Rationally Speaking #169: Owen Cotton-Barratt on, “Thinking about humanity’s far future”

Julia: Welcome to Rationally Speaking, the podcast where we explore the borderlands between reason and nonsense. I'm your host, Julia Galef and with me is today's guest, Owen Cotton-Barratt.

Owen is a mathematician at the Future of Humanity Institute at the University of Oxford, where his work focuses on, I would describe it as theoretical questions involved in trying to improve the far future.

Owen, welcome to the show -- and does that seem like an accurate way to describe your research?

Owen: That's a bit grandiose sounding, but that is what the eventual aim is. It seems like the future could be a very big thing ahead of us and if we're able to do things to make it go a bit better rather than worse, then that could be quite important. It's hard to work that out but it's maybe important enough that it's worth further study.

Julia: What a delightfully affable, British way to describe this very important project, I love it.

Just so we can get clear on our terms, when we talk about the far future, what kind of time scale are we talking about? Is this 10 years, 100 years, 1000 years? What kind of order of magnitude?

Owen: I mean, let's go with 100 years up, but I really do mean up. Not just like capping out at millions of years -- millions of years is roughly how long humanity has been around for. Or even necessarily billions of years.

At the moment we are an apparently unique force in the universe, because we're not just… following existing patterns, but we're building things deliberately as we want them to be. At the moment we can do that over most of the surface of a single planet. Right now we don't have the technology to go out an do it over grander scales, but it is possible that in the future our descendants will develop this technology. And if they do, they could spread out through the universe and be around a long, long time.

Julia: Yeah. This is kind of a perfect time to be taping this episode, since Elon Musk just announced his grand plan for galactic domination, starting with Mars.

Owen: Right, he has an ambitious time scale on that. I don't know enough about the details of the technological side of things to know how feasible that is, but it's exciting that people are thinking about this.

Julia: Right, right, yeah, we'll stay solidly theoretical in this call and let Elon worry about the technology for now.

Are you focusing more on how to cause good things to happen – like, what can we do today to eliminate poverty in 100 years? Or are you more focused on preventing bad things from happening – like, how can we reduce the risk of humanity being wiped out by a pandemic in 100 years?

Or is that not the right way to carve up the space?

Owen: Yeah. I think that there is something in this division. Generally, we think of different ways the future might go, maybe we go down one path or another. And that's often triggered by some event, which could happen either for better or worse. Thinking in terms of what might shift that direction does seem a useful way of thinking about the future.

There's a bit of an asymmetry between events which look like they'll improve the world and ones which could do very large amounts of damage. Generally, people want the good ones and they don't want the bad ones. Okay, so what? Eventually, I think that if our civilization continues and continues to thrive, we will work out how to and then we will eliminate poverty. There're maybe some benefits of doing that a bit earlier than later, but as long as we're going in the right direction, then we may hope to achieve that eventually.

Whereas with some of the bad events, if we all get wiped out by a pandemic, then it isn't the case necessarily that if we avert that and we stop ourselves being wiped out by a pandemic, then it's just going to happen inevitably later. Because there isn't the pressure of people wanting to take actions to make sure that it does happen.

In fact, it's quite the reverse, people would want to take actions to try and ensure that it doesn't happen. As we grow as a civilization and we get better capabilities, we may be better placed eventually to make sure that these things never happen. It's just a period in the middle where potentially a large catastrophe ... We're not yet certain to be able to avoid it. If it did happen, it could totally throw us off a positive trajectory towards a valuable long-term future. This gives some reason to focus on avoiding very large scale bad outcomes.

Julia: Yeah, that is an interesting asymmetry. I had casually picked an example of bad outcome that was about wiping out humanity, as opposed to making the quality of life for humanity a lot worse -- do you think that that asymmetry applies even to bad outcomes that don't wipe out humanity entirely?

Owen: The distinction which seems particularly important to me here is about whether outcomes are just going to cause a bit of a better world or worse world in the short-term -- or whether their aftereffects could reverberate down the future ages, so that they shift us from one long-term trajectory onto another.

