Rationally Speaking #181: Will MacAskill on “Moral Uncertainty”

Julia Galef: Welcome to Rationally Speaking, the podcast where we explore the borderlands between reason and nonsense. I'm your host, Julia Galef, and with me today is my good friend and today's guest, Will MacAskill.

Will is probably most famous worldwide for helping create and popularize the effective altruism movement, which I've talked about in several past episodes of the podcast. He co-founded and is the CEO of the Center for Effective Altruism, and is the author of the book *Doing Good Better*.

But today we're not going to be talking about effective altruism, at least not primarily. We're going to be talking about Will's work in philosophy. Oh, I didn't mention in my list of Will's achievements that he's also ... When he became a tenured professor of philosophy at Oxford, he was the youngest person in the world to be a tenured professor of philosophy.

We're going to be talking about Will's main research focus in philosophy, which is moral uncertainty. This is a topic that's on my list, at least, of philosophical questions that I think are meaningful, and real, and important to actual real-life decision-making -- and that are unresolved, but that I feel like we might potentially be able to make progress on. This is not a long list of topics, but moral uncertainty is on it.

Will, welcome to the show.

Will MacAskill: Thanks so much for having me on.

Julia Galef: I can't believe I haven't had you on sooner. You've been a desired guest from very early on. As I was telling you earlier, I kept having other people on to talk about effective altruism, and then I'd be like, "Oh, well, I can't invite Will now. I have to wait a little while, because I just had Ben Todd, or Peter Singer on." This episode has been a long time coming.

Will MacAskill: Well, I'm thrilled to be on. It's probably one of the very few podcasts in which we can have a conversation about moral uncertainty…

Julia Galef: That may be true! Will, do you want to just explain what -- do you prefer moral uncertainty or normative uncertainty?

Will MacAskill: I think let's stick with moral uncertainty.

Julia Galef: Okay.

Will MacAskill: Normative uncertainty would cover a wider range, including decision theoretic uncertainty, uncertainty about rationality, epistemological uncertainty, but they get even more tricky and tenacious as philosophical issues than moral uncertainty, which also has the clearest practical implications.

Julia Galef: Okay. Great. Moral uncertainty, what is it? Why is this a needed concept?

Will MacAskill: In order to explain moral uncertainty, it's easiest to start with thinking about empirical uncertainty, or uncertainty about matters of fact. In ordinary life, when we're deciding what we ought to do, we don't just think, "Oh, this is what I believe." We tend to think, "These are the variety of things that might occur." You -- perhaps not consciously -- but you think, "These are more likely than not. This is how high-stakes things would be if this were to be the case or not," and then you take an action that is in light of all of that uncertainty that you have.

For example, if you are speeding around a blind corner. Normally, you would think, "Yeah. That's a wrong thing to do. That's an immoral action." The reason we think it's wrong is not because if you speed around a blind corner, you're probably going to hit someone. Instead, we think it's wrong because there's some chance that you'll hit someone, and if you do, then it's very bad. It's a very bad action or bad outcome.

Decision theorists have formalized this using a notion called expected utility theory. Where, in order to make a decision, what you should do is look at all of the possible outcomes. In this case, there would just be two. One is that you speed around the corner and just get to your destination faster. The second is that you speed around the corner and hit someone.

Julia Galef: Well, there are two for the purposes of a simple illustration of the concept, right?

Will MacAskill: Yeah. That's right.

Julia Galef: Okay.

Will MacAskill: Yeah, of course, there's loads of other things you could do. You could jump out the car…

Julia Galef: You could accidentally hit the next Hitler, and then it would be really good.

Will MacAskill: Yeah, of course. Yeah. That's right. Yeah.

Julia Galef: Go on. Sorry.

Will MacAskill: Let's keep the simple example. There's two possible outcomes: Speed and get to the destination faster; hit someone and, let's say, kill them. Then there's values assigned to those outcomes as well. If you speed but get to your destination faster, that's a mildly good thing. You've saved a little bit of time traveling. Let's just, for the purpose of the example, give that a number. Say that's plus one or something, meaning just it's a little bit good.

Then if you speed around the corner and hit someone, well, that's really bad, so maybe that's minus a million. It's a million times worse than getting to your destination a little bit faster is good.

Then the next step, once you've assigned the values of different possible outcomes, is to think, "What's the likelihood of each of those outcomes?" Let's say there's a 99% chance that you wouldn't hit anyone, but the 1% chance that you would.

The idea of maximizing expected value, or maximizing expected utility, is you take the sum of the products of the outcomes and the probabilities of those outcomes, the value of the outcomes and the probabilities of the outcomes. For one action, you speed. That's a 99% chance of plus one, but it's also a 1% chance of minus a million. Let's just say zero is to just go at normal speed.

Now, if you take 99% times plus one, plus 1%, times minus a million, that number is clearly less than zero, and that means that not [speeding] has higher expected utility, expected value, and so that's the action that would be recommended.

Julia Galef: Right. What's interesting is that a lot of people, maybe most people, sort of bristle at the idea that you should be even implicitly doing this calculation of whether the payoff to you of getting somewhere a little faster is worth risking someone's life. They'll say that no amount of time saved for you is worth risking someone's life.

But in fact, they are making that trade-off. They are making that implicit calculation every time they drive, or every time they --

Will MacAskill: That's right. We do this all the time, even ...

Julia Galef: Not just speed. Even if you drive, you have a chance of hitting someone, even if you're not speeding.

Will MacAskill: Even if you drive at 20 miles per hour like a granny going to the local shops, you're still saying, "There's some chance I'm going to hit someone."

