#262: Humanity on the precipice (Toby Ord)

Julia: Welcome to Rationally Speaking, the podcast where we explore the borderlands between reason and nonsense. I'm your host, Julia Galef, and today it's my pleasure to speak with Toby Ord.

He's a philosopher at Oxford University, and one of the founders of the effective altruist movement, which focuses on using reason and evidence to figure out how to do the most good. He's also the author, last year, of the book *The Precipice: Existential Risk and The Future of Humanity*. Here is my conversation with Toby Ord.

Julia: Toby, welcome to Rationally Speaking. It's so great to finally have you on.

Toby: It's wonderful to be here.

Julia: So Toby, you've recently published a fascinating and important book called *The Precipice: Existential Risk and the Future of Humanity*. That's going to be the main focus of our conversation today. So could you start by just laying out the basic idea in the book for my audience?

Toby: Yeah. It's fundamentally a book about humanity over deep time. It's about the 200,000 or 300,000 years of humanity that's come before us, over about 10,000 generations. And about how the history of humanity, even then, might still be just beginning.

Because if we last as long as a typical species on this planet, then we should last for about another million years. And there's not much stopping us lasting for hundreds of millions of years, as some of the more long lived species have done. Or perhaps to outlast the earth itself. If we in the future are able to travel to other star systems where these stars have longer lifespans, or are younger stars, we could potentially carry on in this way for trillions of years.

So the history of humanity may really just be beginning, and we could have a very bright future ahead of us, if trends in increasing longevity and prosperity continue.

However, ultimately, humanity's history has been one of escalating power. And we finally came in the 20th century to the point where our power had grown so great that we could pose risks to our own continued existence. So I date this from 1945 when the first atomic bomb was detonated.

And so with nuclear weapons, and then also now with climate change, we have technologies that are potentially so transformative to the environment

around us, and so potentially able to cause great problems for us, that our own future is imperiled by these things.

And so I think that this question of, "How do we adapt to being in a situation where we pose these threats?" Is a really central issue of our time. And we may be at crossroads. I call this time "the precipice," because we may be inching our way along a cliff ledge where there's a chance of an irrevocable failure.

Julia: Toby, you've been focused, for your entire career that I've been aware of, on doing good, on improving the world. But 10 years ago, you were more focused on more traditional ways of helping the world, like reducing poverty and disease. And now your focus is on improving the long term future of humanity, and navigating us past the precipice.

I'm curious about what prompted that shift in focus for you, from near-term concrete problems, like poverty and disease, to the very long term. Was there a particular argument in the last 10 years that changed your focus? Or can you not pinpoint it that precisely?

Toby: Yeah, the biggest change ultimately was in 2003 when I came to Oxford. And Nick Bostrom had just got to Oxford as well. And we were put in touch with each other and soon turned to the issue of existential risk, an idea that Nick had just published a paper on at the time.

And I thought this was pretty intriguing. That I was very focused on global poverty as one of the biggest issues in the world, and Nick made a strong case that actually protecting humanity's future from existential risks was an even stronger priority for humanity.

And I didn't entirely buy it, but I thought it was very plausible and really important idea…

Julia: Why didn't you entirely buy it?

Toby: So I should say, even since then, I have cared a lot about this. So 10 years ago when I was still in the process of getting Giving What We Can really running, I was donating money towards global poverty and global health related charities — but also in terms of my research time, still doing a lot of thinking about existential risk back then as well.

I think one of the reasons I didn't buy it… there's kind of two. And in some sense, disarming those two reasons have probably been the main reason for my gradual shift towards focusing on this over that time. Whereas I actually went for a very long time with one foot in each camp on this, splitting my

time — in a possibly unproductive manner. It's always challenging if you get too fragmented.

So one of them was a concern about the contrarian framing, I think of this. That it sounded just kind of surprising, and like, "Hey, it turns out you might have thought that helping people who are suffering greatly is the most important thing to be doing. But actually, saving the world is what you should be doing, say from a giant asteroid impact or from a rogue artificial intelligence or something." And it sounds like a comic book plot or something.

I have now come to realize that indeed, this is the type of question that, in terms of fiction, mainly gets dealt with in comic books and superhero movies. Now, I think that's a real shame for literature, that it doesn't actually grapple with these questions, in works in the canon, because I think that these are very important threats. There are some works that do grapple with it, like *Frankenstein*, but generally not.

So I kind of assumed that if this really were so important, there'd be more said about it, by public intellectuals and moral thinkers over time. And only later on, did I really realize that actually, there was this. And it happened with nuclear weapons. And that there were serious thoughts and important thinkers who were writing about the existential risk posed by nuclear weapons.

And that this was being taken seriously by the public as well. The biggest march in American history at the time was against nuclear weapons, shortly after discovering this possibility of nuclear winter and that it may kill everyone.

So I gradually realized, hang on a second, we've already been here. And I've, in some sense, got the misfortune to be from the generation raised just after the end of this Cold War, such that that wasn't obvious to me and it seemed more contrarian. But actually I think that it's possible to make this fairly obvious as to why it is that there are real threats to our future. And that the complete destruction of everyone you love and everything that we could ever achieve, and every tradition that you've ever cherished, why that would be a really bad thing. And you can actually make that seem quite obvious and not contrarian.

So that was one of the things. And a related one, I guess, was that I felt like… to some extent, in my life, that if I focused on this thing, it was taking a risk, in some sense. Because I felt it was a less reliable type of argument. I knew that working to help the lives of people who are much less fortunate

than myself and suffering from a range of terrible diseases that are easily treatable — I knew that that was morally important, with high confidence. Whereas this other thing just felt a little bit more like, "Huh, maybe I could be confused about this, and later on regret and think that that actually wasn't as important, and I was mistaken."

But then I realized that ultimately wasn't the right way to think about it. That instead it is better to think in terms of a portfolio of action by altruistic people across the world, trying to do good.