It's very clear how an event which could cause human extinction would do this. Events which just make people a bit unhappy in the short-term or a bit happy in the short-term, we have much less of a clear story how they could eventually do that.

Maybe there are some intermediate things, where if we had something which improved or damaged international cooperation in the future, we might think: That's the kind of thing which can affect not just whether we have good trade deals and people are prosperous in the short-term but also whether the world enters into a period of conflict and possibly we wipe ourselves out. In that kind of context, I don't see such an asymmetry between the good and the bad events.

Julia: Right, that makes sense. I will want to, at some point in this conversation, get into the relationship between causing good outcomes in the short-term and affecting our long-term future, because I guess it's not obvious to me that aiming for the long-term is the best way to help the long-term. As opposed to aiming for the short-term and counting on that to propagate forward. But I'll put a plug in that for now and focus first on our ability to make predictions at all about the future.

It is commonly pointed out that it is very difficult to make predictions, especially about the future. And I think that goes at least double for the far future. Phil Tetlock, who wrote *Superforecasting* and ran the Good Judgment Project, describes his attitude towards forecasting as a kind of "optimistic skepticism." Where he's skeptical about our baseline ability to make predictions about the future accurately and successfully. But at the same time he's optimistic about our ability to improve at least somewhat on that baseline, through careful work. Which is pretty close to my take as well.

But Tetlock's superforecasters were making prediction on the order of months, not decades or centuries. I'm wondering, first, where you fall on that optimism vs. skepticism breakdown. And then also how you think the much larger timescale that you're talking about should affect that balance.

Owen: Yeah. I have a mixed view on this. One thing which is important to note about the kinds of problems that Tetlock's superforecasters, and other forecasters that they're being compared against, were trying to tackle is that they were selected for being kind of hard on the scale of months. Because they wanted to have things where they could actually get differential accuracy between people.

I think that the difficulty of making predictions varies a lot according to the question. If you ask me who will be president of the United States be in 30 years time, I have no idea. If you ask me will the sun come up in 30 years time, I can be pretty confident in saying yes. There's a pretty broad spectrum in between these things.

Julia: That is an excellent point. What would you say is the least obvious prediction about the far future that you can be reasonably confident in? Does that question makes sense?

Owen: Yeah. When you say be confident, of course we have different levels of confidence.

Julia: I know, that's an infamously fuzzy phrase.

Owen: Here's one which I will take my stand behind. I'm not entirely confident in it but I think I am moderately confident in it, and maybe more so than a lot of people. That if, as a technological civilization, we stay broadly on a path of continuing to accumulate knowledge, and we don't get badly knocked off this course, then I think eventually we will spread out through most of our galaxy, possibly even other galaxies as well.

Some of my colleagues at the Future of Humanity Institute have done some empirical work looking into actually what would it take to get to other star systems and have been looking at what can we say about the limits of technology, what can we understand from what we know about physics, what we know about physical laws. Well, we think we can't travel faster than light speed; we can calculate the amount of energy that it might take to get to other systems; we can calculate the amount of energy that we can gather from the sun. And we can make comparisons between these.

Now, we don't know exactly how efficiently technology will let us convert these things. But we have spent a bit of time asking the question “Does there look like there are any blocks here? Are there any things which are just such fundamental obstructions that even with many centuries of improving technology, maybe even thousands of years, we just are never going to be able to overcome them?”

It looks like the answer is no. For this reason, that's my prediction. I'm not going to make any predictions about the timescale of achieving this, because I think that predicting dates for future technology is often quite a lot harder than just predicting what capabilities may eventually be developed.

Julia: I think I agree with you about humans leaving planet Earth in the long-term being at least plausible. And having it be plausible is all we need to affect the decisions that we make today. It seems to me we don't actually need to be able to pinpoint *when* we expect it to happen, or whether it's a 20% or 60% or 80% chance. Because as long as we think that it's a non-negligible probability, that makes the far future of humanity all the more important. Because there's so much more potential for humanity spreading throughout the galaxy as opposed to humanity staying on Earth.

