally or globally, and perhaps even the extinction of humanity. This was a new mechanism that he and his colleagues had discovered.

He wrote this seminal piece in *Foreign Affairs* magazine, trying to engage with the philosophy and the politics of this. What does this new discovery that nuclear weapons might, through this mechanism, destroy humanity mean for nuclear policy? What does it mean for humanity? What does it mean for being a citizen of one of these countries that is developing and deploying these weapons? There was an interesting mix of disciplines of people who were interested in this: Bertrand Russell, philosopher; Albert Einstein, physicist; Carl Sagan, astronomer; and Jonathan Schell, journalist, environmentalist, and a brilliant philosopher, but to my knowledge, he never trained as one.

Then the next year, in 1984, a fantastic book came out. One of the best philosophy books in the century (widely regarded, and I would agree) was *Reasons and Persons* by the Oxford philosopher Derek Parfit. I'm not a 100% sure whether he'd read those other two things that had just come out. He might've. He included in his magnum opus near the end of the book this idea of how our entire future might be at stake. He also clearly delineated it. He said, "Imagine three outcomes: 1) peace, 2) a nuclear war that kills 99% of all people, or 3) a nuclear war that kills 100% of all people." He said that obviously peace is better than the nuclear war that kills 99% of people, which is better than the nuclear war that kills 100% of people. But, he said, which of those two differences is greater? The difference between peace and 99% of people dying or between 99% of people dying and 100%? He pointed out, as turns out to be correct, that most people would say that the first difference is bigger, between peace and 99% of people being killed. There's way more deaths in that range than in the range below. But Parfit, perhaps independently of Sagan and Schell, noticed that the difference between 99% and 100% was bigger, because it was in that additional destruction that our entire future would be lost, and that the future is bigger than the present. That is the early history of these ideas.

Another philosopher, John Leslie, wrote a wonderful book in 1996 called *The End of the World*, which was the first real exploration not just of nuclear war, but broadening it out into all the ways our future could be lost through extinction, summarizing lots of science. In fact, a lot of the philosophers who've had to deal with this had to grapple with a lot of science, and the scientists with a lot of philosophy.

In 2002, Nick Bostrom, my colleague who works down the hall from me at the Future of Humanity Institute in Oxford, extended this idea from extinction risk to existential risk. He noticed that there are other types of catastrophes that would have a lot in common with extinction. For example, if there was a permanent collapse of civilization across the globe, the type of thing from which we could never recover, we know that extinction would reduce the range of possible futures for humanity down to just one—a world bereft of human life, no more opportunity for human action in the future. A global collapse of civilization from which we could never recover would reduce our future down to an impoverished group of people, a thousandth of the population we currently have, living lives with very little opportunity.

These things would be irrevocable in that case. We could imagine other cases; for instance, instead of a world in ruins, you could imagine a world in chains. If it were possible, as Orwell thought, to have a global tyrannical rule, perhaps a totalitarian regime whose primary purpose is to perpetuate itself, that might not have been possible to go on for centuries back with the technology in Orwell's time. But it wouldn't have to, it would just have to last long enough to have developed new technologies of surveillance that would let it entrench itself even further. We could imagine such a regime starting up soon and perhaps lasting twenty years, which might be long enough to force surveillance into every room in every house. Perhaps using AI technologies to then watch these things and flag any suspicious behavior to human authorities might then be enough for them to last 100 years longer, and then more technology will be developed and they could perhaps bootstrap into the future.

I don't know if that's possible. I hope it's not. But it's an example of another thing where the moment that such a totalitarian regime takes hold would be a pivotal moment for humanity, because at that moment, our potential would have collapsed from this vibrant, vast range of possible futures down to this very narrow range of terrible outcomes.

All of these things have in common that you lose the future, that our potential is destroyed, and that it's an irrevocable loss. Humanity is good at learning from trial and error. We make some mistakes, catastrophes strike, and we learn from sifting through the ashes. We build a better world, something that's more stable and can continue on. But existential risks, whether they be through extinction, irrevocable collapse, or through permanent dystopias, what they have in common is that you can't learn from them. If they happen even once, it's all gone. That means we have to use foresight and forward planning.

We have to be especially prudent. We can't wait until things are emotionally resonant because we all remember the catastrophe. We have to think ahead. This is so much harder. They all have that in common, and thus there are a common class of scenarios for which we need new techniques to deal with. Nick called those existential risks, and they're a major focus of my work. He's right that if you're thinking about the potential of humanity across our entire long-term future, then existential risks are the central threat to that potential. They're the things that can happen in our century, which could have these lasting effects rippling down throughout all of time.