And that rather than thinking of my life and that my efforts had to be divided between things that could be good. so as to not go all in on something… rather, to think that, actually it's probably good — given that there are hundreds of thousands of people who strive to help the world in various ways — it's good that a lot of those people go all in on something, and focus on it. And that these existential risks are much more neglected.

That helped me realize that it was less of an "all or nothing" thing, really.

Julia: I do want to talk more about the challenge of working on something that's kind of abstract — at least, compared to reducing the number of people who die from malaria each year, where you have a quantitative measurement of your impact, and you can get feedback about how you're doing. Compared to that, yeah, there's not really a counter in the sky that tells you, "That conference you held about existential risk reduced our chance of extinction by 0.001%," and then you can measure the cost effectiveness of that. We don't really have that. That is on my wishlist of questions to ask you.

But before we get too far into the weeds, I wanted to just talk about some of the things in this category of existential risks that you discuss in your book. So you catalog various things that could potentially pose an existential threat to humanity, like asteroids or nuclear weapons. Could you first explain what you mean by an existential threat? Because it's a little more complicated than just wiping out humanity, right?

Toby: Yeah. So an existential catastrophe is something that destroys humanity's long term potential. That's how I put it in a pithy way.

And then to unpack that a little bit, this would include something like extinction, where if humanity went extinct, it's clear why it is that our long term potential has gone. And that there's basically only one path forward, which is whatever would happen without any human action.

Whereas, it's also somewhat similar though — and this was a big contribution by Nick Bostrom — to realize that, what if instead, we had some very severe collapse of civilization and we were reduced to a pre-agricultural state? So, back to a kind of foraging past. And perhaps through say the climate being ruined, or something else, we could never find our way back to civilization.

In that case, there would be something very similar would've happened. In that our once soaring potential would've been reduced to this very narrow range of meager options.

And also that it's the same kind of catastrophe. In that if it happens once, then that's enough. Such that you can't learn by trial and error when it comes to things like this. You have to avoid it happening for the first time. Which means that if we did want to survive for a million years, we'd have to get through 10,000 centuries without ever once falling victim to such a risk. And that requires some different ways of thinking to what we're typically good at.

So you notice that these things would have something in common there. And then also you could include dystopian outcomes that are locked in, in some important way. Where, say, a global totalitarianism that reinforces itself, such that there's no or almost no way to escape it, once it's been established… In that case, the moment it gets established would be that the moment where our potential is crushed.

So I'm not saying that it has to go down to exactly zero long term potential. You could imagine something like, we lose more than 90% of the ceiling of what we could have ever achieved because a certain thing happens. That's an existential catastrophe. And then an existential risk is just the risk of something like that happening.

Julia: When you talk about humanity achieving its potential, I suspect for at least some of our audience that conjures up *achievement*. Like building things and inventing things and exploring things.

And I suspect that's part of what you mean, but that you also mean… feeling happiness and fulfillment. And just the subjective experience that we could end up achieving. Is that right?

Toby: Yeah. So what I'm trying to do here is to be very open to different ways that this could be cashed out. I'm probably most sympathetic to theories that say that it's mainly to do with well-being. So, how well lives go. And perhaps to do with the happiness and suffering in those lives, or some other

components of what makes a life go well, that we could broadly call "flourishing."

But there are also views that focus on achievements such as the greatest discoveries that we have. Whether we finally create truly just societies, whether we discover the workings of reality, or whether we make truly great art that surpasses anything that's come before. So if you care about those things, you should also care about the future, because that's where most of that would lie as well.

So I'm trying to have this broad approach. And I also want to be clear that it also cares about animal well-being, and also about the well-being of any entities that come after humanity. Any of our descendants, either future species that we evolve into, or design, or anything like that.

So my focus on humanity is not because I think that all the value in the world resides inside human lives. Rather, it's because we are the key agent here. It's not monkeys or blackbirds that are going to make these decisions about how to design AI systems, or how to protect the future, or whether life gets taken to other planets around other stars. Ultimately, it's all up to us. So it's us qua moral agent, rather than as the moral patient.

Julia: Right. Well put. If you had to pick one risk that you think people worry too little about, and one risk that you think people worry too much about, what would they be?

I guess this is like a round of "Overrated, underrated," but with very high stakes.

Toby: Yeah. I would say it depends which people.

Julia: You can answer separately, for separate meanings of "people."

Toby: Well, here's a kind of answer. The risk that they overrate is climate change. And the risk that they underrate is climate change.

Julia: That's a great answer.

Toby: For somewhat different people.

I think that there is a perception sometimes, in the popular media and popular discourse, that climate change poses, say, a 2/3 chance or something like that, of destroying humanity's entire future. Something along those lines. Or that if we don't act within five years, then more than

something degrees of warming is locked in, which means that we basically have no future. Or something along those lines.

I don't think it works like that.

I think that climate change is extremely serious and severe, mainly through most of the possibilities that don't involve existential catastrophe but just involve severe hardship and loss and environmental damage, perhaps lasting for hundreds or thousands of years. So that's not to belittle any of that, but to say that it's not the same as existential risk.

But that said —

Julia: And those are scenarios where there's warming on the level of a few degrees?

Toby: Yeah. Even if there's warming on the level of say five degrees, it's very hard for this to lead to human extinction.

And I believe, although it's more debatable, it's very hard to lead to a situation where we can't maintain civilization. I think people have different thresholds for that. When I talk about whether we have civilization, I mean, do we have writing and cities? The kind of thing we had 5,000 years ago. Rather than do we have industrial revolution and modern liberal democracies or something like that? Which is I think a more common bar, but that is not required in order to have civilization.

And then within rationalist and effective altruist circles, I think that a lot of people underrate climate change as an existential risk. Because they just assume that these kinds of more typical scenarios are all that there is. Or that because they can't see a way where it could destroy our whole future, that there can't be such a way.