I don't quite know how to put this -- but I guess if we thought that humanity was just going to stay on Earth, then the chance of our entire species getting wiped out is much higher and, therefore, there's less potential value that we could capture in the far future. Whereas if we’re going to leave Earth, there's a lot more at stake there and it's really important that humanity doesn't end in the next couple of centuries. Does that makes sense?

Owen: There's both less chance of making it through and having an extremely long future. Also, it's a smaller future and there're maybe less people living flourishing lives in a future where we're confined to just one planet, than one where we can actually spread out and use a much larger fraction of the resources of the universe.

There's a question about whether that matters. But I think if we are in a position to create a lot of people who actually just have very high quality lives, then most people would think that it's better to have more of them rather than fewer.

Can I just go back to another thing that was in that question that you are asking? Which is saying it doesn't matter that much about the timeline… I can think of situations where we might think, "Okay, that is actually relevant for our decisions today." If we thought that in the next five years we were going to be able to get off the planet and make it to other habitable planets out there in the universe, and start getting lots of people to build new homes, we might start being less concerned about climate change on our planet today.

Julia: No, that makes sense.

Owen: Yeah.

Julia: I think I just meant to say that we don't need to be confident in timelines in order to be able to act in *some* way on that prediction. I agree that the more confident we can be, the more decisions that will affect in the near-term.

Owen: That's right. It matters for understanding when we need to say collectively, "Okay, now is the time that we really need to start looking into space travel and doubling down on this."

Julia: I'm wondering in general how quantitative you think we can be when we're making predictions about the far future. Both in terms of how likely something is and also in terms of how big the effects will be.

When people ask me about why I'm interested in promoting rationality, or developing strategies to help us get better at improving our judgment and changing our minds and so on, I can tell a very plausible story about why that will be important for the future of humanity. Not to be too grandiose! And I believe that story.

But I would have a really hard time quantifying the magnitude of that effect, like the magnitude of how likely it is that my efforts or the efforts of everyone working on rationality will actually make a difference. And if so, how large the difference will be.

As a result, I would have a hard time comparing that intervention to other interventions I could be doing instead. Like, I don't know, working on cancer research or working to reduce poverty or reduce the risk of a pandemic. Can we quantify these things at all? If not, can we still make a comparison between different interventions?

Owen: Yeah, this is a tough one. I mean, you're right, it is hard to quantify anything like this. But ultimately, we have to make decisions between the different things.

We can try to do that in a non-quantified, just informal, “Well, I think about it and this one feels better.” Or we can try to have slightly more explicit comparisons via quantification.

When we think of quantification, often we think of measurements: I'm going to go out and collect this data, and now I have data to back up what I'm doing. If we're thinking of improving the rationality of the world as a mechanism of making there be better decisions about whatever the big important issues are, decades down the line, and changing the long-term future of humanity, there's no way we can do a randomized controlled trial on this. The numbers that we're using for quantification are, themselves, not going to be solidly founded.

But I think we can still get somewhere with this. We can try and break the big question of “How much does this work we're doing now help in the long-term?” into smaller pieces, that even if we can't measure exactly, are a bit closer to things we understand. And so our intuitions are a bit better trained on them. Then we can try to make our judgments about those and combine all of this at the end. That's going to get you numbers which are still pretty uncertain, but I think it can be helpful for having orders of magnitude.

Like “Actually, this whole direction does look a bit better than I had intuitively thought.” I give some weight to my informal intuitions, which were saying that they didn't think this looks so good. Then I go back and I look at the numbers I estimated for the different components, and I'm a bit skeptical about this, and then I shuffle all the numbers around, until I feel happier with it as a whole.