We're not often allowed to explore things like this in academia. People don't want to publish these things. This isn't the normal type of stuff we do, but we've been trying to make a bit of room for it. I work at the Future of Humanity Institute at Oxford, and we try to create a bit of space to explore ideas like this and find a way to get people together that care about the science, to understand the science and understand the philosophy and the ethics of this.

A lot of people ask what moral philosophers do. If you look at what most people in moral philosophy do, it's trying to find a theory that would explain what we should do. In a lot of instances, they're looking at human behavior, particularly human practices that seem morally loaded in some manner, such as, say, theft, murder, or lying. Maybe something like charity on the positive side, or empathy. Then they look at these practices or emotions and try to understand what makes them appropriate, thinking about all those things to try to work out how to live a better life, and why we should even be trying to live a better life.

There's also an area of meta-ethics where they ask not just what should we do, but what does it even mean to say that there's an answer to that question. If we shouldn't kill someone, what does it mean that we shouldn't kill someone? How do you cash that out? One thing that doesn't happen often is asking bigger and more revisionary questions, such as what are the most important problems in the world? Why are those the most important problems? Is it possible that we're in the midst of a moral catastrophe? If you think about the times of slaveholding, the citizens were in the midst of a moral catastrophe—they were part of an institution that was causing immense injustice and suffering.

Are we doing something similar now? Some people would say yes. One example that people would suggest would be the environmental destruction. This was going almost unnoticed until the 1960s, when all of a sudden there was dramatic understanding of this. It moved from something that wasn't even considered part of ethics for most people, or a part of morality, to something that's considered a central aspect of being a good person. People often use whether someone recycles as a litmus test for being a good person. Similarly, animal welfare was not considered a central part of morality, at least in the West, until late in the 20th century when there was a big change in that. Similarly, a lot of people use whether someone is a vegetarian as a kind of litmus test.

These are interesting ideas that show maybe there are some big questions at stake here. Maybe there are things we're doing that are very wrong, or maybe there are things we could be doing that would be extremely good that we're just failing to do because we're blind to it. Presumably, a lot of approaches are going to be misguided, but I wish more people would ask these types of big questions.

One school of thought that's related to that is something that I've been developing along with my colleague William MacAskill, which is an idea that we call long-termism, which is thinking about the long-term future of humanity, and maybe the long-term past as well. It’s trying to get people to think beyond the present. Normally, when people think about moral action, whether something's right or wrong, they're thinking about the consequences, or the motives behind the act that are present right now or in the very near future; for example, whether it would hurt other people who currently exist. But a lot of our actions affect the long-term future of humanity. And because this long-term future could be so vast and some of these actions might have lasting effects over this future, it could be the long-term effects of our actions that are of central importance, perhaps more important than the effects over the next few years. That's something we've been exploring. A key part of that would be these ideas around existential risk. It could be that living a good life, or at least one of the best lives that you could lead today, is about helping the long-term future.

I've often tried to tackle some of the big picture questions facing humanity, tried to make a contribution rather than trying to tackle something that's a small part of life. I worked on global poverty and global health, trying to work out how that fits into our lives and what we, people in rich countries, should be doing about it. One of the things I first noticed was that some ways of helping were much better than others. When you ask people what we should be doing, their answers are often about giving money to charities that are working to help people in poor countries. That's right. But what they don't notice is that some ways of doing good in poor countries are much more effective than others. That observation has been noted. That's one of the reasons for skepticism about aid, to point out that a lot of aid does a lot less good than we think it would. Perhaps it gets wasted. In some cases, there are negative effects, but that's not a good reason not to donate to charity. It's a good reason to donate to the right ones, to the ones that are having positive effects.

When I looked at some of the evidence on this and did some mathematical analysis (something that's not very common in ethics), I noticed that if you look at the health interventions, ways of helping people across the world, some of these were ten, or a hundred, or even thousands of times more effective than others. This was data through the Disease Control Priorities project. Their data suggested that it wasn't the case that most things you could do are within, say, a factor of two of each other in terms of how much they help. But it was easy to find things that were ten or a hundred times more important. In fact, one piece of analysis showed that if you took any two health interventions at random and funded them to the same amount, on average, one of them would do a hundred times as much good as the other.