And in my view, it's possible that in, say, a couple of 100 years time, when this science is much more developed, we'll look back at this and see that there really wasn't a way that it could destroy our future. But we don't know that yet.

And it would be making such large changes to the world that… I kind of imagine a representative of humanity appearing at the pearly gates, and Saint Peter accosting them and saying, "Okay, so you destroyed everything through climate change." And you say, "Well, but we really couldn't see how it could have killed everyone," or something.

I just think it'd be a very lame response. Like someone who was extremely negligent in an industrial accident, or something. Where you'd say, "Sure. We knew it was the largest change that we'd ever made to the Earth's environment, by a very large margin, and so on. But our models said that it couldn't go quite that bad." And it's like, "Well, have you ever had problems with your models before?"

So I feel that there is a realistic chance and that that is underrated.

And also I admire the people who care about this, a great deal. Because I think there's generally neglect for existential risk in society these days. And also for really long term thinking about these ways that there could be irrevocable losses that are felt over the entire future. And that that gives a special leverage and importance to our actions now.

And I think that environmentalists are among the first groups to really notice such things, when it comes to either extinction of species, or of ecosystem collapses. That there are things that have this structure. And that potentially could even rise to the level of risking human extinction.

And I think that they deserve credit for that. And I think it's the right style of argument. I'm not sure exactly how much risk climate change poses, but I think it's more than some of my colleagues might think.

Julia: I actually think that climate change is both overrated and underrated among the general public, not just effective altruists and rationalists.

Because even the mainstream climate scientists, and the IPCC reports… They tend to just focus on the 85th percentile case of badness. I could be wrong. I haven't really closely followed this discussion, but it sure seems like there's not that much attention paid to the 95th percentile badness scenario, and trying to model that and how bad that would actually be. And even if there's only a 5% chance of something that bad happening, if it's really bad, maybe the bulk of our worry should be directed towards that.

There's a weird disconnect that I've noticed where people talk about the modal probability outcome of climate change, as if it's existentially bad — and it's not. But they could be talking about the 95th percentile outcome of climate change. And that might actually be existentially bad, but it's unclear. We haven't really discussed it very much.

Toby: Yeah. You've said it better than I would've, actually.

And you can see why having spelled out like that, why that's an area that is ripe for public confusion. It is a complex message to say, "Climate change,

on the whole, with the most likely scenarios, are going to be somewhere between bad and extremely bad, or something. But then also there is a lower probability chance that things really are a lot worse than we expect, in which case it could be beyond extremely bad. It could be the end, the worst things ever happened in humanity's very long history. And that, even though that's a small chance and one that's very hard to quantify, that could still be the most important thing about it."

And that's a very nuanced message, which is hard to maintain.

Julia: Right. It's hard enough just trying to convey that "Existential risk is not equivalent to really bad risk." That's already a level of nuance that I find difficult. And what we're describing is like four or five notches harder than that. So I understand why the point hasn't gotten successfully conveyed yet.

I wanted to talk about one of the risks in your book that you put the most weight on, in terms of the probability of it happening and the probability that it could in fact, be an existential catastrophe. Which is the risk from advanced artificial intelligence.

And I have a couple questions for you about that, but I was hoping you could just first briefly summarize the case for why that's a risk. And I know there are whole books that have been written about this, so we don't need to go into too much detail, but you can give the high-level, one to two paragraph summary.

Toby: Yeah. I think the simplest case is something like: We sometimes forget, when we look at the AI products that are being produced at the moment, that are often quite narrow, that the original dream of artificial intelligence by the early pioneers was to create systems that had the general intelligence and ability to achieve a very wide range of goals in a wide range of environments, that humans have. And that was the dream.

And in the last decade or so, we're actually coming a bit close to that. We've had things like DQN, DeepMind's Atari playing agent, and also even more so AlphaZero. Another DeepMind product that can learn a bunch of different games, including Atari, but also chess and go and be superhuman performance at all of them. And systems like GPT-3, which is a conversational system from OpenAI. That are much more general than anything we'd seen before.

When surveyed about this, the AI researcher community give a very wide range of different answers as to when they think it would be 50% likely that we will have advanced AGI systems, artificial general intelligence, that can

do most of the jobs that humans can do. But they think that it's something like 50% likely to happen this century. So that's to say that there's a pretty decent probability, this century, that we reach such a threshold.

And if we look back at human history and we ask, why is it that humanity is in this very privileged position? Where our fate is in our own hands, really. Where for every other species, their fate is generally in human hands. But we are in a situation where we have this potential to achieve whatever we want, to shape the world in all kinds of ways… And it ultimately comes down to something like our intelligence. Our cognitive abilities. Rather than to do with how fast we can run or how strong our claws are or something like that.

And yet, the AI researchers are predicting something like a 50% chance that by the end of the century, we're no longer at the top of this totem pole. That we've created systems that can outdo us at the thing that we're best at. And so then the question is, why is it that we would maintain our ability to call the shots in that situation? Why wouldn't it be that our future could be crushed if these other systems so desired it?

And I think there are some answers to that. We certainly want to try to make such systems safe, to make them either aligned with our values, such that in creating their own ideal world, they create ours… Or to make them controllable, such that they will do what we ask them to do.

But the people who are working on those things, are finding this extremely difficult, to work out how to program that. And it's not clear that they'll be able to do that on the timelines before we have systems that are so powerful, such that those systems might be uncontrolled.

So if you put those two things together, something like a 50% chance that we'll have such systems this century. And then maybe something like 80% chance that we manage to solve the problem of controlling them… that would still leave overall a 10% chance that we've built such systems and fail to control them. And that our future is no longer in our hands, and we're at the mercy of these systems.

That's a rough case for this.