Julia: Yeah. It reminds me a little bit of these papers, I think they were by Kahneman and Tversky, but certainly someone in that field. Where they were looking at hiring decisions. And they were comparing our default gut hiring decision where we just use our intuition to decide who's the best candidate, and they compare that to a much more quantitative method. Where we come up with categories that we think are important to a good hire, and we rate people in those categories and then we give them the total score across those categories.

That more formalized method did better than the intuitive guessing, the default method. But what did even better than both of them was the method where you use the formal approach to sort of call your attention to things that you haven't been paying attention to before, you haven't been giving weight to. And then after you go through that whole exercise, you *then* go with your gut -- except now it's an informed gut instead of an uninformed gut.

Owen: I think that's exactly right. I see there as being two major mechanisms of benefit from using these formal models.

One of them is making sure that it does call our attention to the relevant factors and helping to avert scope insensitivity, where we don't pay attention to exactly which are the most important factors or how much a large difference in one factor matters, compared in another place.

The other mechanism which I think is quite important is it allows us to discuss these things better. If I have a disagreement with a colleague where we both just have intuitive judgments about a thing, then we can notice that we disagree and we can point to things that we think, and maybe we can uncover a lower level disagreement. But actually, it's pretty hard. Because we often have similar thoughts, and we're just shading things a little bit differently.

But if we have got an explicit breakdown and now we notice we disagree, then we can go into the explicit breakdowns and find out, actually, where is the disagreement coming from? Now we can have a productive conversation about the part that we're at least on the same page about.

Julia: Right. Also, potentially identify what information we’d need to get in order to settle the crux of the disagreement.

Owen: Exactly. Well, even if you don't have somebody with a disagreement, you can just keep track of the different variables that are going into your personal model and you can keep track on how large your error bars are, in an informal sense, just capturing how uncertain you are about each of the different variables. Then that can help you to know at the end which variables would you have to dive into and explore more to actually decrease your total uncertainty.

Julia: Right. Speaking of uncertainty about the future, I'm wondering if you think that that is an argument for discounting interventions aimed at the far future, as opposed to interventions aimed at helping people now.

We were talking earlier in this conversation about how there's so much at stake in the future of humanity and so it really matters that we not die out now. Which I buy. But, at the same time, it's just much less obvious that any one thing that we try to do now will actually have good effects for humanity in the future, as compared to saving the life of someone who exists today. Does uncertainty play a role at all when you're trying to compare different causes to each other?

Owen: It plays a massive role --

Julia: Sorry, that was a dumb question. Of course, it plays a role! I guess I just meant to ask how do you incorporate it, or how do you think about it.

Owen: Yeah. In some ways, the whole game is working out how to deal with uncertainty. This question that you're asking is specifically about: When we're thinking about helping people further in the future, does uncertainty in fact wash out the benefit that we might be able to help more people? I think there's something to that. It bears thinking about pretty carefully.

Even over short timescales we can see this. If I have a plan which is going to help somebody next year, as long as the steps in the plan are reasonably solid, I can be pretty confident that it will happen. Because a year is not a large timescale and so all of the supporting institutions that it might rely on are likely to still be around.

But if I have a plan to help people in 30 years, it's more likely that something fundamental would have shifted the ground underneath, so the plan is not going to be well-founded. Maybe I'm planning to help people with a health condition in 30 years, but it relies on this specific hospital. And the hospital goes bankrupt and closes down. Or maybe in the meantime we discover a cure for the condition they have, and now we've invested this money early in a long-term preventive measure, but the cure is quick, easy and effective, and we didn't need to bother with that. This is a low level background uncertainty which cuts through a lot of things, even when we have well-understood mechanisms for how we eventually want to help.

I do worry about this when I'm thinking about things which might affect the very long-term future. If we're looking at tens of thousands of years, or millions of years, or billions of years, then certainly that's an extremely long time and I just don't think we can have confident predictions about how things will proceed.