You could see that in a positive sense, one does way more good. Or you could see it in a negative sense, that in one case you'd be squandering 99% of your money compared to how you could have effectively used your money. This struck me as a supremely important aspect about this question. It's not just about giving more, but giving more effectively. And those things are multiplicative. If you do both, they're better than the sum of their parts. They're the product of their parts. A lot of people could give ten times more to help those in need and to give it a hundred times more effectively, increasing their impact by a factor of a thousand. I tried to put ideas into practice in my own life and work out how that would play out. I found charities working on some of these interventions that were found to be the most effective and started donating my money. I heard from other people who wanted to join me, so I set up a charitable organization to do that. It's a society of people who make a pledge to give at least 10% of their income to help others as much as they can.

That developed further as well. I was joined on that project by my colleague William MacAskill, and together we and others founded a broader philosophy called “effective altruism”—people trying to use their money to do as much good as they can, and also trying to use their careers to do as much good as they can. When they find out that, say, $1,000 can save a life, their reaction is not okay, here's $1,000. The reaction is, what would $10,000 do? Save ten lives? How many lives could I save over my whole life if I took this seriously? If you do these calculations, a typical person in a rich country, if they took this seriously and lived on a very modest amount, might be able to save about a thousand lives during the course of their career, which is about as many as Oskar Schindler saved.

When we think about the history of ethics and ethical problems, we often think of these small moments where there was this possibility amidst tragedy for heroism, such as with Schindler, where someone could take these risks in order to save a thousand lives. That was a silver lining in such a dark time, that there were moments for this kind of action. But we could do these in our own lives. You'd have to make a considerable sacrifice in order to save a thousand lives through your career in most cases. But it could be done. And the sacrifice is probably less than that facing Oskar Schindler. There are hundreds of millions of people in a position to make that choice. By trying to think about this effectiveness and take these numbers seriously, it opened up all of these new possibilities for how to think about an ethical life, less focus on the age-old questions of how to avoid wrong action—lying, cheating, stealing, killing. Not that people would do any of those things in effective altruism, but that's a low bar, not doing those types of behaviors.

We should be thinking bigger than that, not just about how can we avoid those wrong actions, but how we can make the world a much better place. How can we help others as much as possible? How can we use reason and evidence in light of doing that? How can we use the best science and mathematics to try to make a difference? That's been something I've found to be very neglected, partly because it involves thinking about ethics, mathematics, and science all together. These are things that a lot of those people I mentioned did in the early days of thinking about extinction risk, existential risk, and nuclear risk. In order to ask and answer these big questions, you need this quite interdisciplinary approach. Maybe that's why it doesn't happen more often.

I started off my studies in science, not in philosophy. I was in computer science, particularly interested in theory of computation logic and artificial intelligence. I was interested in creating other minds. But I also was interested in politics. The political debates made me wonder what the fundamentals were. I noticed that there were a lot of things that were disputed, some of which were facts, but other things were values, where people would sometimes agree on the outcomes of a policy, but disagree on whether it was a good idea or not. It brought me into ethics, which I did alongside my science. As I kept going, I did more and more philosophy. I found that you could ask a lot of the questions I was interested in about the nature of minds, the nature of logic and computation within philosophy as well. I ended up doing both. And there are various people who inspired me within philosophy, who made me realize that there was something there I could do.

There was lots of philosophy that I thought wasn't so inspiring, where I worried that it was all a bit of a game and that there wasn't that much at stake. Maybe some of the questions were interesting and it was fascinating trying to see whether you could rebut the skeptic who thinks that no external world exists. What could you say to them? But on some other level, we would go on assuming that an external world existed, regardless of whether we could answer that question. It wasn't clear that there was that much at stake. Peter Singer was someone, another Australian philosopher, who showed me that you could turn to key ethical questions about the world around us and make a lot of progress on those questions, both in terms of the ideas and in terms of taking them to the world around us.

His key contributions were the book *Animal Liberation* and his work on animal ethics, which started a movement, and also a key paper of his called "Famine, Affluence and Morality." It was one of the first papers he published, which was pivotal in terms of the ethics of global poverty and what we could do to help others less fortunate than ourselves. One thing that connected with me was that he didn't just have these nice arguments about what we should do or how it might be a moral obligation on ourselves to donate, it was that he set a high bar for himself. He took these ideas seriously. He knew his own life, both in terms of animals and in terms of global poverty and donating his money.