Julia: So there's definitely a lot more mainstream agreement now about the potential risks from advanced artificial intelligence than there was 10 years ago. Prestigious professionals in the field of machine learning, and politicians, and other public intellectuals take this seriously as a risk, when that would've been unthinkable 10 years ago.

In fact, I was chatting with a friend of mine who runs a tech company, and we were talking about risks from AI, and he actually said, "One reason I'm skeptical is because there's such consensus that we should worry about this, that I feel like there's probably a lot of social pressure to believe it." And my jaw dropped. I was like, "I can't imagine someone saying, even a few years ago that there's 'so much social pressure to worry about AI risk.'" And he wasn't like a rationalist EA nerd. He was a mainstream person.

So the landscape really has changed just over the past few years — but there are still plenty of smart and thoughtful people who think worrying about risk from AI is a pretty silly thing to worry about. Do you understand why they disagree with you?

Toby: Yeah. I don't think there's just one reason, but here are a few.

One, is that I think a lot of people see the probability of being able to build such an advanced AI that has this general intelligence as very low. Now, I think, as it happens, I think it's something like 50-50 chance this century, probably a bit more. And the machine learning researchers think it's about that. And also the general public think that as well.

But that average is taken across a lot of people. And so many of the people who are in those categories think it's much lower than that. And I don't think they should be so confident, given that there's so much disagreement with them, that they're far from the median on that…

Julia: And given how many examples there are from the past, some of which you talk about in your book, of people who were very confident that a technology would never be developed or would only be developed in the far future — and then it was developed a few years later. Like the plane.

Toby: Indeed, or the nuclear reactor.

Julia: Right.

Toby: In both cases by the people who invented them, so particularly telling.

So yeah, perhaps mistakenly, some people thinking that the probability is very low. Also, thinking perhaps just that there would be plenty of early warnings. That it's not the thing that goes from no major industrial accidents, to the end of humanity — but rather, there would be some warning shots where some really serious things happen that will give us time to wake up to the risks and deal with them.

I think that for many risks, that's a pretty good argument. I think that there's a case to be made that with artificial intelligence and the types of risks that we're envisaging, that you may not get such warnings. But I think that's a reason that they might think, "Well, this isn't one of those cases where we have to plan in ahead, 20 years in advance of developing the technology. Rather, like most technologies, it'll be fine to have things develop in concert with the technology."

Perhaps being unaware that it's so far proved extremely difficult to control and align AI systems. And that the people who are working on those abilities at AI companies and in academia are having a lot of trouble doing it. And there seem to be some reasons to expect it to be very hard. So I think that a lot of the critics are not aware that the people working on this are finding it challenging. It's not as simple as many of the things that they suggest.

And then finally, I think one aspect is just not thinking about this probabilistically. I've said that I think there's something like a one in 10 chance that humanity doesn't make it through the century with our potential intact. That we suffer one of these existential catastrophes due to advanced artificial intelligence. And maybe a one in 10 chance spread out over a whole century.

And maybe that just is compatible with what some of the critics think. They think, “It's not going to happen.” Well, maybe it won't happen. I say there's 90% chance it won't happen, or that we will avoid the downsides.

The thing is that… how low a chance does it actually have to have before it's a good idea to be trying to make public arguments with people saying not to worry about it? When there's very little worrying about it, and there's quite a lot of powering ahead. It seems like you'd have to be pretty confident that the probability was less than say one in a 1,000 or something. And that you are sure that you are right, and not your peers who are disagreeing with you, such as AI luminaries like Stuart Russell or Demis Hassabis, or people. And I don't really see how one could be in that epistemic state.

It's also a bit suspicious, if you are someone who works on building a technology that other people are saying could destroy our future, including the children of your critics and so on. And that you're like, “I don't need to listen to your arguments,” or something. I don't know, there's something there, where there's some extra onus to take the other person's argument seriously, if you are the person who is potentially unilaterally imposing the risk.

Julia: Yeah. That's a good point.

I guess a variant that I think seems pretty reasonable is just that… it's not that AI won't be a big deal, and couldn't have some potentially transformative and serious repercussions for society, but that it's just way too early to productively worry about it. That we have no idea what form AI is going to end up taking. We don't yet know how to build it. And so there's really nothing we can do now to try to reduce our risk from AI, while the technology is still in its infancy.

That's an argument that I've heard from a number of people I consider very thoughtful and reasonable, who were not just dismissing this out of hand as weird futurist nonsense.

Toby: I think that that's a pretty good argument. I think that's, that in as far as it goes. I think that in general, the further out a risk is that we're trying to deal with, the more our efforts are likely to be wasted — by either inefficiency, because they're not quite targeting the right thing, or perhaps just being completely useless, due to additional major things that happen that we just weren't aware of.

So you could think of this as a kind of nearsightedness problem — that the closer something is to the present, the more accurately we can perceive it. And the further away it is, the less accurately. So therefore things that are further away, there's a kind of multiplier or something, where we multiply our impact by 0.1 or something. We end up having a lot less impact than we thought.

It's even more so if you're trying to deal with existential risk in general, and it could be the case that AI was not the one you should have been tackling, and that we later on find that out. And that's a way in which efforts that are done today could end up being less important than the same amount of effort done later.

That said, there are things that point in the opposite direction. So there are ways that we can try to steer the course of a discipline. For example, Stuart Russell talks about this. He thinks that the idea about controlling AI systems and being able to point them in the right direction at a very nuanced goal that they're trying to achieve — that that is just part of AI. It's part of what it means to try to develop intelligent systems, and that this should be made a more core part of the discipline, rather than as a kind of strange add-on that no one really wants to do. So his attempts to do that, it's better the earlier they start, if he's trying to actually steer the direction that the field goes.

So that's one type of thing there. Steering things is better to happen earlier.