But there are cases where we can see a broad enough route to either benefit or harm, that we don't need to worry about tracking the specifics. If there's an event in 30 years' time which threatens to cause human extinction, I'm pretty confident that if it does cause human extinction, there won't be humans around 100,000 years later. I don't need to discount my uncertainty between that 30 years and 100,000 years. It's possible that if we avoid it and we don't have human extinction, then humans could still go extinct in the following 100,000 years; that may be likely. So there's some discounting that would occur on account of that.

There's another level of uncertainty which I think also bites here, which is: There isn't much that I can do where I have a very solid definite plan of how to reduce the chance of extinction from an event in 30 years' time. And this is a case where uncertainty is biting pretty hard. Now, it's only biting over a timescale of something like 30 years, so that's maybe manageable. But it is dealing with quite unprecedented events. We haven't had humanity go extinct before, we don't know exactly how to manage and deal with this. So that could lead to quite significant amounts of discounting, particularly if they're very precise plans which need things to go in a particular way in order to eventually be effective.

Julia: Do you think that there are any other theoretically sound reasons for discounting the future relative to the present, aside from uncertainty? I mean, I know in practice, people do, in fact, prioritize the present far more than the future. To what extent do you think that's rational versus irrational?

Owen: This is getting into large debates in economics about this question. There are a number of different reasons why people end up using discount rates.

One of the major ones is: If you're thinking about something like money or something you can easily convert into and out of money, then we think about the opportunity cost of investments. If I can stick the money in savings, then I'll earn some interest on that. And that means that I ought to care less about £10,000 or $10,000 in 10 years time than I care about it today.

That doesn't necessarily apply to things like people's lives.

Julia: Yeah. I was trying to figure out the parallel there, but it's not ... Yeah, I don't see how that would carry over.

Owen: It's a bit complicated. Another reason that, in some cases, people discount is they say, "Well, actually, we observe that people have preferences for the present over the future. So what we ought to be doing as a society to promote the social good is to promote the preferences of all the people around, to the extent that they prefer the present over the future…"

Julia: Well, that’s a little circular. I mean, part of what we're trying to do is figure out whether our current preferences make sense or not. So to use them as evidence that they make sense seems circular to me.

Owen: I am pretty sympathetic to your view here. I'm just trying to be fair and mention all the reasons that people might give for this.

Julia: I appreciate that!

Owen: There's a defensible position which says it shouldn't necessarily be our place to judge what people *ought* to want, and we should just act on the basis of what they *do* want.

I think it's a bit funny and I think it also has this artifact that it doesn't necessarily weight the preferences of the future people. If you think that future people matter comparatively much to present people, then that can be an issue for you.

Julia: Right. I also heard an interesting argument recently, that we should expect that if the world continues to become more developed and wealthier, then far future people will be better off than people today. So to make sacrifices today, like to forego economic growth today to reduce pollution that could affect the quality of life in the future – that's a little equivalent to forcing poor countries to make sacrifices to benefit rich countries. Except the countries here are actually generations, so they're separated in time and not in space. What do you think about that?

Owen: I think there's something to this. This is a pretty empirical question about how much richer future people will be than we are. But certainly we are much richer than people were 300 years ago. If it were just the case of moving money or moving physical resources between one time and another, then I think this argument is pretty clear-cut.

But nobody thinks, "Well, the present generation should make sacrifices for the future generation. Let's take something useful like cars, and we'll bury a load of cars, seal them up properly so they'll be able to get them out. Then in 300 years time, they can unbury the cars and they'll have all this wealth from us."

It's a little bit less clear in cases where there are trade-offs where we actually just have much more leverage over the future than they will.

There are differing degrees of this. In cases where it's just about wealth, actually we can do better than burying a car for them -- we can go and put the money in savings and to have more money for them in the future than we have now. But at least it's not clear whether that is a large enough growth in wealth to cancel out the fact that they'll be richer and will care less about any particular dollar.

In the case of something like, if we're actually talking about taking steps to reduce the catastrophe which could cause extinction, then this leverage argument becomes much stronger.