That kind of moral seriousness, that kind of willingness to follow ideas where they go and then adopt them in your own life, is something that showed me that there was something serious here. I was also inspired by the work of Derek Parfit, his book *Reasons and Persons*, which I read when I was studying in Australia. He did not do any mathematics and has always claimed to be symbol blind. If he sees some kind of formula with a capital Sigma in it, his eyes just glaze over and he doesn't get it. Yet he was so precise and clear in the way he did philosophy that I couldn't believe it when he told me that. I assumed there must have been all of these things in these books, so I went back and looked and indeed they weren't.

He was doing mathematics, he just didn't use the symbols. He was very clear and conceptual, and ultimately had a very mathematical mind. That concision and clarity inspired me in how I was going to approach philosophy. When I came to Oxford, I was lucky enough to have him as a supervisor for my dissertation. I was also inspired by my other supervisor John Broome, who used to be an economist and then switched to ethics. He showed how a lot of mathematical tools from economics could be used to make central points in ethics. He ended up in the White's Chair of Moral Philosophy at Oxford, the most prestigious position here in ethics, despite being an economist by training. He saw that economists were very good at what follows from a key set of assumptions, a set of axioms.

They had the methods that philosophers didn't to find out what follows. If you start with these four assumptions, such as Arrow's impossibility theorem, about voting, he showed how some of these could be applied to ethics and things like the von Neumann-Morgenstern axioms of expected utility theory, in economics, which shows how you can take people's preferences—do they prefer chocolate ice cream or vanilla ice cream?— not just over particular outcomes, but over gambles, chances of getting different outcomes, you could turn those kinds of ordinal preferences into this cardinal structure. You could start to say things like the difference between chocolate ice cream and vanilla is ten times as important as the difference between vanilla and caramel.

John Broome showed how you could take the mathematics of that and apply it to ethics, not just to the structure of human preferences, but to the structure of the good. You could start with a fundamental idea about what's better than something else, which chances of outcomes are better than which others, and from that you could create a cardinal, a numerical structure, of the good. He made a lot of progress with these ideas, showing how if you take some kind of expected utility theory over chances, that affects how you should distribute benefits between different people, building on the work of others, such as Harsanyi. There are amazing groups of people who took these ideas seriously and sometimes saw them from these different perspectives.

I'm also inspired by others, like Carl Sagan. His work and writing showed me that you can be a scientist and you can write books for a popular audience, books that are beautiful. At the same time, if you look at what the sentence is saying and the maths behind the science, you can see that the sentence is exactly accurate, that it's a poetic way of describing reality, but it's still a description of reality. It's not popularizing the science, it's doing new science in a way that crystallizes out the essential aspects and presents them clearly and beautifully. That blew my mind. That's been an inspiration to me in trying to write in that way.

What I'm excited about at the moment is thinking about this long-term future of humanity and thinking about institutions, such as our democratic institutions. Most of the people affected by the decisions of our government don't get a vote. We've had progressive expansion of franchise, for men, women, people of all races, everyone above eighteen. Yet most people who are affected by our actions are people in future generations, people who don't exist yet but will benefit or suffer under the effects of these choices. And when you think about what justifies a democracy, it's this consent of the governed. But how do we deal with that? Are there ways in which we could take steps to give a voice to the people of the future? If we could, that could help to resolve some of these challenges about taking the future appropriately seriously by giving it some kind of political power, whether that be soft power, as it's been tried, say, in Wales where they have a commissioner for future generations who can ask these questions of government and demand answers, or perhaps even hard power. Perhaps having a political chamber to review legislation in the interest of future generations representing those generations. How would you design such a thing? How would you make sure it didn't get captured by current political interests or corporate interests? Is that impossible? Are there ways to do it? What about at the international level? How can we guarantee our future?

Now that we've realized the fragility of our present time, how can we put into motion the institutions that are needed to get risk down and keep it down forever? Could we, through new international institutions or changes to existing ones, write a constitution for humanity that would put in place safeguards to keep us away from the edge of the cliff, but would leave other things open for us to decide in the future what kind of possible future we want to realize. Is it possible for people this century to literally or figuratively write a constitution for humanity and be the founding fathers and mothers of the future, and thereby perhaps to ensure that we set it on a course towards achieving this bright potential that we have over the distant time? That's what fascinates me at the moment.