Also, there are some things to do with growth — if you're trying to build a field, and you think that ultimately, say by the time we develop these systems, that something like a tenth of all AI research should be on controlling these systems, and you start with much less than that… Then maybe you need to start this process of exponential growth, among the field of people who are looking at controlling such systems, you need to start that earlier.

So I think that the idea that we're early compared to the crunch time, it cuts both ways. And so while I think that they've hit upon a valid and important argument, the bigger picture shows that maybe that's an argument that they should be spending more time now thinking about steering things, or about growth of fields, and less time on object related work. Except inasmuch as it's needed in order to tell us where to steer things or to help the field to grow.

Julia: I guess a common theme, to both this case, and the case of 95th percentile bad climate change, is that… there's a common probability error, I think. Where the reaction people often have is, well, if something is really uncertain and in the future and hard to reason about now, then it's silly to try to worry about it. And your argument is, no, if something is uncertain and far in the future and really high stakes, then we should be spending at least some effort trying to reason about it and see if there's anything we can do, even if nothing is obviously jumping out at us right away.

Toby: Yeah, I think that's right. And there is a question about how much of that effort… That's a fair question. And I think that in part, there are people who are concerned about this have maybe been in bit too successful in attracting attention, or something.

Because some of the arguments that say that "It's like worrying about overpopulation on Mars," or what have you, then continue to say that it's unfair that so much funding is going on AI safety.

Julia: So much funding?

Toby: Yeah. So you hear this a bit, and it's not based on the facts, but it might be based on some kind of appearance of things. So yeah, there's some kind of interesting disconnect there — that maybe it's more visible than is its share of actual work going into it, or something. So that can fool people into thinking, "Hey, why is that thing so big?" when people are not aware of its actual size.

Julia: Right, that's a good point. It reminds me a little bit of this thing that happens, where if people feel like some person or some idea is getting more status than it deserves, they will talk about it as if it deserves *no* status. As almost an instinctive corrective, to what they feel is an unfair allocation of status. And so maybe there's something like that going on.

Toby: Yeah. I think you're exactly right. It's confounded by the fact that there are often these articles, these kind of what we call these Terminator articles that appear in newspapers with a giant picture of the Terminator, and then say something very ill-considered about AI risk.

Julia: Right.

Toby: And the people who are concerned about AI risk and the existential risk from it, they don't count these Terminator articles as being on their side.

Julia: Right.

Toby: They think that they're this annoying thing that keeps happening.

Whereas the people who are trying to focus on AI capabilities and just making the systems work at all, they think that these Terminator articles are part of the safety landscape and thus that there's heaps of stuff going on there, and there's heaps of attention and so forth. I think that's part of this disconnect. Therefore, they think that it's really overrated. Because if you counted those articles in the newspapers as being part of the core, understanding it, then maybe it would be overrated.

Julia: So I want to zoom back out to existential risks in general. And a difficulty in thinking about how likely they are is that normally, if we wanted to estimate how likely something is, we would look at how often it has happened in the past, to get a base rate.

And we can't exactly do that with existential risks. Because by their very nature, there's no track record of humanity going extinct, because if it happened even once, we wouldn't be here reasoning about it.

So how do you think about getting an estimate of the base rate of existential risks?

Toby: Yeah. I think this is a fascinating question, and it gets into the questions about the nature of probability as well, which I'm sure you and your audience are very interested in.

So one can look at human history. So as I said, there's been about 200,000, or we now think maybe 300,000 years of Homo sapiens. So you can use the fact that we've survived for, let's say 300,000 years, as some way to get an idea on what is the per century risk that we go extinct. And I think that idea works.

You can't quite do it the naive way. So if you say, "Okay, we've had zero extinctions in 300,000 years. So the chance per year is zero," that's not correct. And this is a problem called the problem of zero failure data. How do you estimate the probability of an event happening if you've the chance that you're going to draw a black ball from the urn, if you've drawn 300,000 white balls so far and never a black ball.

And so there are answers to this, and basically these answers end up giving you things somewhere between, in this case, one in 300,000… Clearly, one in 300,000 would be an overestimate, because that's the chance that you would be saying if there had been a case. So it should be somewhere between zero and one in 300,000 in that case.

And there's some nice principled arguments for it being either just slightly below one in 300,000 — that's Laplace's law of succession — or that it should be about half of that, which is Laplace law of succession with a Jeffrey's prior. For the probability nerds out there. And so there's interesting ways that you can geek out on that and I've thought about this quite a lot, although I had to constrain that a little bit for the book.

Because it turns out that we can also think about other species where we don't have the problem of zero failure data. We can say, well, okay, what about other hominid species? How long do they tend to last? And what about if we look at other mammals, or just species in general?

And the answer that you tend to get, it varies a little bit, but is something like a million years is the typical time before going extinct. In which case you avoid this kind of strange anthropic effect that you can never witness your own extinction. And you can actually step outside that. So that's helpful.

And it's also helpful to note that when we try to do the estimate without looking at other species, we end up with a somewhat similar answer because our track record is not that different from a million years, in the scheme of things.

Julia: That seems so surprising though, because the situation we're in as humans is so different than the situation that animals are in. Isn't it coincidental that we would end up with a similar estimate?

Toby: Oh, not that coincidental, because what I'm imagining here is the chance of natural risk. So the chance that we would make it so far, and for a lot of that time period we were more like the other animals.

Julia: I guess that's true. Yeah. We don't really have an ability to control a new ice age, or... Well, we have more ability than animals do to adapt to the new conditions.

Toby: We do. And even if you just treat us as animals that have a very wide species range, where we're on all... well, we're on six of the seven in continents, in sustainable manners and so forth… Then it's much harder for such a species to go extinct. We also have a very large number of different types of foods that we eat instead of a species that relies on exactly one other species for food.

So there are a bunch of things that suggest that we should do better than average.