If we could prevent an extinction event in 30 years' time, then people in 35 years' time will benefit a tremendous amount from this. But they have no ability to spend their wealth to try and buy this good. It's only people in the next 30 years who have any leverage over that at all. It might be that people earlier in that 30-year-period have even more leverage than people later. For instance, if there's a long string of actions that you need to take, or if you just need enough time to gather attention to the issue.

Julia: Right, good point. I had hinted earlier in the conversation about wanting to delve into what kind of approaches are most likely to have a good effect on the far future. Because, as I said, it wasn't obvious to me that aiming to affect the far future is better than just aiming to do good things now which will sort of indirectly improve our far future.

You've talked about looking back over the history of our civilization and how we've gotten more well-off and better developed and so on, and I think we've also had a fair amount of moral progress. But it seems to me that all of that progress was not the result of any one person or entity trying to make the far future better. It was just the result of individuals doing things that seemed good very locally. Like, this invention will improve the lives of current workers. Or this invention will make me rich. Or I want to cure this particular disease.

But those interventions had a snowball effect. The more you reduce disease, the more you reduce poverty, and make it easier for people to go to school and get educations. And that increases their ability to develop new technology that, in turn, helps us reduce more diseases, et cetera, et cetera.

It seems like we do have some good track record for doing near-term good things that then have these flow-through effects that end up powerfully shaping the future, even though we weren't aiming for that. And as far as I can see, we don't really have any examples in this other category of *aiming* to affect the far future.

So I mean, how is that not an argument for just continuing to do what we've been doing successfully so far?

Owen: I agree, we've done extremely well out of people often just following this local "what looks good" [approach] and pursuing that, and then producing good effects. I certainly don't think -- and I don't think anyone thinks -- that we should stop doing that.

Quite a lot of our society is set up around encouraging these things. We provide incentives for people to do things which help other people, and people generally feel good about themselves when they do that. We have a lot of resources going into this and that's fantastic.

But we don't have many people explicitly thinking about whether there's anything we can do to help the long-term future, or even just maybe among these shorter term things where we're helping people, whether some of them would be more useful than others from a longer term perspective. That may mean that there are good opportunities here to actually help in quite an effective way, that nobody is taking because nobody is giving attention to this.

Julia: We have, as promised, been speaking very theoretically. Before we close, I wanted to give you the opportunity to talk in a little bit more specifics about whether there are any interventions that you feel have a fighting chance of impacting humanity's future in a positive way.

Owen: I think that there are a lot of different things we could do, which may have some positive benefits on the future. Which include just generally looking around for things to help the world go well in the short-term.

The ones which I feel most optimistic about are things which target possible ways that things could go wrong in the coming decades. This pandemic scenario, hopefully is unlikely but I think it is a real issue.

And the advent of artificial intelligence -- again, it's hard to predict timelines, but maybe this is a thing which could come in the next few decades. If it does, if we get really powerful artificial intelligence, then that might mean that our world looks radically different from how it does today. And that could be a great improvement. But people have also outlined scenarios where that could lead to worse outcomes. And trying to make sure that we are placed to get good outcomes from that seems valuable to me.

Then other interventions which try to position ourselves better for facing future challenges. So this could be trying to do research into questions which look like they're going to be particularly important for understanding what's coming in the long-term future, and what the dimensions are. It could be trying to improve cooperation at an international level. It could be trying to improve rationality, as you were talking about earlier. We can draw a more direct route from improving rationality, particularly among people who may go on to be key decisionmakers in coming decades, to making sure that the world goes in a good direction.

Then we can form some other goals which look like they would just be fantastic in the short run, like curing cancer. There would be a whole lot of good knock-on effects from that, but they're less direct and more diffuse.

Julia: Yeah, that's my sense as well. Although it is really hard, emotionally and strategically, to make the case for long-term indirect things being more important than saving lives from a disease like cancer now.

Owen: Actually, part of the thing which strengthens the case for it is that it *is* emotionally hard. There are lots of people who go out and do things to help people. There's a pull towards helping where we can see this direct way, of how “This is definitely going to help, and I understand how, and I lost a relative to cancer and I don't want that to happen again.”