But that's if you... So there are a couple of ideas there, if you look at the natural risks. And you also need an assumption there, at least the way I do it, which is that the risk over time hasn't been increasing. If you can just assume that the risk per century hasn't been increasing over human history, then you can use some of these assumptions to bound the risk.

And I have a go with that in some of my writings and end up with bounds of roundabout one in 10,000 chance of natural risk wiping us out per century. Maybe one in 1000, but not higher than that. Because otherwise you just can't explain why we would've lasted as long as we have, and why other species would.

But there are a bunch of risks where you can't make this assumption that the risk hasn't gone up. And that includes a lot of risks that we think of as anthropogenic such as the risk of climate change, or the risk of nuclear, winter destroying us, or AI. But it also includes the risk of pandemics. Even the pandemics that we think of natural, because they're naturally arising or something. They're more likely to arise because we keep a lot of livestock, and also they spread faster across the world because of travel and we live in much denser societies and so on.

We also have things that go in the opposite direction, things that protect us. But it's not clear whether the protective things or the exacerbating things, which one wins out. And thus, we can't just rely upon this argument there.

So I tend to put the pandemics in with the anthropogenic risks. And for those, it's kind of anyone's guess. Then you're in the situation that you mentioned when you asked the question, of if you take, say, things since the industrial revolution, about 240 years ago, then you've just got two-and-a-half centuries where we survived. That's the track record. And so how many centuries more should you expect we survive? Well, it doesn't really help to bound it very much. You could have a risk of say 50% per century and get through two centuries and not be that surprised by it. I think the risk is probably smaller than that, but the point is that you can't really bound it very helpfully.

Julia: In your book, you talk about looking at near misses — where we didn't go extinct, but we got kind of close. Or at least it seems that way from looking at the details of what happened.

How would we incorporate near misses into the estimate?

Toby: I don't think there's just one way of doing it. But it is an example of how you can try to tackle this problem of not having the long run historical track record.

In general, with science, we work on an assumption that even if you are Bayesian about science, and you think that you have to start with a prior and then you update it based on evidence, that we should perform enough experiments before we announce our results. The prior is mostly washed out. Which means that for any reasonable choice of prior, you would be compelled by the evidence to pretty much the same answer about the probability. We like it when that happens. And we can often do that. And so we tend to hold that up as a kind of gold standard for announcing a result.

But when it comes to existential risk, we just often can't get to that position. And Carl Sagan has some nice line about, we don't have any more spare Earths that we could destroy in a laboratory in order to confirm the theory of nuclear winter before we face it. And thus it's unreasonable to use that standard on this question. And I think that that's right. Although it may also suggest that we shouldn't treat these claims in the same way we normally treat scientific estimates of probability. But scientists have a lot to add to the question because they know some of these good techniques, but it is a bit different.

And so how can we use near misses on this? Well —

Julia: Well, maybe we should give an example of a near miss before we go into the abstract theory. What would be an example of, like, a nuclear near miss?

Toby: Yeah. So we could use things like the Cuban Missile Crisis. That was a period of heightened risk in, I think, 1962, where the nuclear tensions were high.

And there were also a number of things that got very close to precipitating a nuclear war, such as a Soviet submarine that had a nuclear torpedo, and the captain ordered that it be used to destroy the fleet that was firing depth charges on it, but was overruled by the flotilla commander, Vasili Arkhipov, who was on the submarine.

And when we look at people like Robert McNamara, who was the minister of war at the time, and what he thought about this, and how Kennedy was thinking about it… It does seem like if this had have been fired, that the only retaliation plan that they had was full scale nuclear war.

So there's things like this that suggest we got really quite close. I think that there's a lot of fascinating and tempting ways to try to eke as much information as you can out of anecdotes like this. For example, Vasili Arkhipov was randomly on this particular submarine. There were four submarines in the flotilla, and there's a kind of one in four chance that he happened to be on this one, and therefore could overrule the captain and the political officer who both agreed to use the torpedo.

So maybe you could kind of suggest that there was therefore a three in four chance that this particular incident would've led to a nuclear war, and that we got lucky there. I don't think you can quite do that, although the reasons why are subtle… but you can see that you can explore these types of scenarios. You could try to get experts to rate the chances.

I think that one of the big challenges in terms of understanding probability is that when you've got probabilities about events in the future, and you don't know if they're going to happen, we kind of know... at least we've got some pretty good understanding of what probability means. We've got some sense in which it's, say, zero or one just based on whether it ends up happening or not.

We've also got some sense of, like, objective probabilities, like coin tosses and how they work. And then we've also got these kind of evidence-relative probabilities, like the Bayesian probabilities based on your degrees of belief.

But when you've got events in the past, and you're asked, what was the chance that the Cuban Missile Crisis would've turned into a nuclear war? It's difficult not to just say zero because we know it didn't happen.

And you can then try to set up some counterfactual where it's like, well, what if you were back then? And what would you say? And it's like, well, I guess I would say what the people at the time said, if I had the evidence that the people at the time had. Say, Kennedy said at the chances between one in three and one in two.

But then we also have these revelations that came afterwards. So I guess we could try to say, what would Kennedy back then have said if he'd happened to have known that these revelations…? But if you knew *all* the information, he would say zero. So you have to know some of the information and not other things.

And so, yeah, it's not totally clear that it's coherent... I think that there are answers here, and that there is some coherence, but it's very easy to start saying things that are incoherent about it. Which is a challenge.

But to give you a different example of a kind of near miss thing that you could do… you might think that prediction markets or insurance wouldn't be helpful when it comes to existential risk, because of this issue, that if you correctly predict that there will be an existential catastrophe that you can't collect on your bets. There's the kind of Tom Lehrer line that "Lloyds of London will be loaded when they go."

But there are some ways that you could do it based on a kind of near miss. So previously, we talked about near misses where it seemed like things were getting hot. The situation was breaking down and we're in a situation that got close to the brink, let's say. But you could also have a situation which is more like a literal near miss.

So suppose that you had a set of predictions on whether an asteroid will come within certain distances of the earth. Say, a big enough asteroid, a 10 kilometer across asteroid. Then you could imagine a series of distances — let's say, come within a thousand times the diameter of the earth, or within a hundred times the diameter of the earth, or within 10 times the diameter of the earth. And you could see what probabilities people are estimating for these different things, where they can actually collect on their bets. And then you could use those to extrapolate and work out the implicit probability that it would hit the earth. So there are ways that you can try to leverage the possibility of actual near misses in that sense to do this.

Julia: I guess, it's a little easier with an asteroid, just because it's a much more physically well-defined phenomenon that we can extrapolate from. It's not not messy like human behavior.

Toby: That's right. And you could also do a similar thing based on the size of the asteroid. You could say, what's the chance that a one kilometer asteroid comes within this distance. What's the chance that a three-kilometer, that a 10-kilometer and so on. And you can extrapolate up like that. And because we understand something about the size distribution of asteroids, we can make some progress.

We could even treat smaller asteroids, hitting the earth as near misses for larger asteroids. Where it's not that it nearly hit us. It's that we were hit by something smaller than the thing we were worried about.

And similarly, that could happen with things like pandemics, where maybe in some ways the current pandemic is a near miss or a warning shot or something like that for a bigger pandemic that could actually pose an existential threat.

Julia: Something that has always seemed weird to me about reasoning about near misses, just using my own intuition, as opposed to these more formalized methods you're talking about, is that if you had asked me to put a probability on each case, like taking nuclear near misses, for example, like the Cuban Missile Crisis, or the time that the Russians could have retaliated against us, if not for Arkhipov… I would've put a probability of maybe 20%, 25% on a lot of those.

But it starts to add up. If you take all of the near misses that I would've assigned 20%, 25% to, they add up enough that it starts to look really surprising that we *didn't* end up with a nuclear war. Which throws into question my ability to assign good probabilities to all of these near misses. You know what I mean?

Toby: Yeah, I think that's right. Once the probability that we could have escaped all of these near misses starts to get too low, let's say, lower than one in 10. Then there's a factor of 10 disconfirming — the evidence of the world disconfirming your estimates.

You could take that too far though. So Stephen Pinker, when he writes about existential risk, has appropriately lambasted some of the people who during the Cold War said, it's "almost certain" or it is a "certainty" that it will lead to a hot war. And indeed they were wrong.

But if the thing is that in order for this argument to go through — that it was an extremely important issue — there being a 10% chance, over a 50 year period, that it would lead to a hot war, would be ample for it to be the most important issue facing humanity at the time. And isn't disconfirmed.

But you're right that if your attempts to do this are producing very large numbers, then that does suggest that the attempts might be going wrong. And I think it's very easy to have them go wrong. I think it's an extremely difficult thing. And I think there's been very little written really about how to do this kind of retrospective prediction, trying to assign probabilities to past events where we actually know what happened.

Julia: Toby, I alluded to this topic earlier in our conversation, but as a lot of people have shifted their focus — like you — from near term issues like poverty and disease, to the project of trying to promote humanity's super long term flourishing… We've kind of shifted away from the ability to do rigorous evidence-based evaluations of things. Like the sort that Givewell does, where you can measure the cost effectiveness of donating a hundred dollars to the Against Malaria Foundation.

And a lot of people in the effective altruist community have said that even though this shift in focus may be good and warranted, an unfortunate side effect is that these discussions have become squishier, and less legible to outsiders. Like, it's less obvious to someone not in this community whether or not we are being truth-seeking, and getting the right answers, or making progress. Because we don't have a track record of impact.

And that also, it can make us more vulnerable to motivated reasoning. That because these topics are so abstract and we're not getting feedback from the world about whether we successfully prevented existential risk, it's easy to fool ourselves into thinking that we're making progress. Or into thinking that our particular pet cause is the thing that everyone should be devoting more time and funding to.

Do you agree with this kind of diagnosis? And if so, does it concern you at all?

Toby: Yeah, I agree with that. And it does concern me. It's definitely a downside of focusing on an area where the feedback loops or the ability to get really solid evidence is much less.

So in the case of say global health, there's a situation where we could use the ideas of effective altruism — both to select that cause out of all the causes we could work on, because we know something about how much it

can change people's lives and that there's an evidence base there… And also we could use these tools in order to select each intervention, and maybe even, which organization you want to fund among all the different ways of improving global health. And to get a big boost there, such that you could, I think reasonably think that you could do better than the typical person who's funding global health.

In the case of avoiding existential risk… I think that that's quite a lot harder. We use the tools of effective altruism to select this idea of that we should really strive to protect humanity's long term future and avoid these existential risks. And we get a lot of value from those tools in selecting that cause. But then within that area, it's harder to think that you can systematically do better than others who are trying to do the same thing.

So that's how I would see it. And so you get perhaps a bit more oomph from these tools in terms of the cause selection, but substantially less in terms of the intervention within that cause.

Julia: Yeah. You know, one way in which I think the "long-termist" framing of this project is unfortunate is that it makes people think "How could you possibly think that you can reason about how to influence society in 500 years, or 1000 years? That just seems so abstract and hopeless."

But that really isn't what long-termism is focused on. They're mostly just trying to prevent humanity from going extinct in the near term to make sure that we have a long term. It's still hard and abstract of course, but it seems much less crazy to think that that could be a tractable project, than trying to shape what society looks like in 1000 years.

Toby: Yeah, I think that that's right. And I think we're slowly beginning to understand these distinctions a bit better. So I would make tentatively a distinction between "long term thinking," where maybe that's the things like building cathedrals. Can we do long term planning where you have a complicated set of dependencies and reliances that span 100 years or more, or something along those lines? Can you do something like that versus long-termism?