That means that we are collectively already investing in opportunities like this. I definitely don't think we should move all our resources away from things like this.

But if we are just controlling a few marginal resources, and just taking up a couple of extra opportunities, then going for things where we can see a strong, reasoned argument that they could be effective, but there isn't so much emotional pull, we may find things where the low-hanging fruits are yet unpicked.

Julia: That's a cool and kind of counter-intuitive argument for doing counter-intuitive things. Which makes me very happy.

Owen: You have to maintain a bit of skepticism about where you're applying it, but I think there's something to it.

Julia: Yeah. The argument on the margin seems to be such a stumbling block when I have conversations with people about the most valuable ways to help the world. Because whenever I try to make the argument on the margin for research in trying to prevent existential risks, or investing in infrastructure to make us better at handling risks when they come up, et cetera, people often jump to, "Well, but if we invested *everything* in that, then what are we going to do, just let people die from poverty and disease now and not try to save them at all?"

Of course, we're already investing tons of resources into those causes, and I'm talking about a small change on the margin.

Owen: That's right. I think that this is partly because the idea of margins isn't one which is quite in the public consciousness.

Julia: Yeah. I keep looking for another way to phrase it and I haven't found a good one yet.

Owen: I'm not sure it's even just about the words, it's about the ideas. Often, people don't have a distinct concept of absolute priority -- how much ideally would we devote to this problem collectively -- and then marginal priority, given all the decisions other people are already making, what's particularly valuable to do now.

Julia: Right. We were ending on a hopeful note and now we're on a frustrating note.

Owen: I think this is an opportunity. There's an opportunity here for people to learn a new concept which you and I both think is important, and spread this out so it becomes common social knowledge, and then we'll get better decisions as a result.

Julia: Excellent. I like that, perfect. We'll end here, before I find a way to accidentally make things frustrating and depressing again.

Owen: Thanks, Julia.

Julia: Cool. We'll move on now to the Rationally Speaking Pick.

[interlude]

Welcome back. Every episode, we invite our guest to introduce the Rationally Speaking Pick of the episode, that's a book or an article or website that has influenced our guest's thinking in some interesting way. Owen, what's your pick for today's episode?

Owen: My pick is *Probing the Improbable*. It's a paper which looks at the question of what should we do when we have a model which tells us that an event is extremely low probability.

It raises this interesting point, that if our model says that the probability of an event is one in a billion a year, well then, straight off, we're pretty confident this event isn't going to happen. But if it did happen, we would think not, “we just got extremely unlucky” but, “I guess our model was wrong.”

And we should actually think about that in advance, we don't have to wait for that to happen. Beforehand, we should design some probability to the model that we're using being wrong, and factor that into our assessment of likelihood.

Julia: Interesting. Do you think that the way that people typically think about these very low probability events is they never step outside the model, they just go with that estimate?

Owen: That's even one level of sophistication up. I think that often at an intuitive level, thinking of very low probability events, people think, "Well, it's never happened," and then they just don't think about it anymore.

Then I think the next level up is they build the model as a way to understand what's going on there. We talked about the advantages of models a bit earlier in the discussion. Then I think, yes, they often will say, "Well, I've put my knowledge into this model so that tells me what I should think," and they don't notice the ways that it might go wrong.

Julia: Right. Yeah, excellent. Maybe I should do a whole episode on the promise and perils of models or something, because that's a really interesting thread that just keeps coming up in these discussions. Cool. We'll link to *Probing the Improbable*, was the name of your article?

Owen: Yup. It has a subtitle as well but I forget what it is.

Julia: That's all right. Well, we'll put that up on the website and we'll link to the Future of Humanity Institute page as well. Owen, thank you so much for joining us. This is a really fascinating conversation.

Owen: Thanks, Julia. It was a great conversation.

Julia: This concludes another episode of Rationally Speaking. Join us next time for more explorations on the borderlands between reason and nonsense.