Where with long-termism the idea is that the acts that we're choosing amongst, that we're particularly focused on the extremely long term consequences, because we know of acts like trying to avoid extinction, that would have a very reliable chain of consequences after a certain point.

So the idea is that with long-termism that the value that this act would have, such as lowering the risk of a bio weapons attack or something like that…

That the value that would have is in virtue of the long term future of humanity. It's kind of its value in terms of future generations, and allowing them to exist. But not that it involves a whole long complex set of causal links that you have to deal with across that time.

Instead, the causal story is, "What do we do now until this point where such a risk could strike?" So for example, maybe you're working on trying to deal with a risk that could strike in five years time, or 20 years time. And so you're really doing stuff on that time period. And then the connection — that if we're all dead, we don't have a flourishing future — is pretty obvious and should just be granted by the interlocutor in this conversation. So I think that's a distinction that I think is important.

I feel like I would like humanity to be better at long term thinking as well, this other side of things. But I have less to say on that when I think about it. And one of the things that makes me more confident when it comes to questions about existential risk is this pretty clear connection, that if we avoided a nuclear war in the next five years, then it's easy to see how that could be making people's lives better in 1,000 years time.

Julia: I'm actually curious how much you think it matters, whether people agree with you about the enormous value of preserving humanity's long term potential.

Like, I think it's a very common view that, "Look, I don't really care about whether humanity still exists in a thousand years. I care about whether people who do exist are happy. And I care about the people alive today. I care about their children, their grandchildren. But if we're talking a thousand years in the future… I don't know, I don't really care."

How much hinges on that, on whether people agree with you about that? For someone who just cares about preventing massive amounts of death and suffering, would the policy prescription be so different? I feel like a lot of the risks that you talk about in your book are things that it would make sense to care about and devote more funding to, even if you didn't care about humanity's long term future.

Toby: So I think Carl Shulman has written and said some good things on this. And he makes the case that the probability of some of these risks — say, bio risk, particularly with a bio weapon related pandemic — that the probability that that could destroy humanity… and then, just in terms of the lives that would be lost in that event, and in terms of the tractability of what we can do about it, suggests a cost-effectiveness of working on that, that even in terms of the presently existing lives would be cost effective

enough that the US government would normally fund it based on how the treasury determines such things.

And I think that that's a useful and important thing. And perhaps something I didn't stress enough in my book. But there are a bunch of things that the government should be doing that would meet its standard cost effectiveness criterion. But we don't have a movement of people based around them and so on.

And I think that based on the presently existing lives, there's enough of a reason to do something about it. But it's not as clear that there's a reason why this should be one of the chief priorities of our time or something like that, which is a level of concern that I think this justifies. That's another reason.

And some people say, not Carl in this case, that we should think a lot about co-benefits with something. So when someone talks about, say, saving the environment, that some altruistic people care about that… but maybe the politicians should focus on what would be good for you, in terms of that there'll be less pollution in your local stream or something like that. And so you could, again, maybe you are motivated by long term effects, but you should tell people just about the near term death and destruction.

I think that —

Julia: I don't love that.

Toby: Yeah. There's something to be said for that, I guess, but… it's not quite deceptive, but it's a little bit in that direction. It's not using the actual reason that motivates you instead, telling the person a reason that would motivate *them*, which is a little bit iffy.

And I also think that it's much more brittle, because it could be that the facts change, as to what is a risk. Or maybe it ends up saying you should prioritize a certain risk. Perhaps one of these risks that disproportionately has a chance of killing 90% of people, rather than 100%, and it poses, say, much less existential risk than some other thing, but it ends up getting prioritized according to this metric, because there's more chance that it kills 90% of us.

Or something like that. Where, I do worry about saying rather than here's the actual strong, moral case that we should accept, here's a different case that happens to be aligned with it.

Julia: It seems a bit myopic.

Toby: Yeah. So I'm generally a bit suspicious about that. Yeah, myopic is a great word for it actually.

I think that one of the things I'm trying to do when I talk about this is to actually change what's considered part of ethical discussion and to get future generations — in particular, the long term future of humanity — onto the table, as the type of thing that we should be caring about. And so approaches that note that it's currently not something we care about, and so don't talk about it, will never achieve that.

And I think that part of what we need here is actually a moral revolution. And we've seen this before. So if you go back to say 1950, the environment just really wasn't considered part of ethics or what it was to lead a good life and so on. And it's fascinating.

I've got a Richard Scarry book, *Busy Town* that I was reading to my daughter, that was a book that I loved as a kid. And I was shocked that there's some page that just… where they, I think, bulldoze a forest, and then they put this road through, and then they build hotdog stands and so on. And it was like a parody of a Joni Mitchell song or something. But there was just totally straight faced. There was no awareness that this was something that a generation later would be seen with horror and that it must be sarcastic or something.

But this radically changed in the '60s and '70s, to be something where I think that it's the largest part of moral education in schools in the west, is environmentalism, more so than even how to treat other people.

And it's considered a standard issue. There is a cabinet level position for it in I think all English speaking countries. It really has become part of what we conceptualize ethics and our ethical responsibilities as. And I think that future generations could go the same way, in thinking about humanity and this long term future, as a thing where we have special responsibilities. So I do think that even though that's not easy to make happen, there is quite a bit of reason for hope on that. And that if we could make that happen, I think a lot of the things would be a lot easier if people only actually cared about them.

Julia: Well, Toby, thank you so much for coming on Rationally Speaking, it's been fascinating.

Toby: Oh, it's been great to talk to you.

[musical interlude]

Julia: That was Toby Ord, philosopher at Oxford University, and I highly recommend his book, *The Precipice: Existential Risk and the Future of Humanity*.

That's all for this episode of Rationally Speaking. Join us next time for more explorations on the borderlands between reason and nonsense.