Julia Galef: Right.

Will MacAskill: The idea that sometimes people say, "Oh, life is of infinite value, so you shouldn't do anything," just doesn't make any sense.

Julia Galef: Right. Right. So would it be fair to say this is a utilitarian ... You're making this calculation under the utilitarian framework?

Will MacAskill: I think, importantly, not. I think absolutely anyone should be thinking in terms of expected utility. You don't need to be a utilitarian.

Julia Galef: Oh. Maybe I'm using the world very broadly or something.

Will MacAskill: Yeah. As I understand, utilitarianism, it says ...

Julia Galef: You are the philosopher of moral philosophy! I don't think you need to caveat that. Go ahead.

Will MacAskill: Perhaps that was me being very British!

Utilitarianism is the view that you are always to maximize the sum total of well-being. There's three ways you can depart from that.

One is that you could value things like art or the natural environment, things that aren't related to people or people's wellbeing. The second is you could think there are side-constraints, so you ought not to kill one person in order to save even five people or a hundred people.

Then the third is that perhaps there are some things it's permissible for you to do even though there's some other course of action that would allow you to do even more good. So if you're just above the poverty line, perhaps you're not required to give even more than you already have done in order to save people who are even poorer than you. You have some realm of the merely permissible.

Julia Galef: Got it. We've started to get into different moral theories diverging from pure, total utilitarianism, but I derailed you a little bit in venting about people's unwillingness to acknowledge that they are in fact making trade-offs.

You were explaining ... You were starting to talk about empirical uncertainty, and then distinguish moral uncertainty from that.

Will MacAskill: That's right. Yeah. Decision-making under empirical uncertainty is extremely well-studied, going back all the way to the '50s, '40s, even going back to Blaise Pascal. We have a very good formal understanding of decision-making under empirical uncertainty.

We haven't solved all of the problems. I know you talked about Newcomb's Problem in a previous episode. There's still some underlying philosophical issues, but in general, it's pretty well-accepted that the rational thing to do under empirical uncertainty is something like maximizing expected value.

Now the question is, if that's the case for empirical uncertainty, uncertainty about what's going to happen, is it also the case for uncertainty about your values, or uncertainty about what's morally the case?

Now consider another example. Suppose you're at a restaurant. You're deciding you want to order for dinner. You have one of two options. You can

order the steak or you can order a vegetarian risotto. You think probably animals don't really have moral status. It's just ...

Julia Galef: Meaning that there's no point in worrying about their suffering or their death.

Will MacAskill: Exactly, yeah. We shouldn't worry about harming or killing animals, in just the same way as we shouldn't worry about destroying rocks or something that are non-sentient.

Let's say that your preferred moral view. Maybe you've even thought about it a lot, and feel quite committed to this.

However, it would seem quite overconfident if you were to be absolutely certain in this, 100% certain. In fact, if you were 100% certain, that would mean that there's no conditions under which you'd change your mind.

In fact, there are all these animal welfare advocates, and vegetarians, and they have what seem like fairly compelling arguments, so it seems like you should put at least some degree of confidence that they have the right view and you don't. Let's say you just put 10% confidence in that view.

Well, now, this looks kind of similar to the speeding around a blind corner example. Where if you choose the steak, you're doing something that, by your own rights, is probably morally permissible, is probably fine. Let's say it's even a slight benefit, because you enjoy eating meat.

But you're taking a risk, because there's a 10% chance, given what you believe, that actually animals do have moral status, and it's severely wrong to be incarcerating them in factory farms, and then buying and eating their flesh.

Julia Galef: This is a bit of a confusing example, because it's so tied up with empirical uncertainty. Because I can have uncertainty about whether a chicken is conscious, or has the capacity to suffer, and that's an empirical question that might determine my values.

What if it were ... Feel free to reject this, but what if the situation were, let's say I know how conscious chickens are, and then the question is just, is it wrong? Is it morally wrong to kill, to take the life of an animal with that degree of consciousness or something?

Will MacAskill: Yeah. Yeah. That's right. In all of these examples, we can imagine that we just know all of the empirical facts and matters of fact.

Julia Galef: Right.

Will MacAskill: Yeah. If you want to make it really visceral, the chicken is right there in front of us, and we can just wring that chicken's neck. I don't know whether it's bad to kill the chicken or not. I don't know whether that's a wrong thing to do, even though I know all the facts about chicken consciousness and so on. There's just this further moral fact about given all those empirical facts, is it wrong to act in a certain way with respect to that animal?

Julia Galef: Right. I was complaining a few minutes ago about people's unwillingness to acknowledge empirical uncertainty, at least consciously acknowledge it, but I think with moral uncertainty, it's even more so. it's not just that they won't acknowledge that there's a chance that their moral theory might be wrong. I think they act as if they're positive their moral theory is right.

Will MacAskill: That's exactly right. That's why, when I had this idea seven years ago now, that, "Why don't we take moral uncertainty into account in the same way we take empirical uncertainty into account?" the reason it was so interesting for me was precisely because so few people seem to think in this way. Instead, people have what I call the football supporter approach to moral philosophy, where they ...

Julia Galef: The football supporter?

Will MacAskill: Yeah.

Julia Galef: Do you mean like the football fan?

Will MacAskill: Yeah, as in football fan. Yeah.

Julia Galef: It's so adorable that your term for football fan is a football supporter. I'm sorry. Go on.

Will MacAskill: Yeah. Also, in my head, I'm thinking of soccer. I'm not thinking of NFL, as well.

Julia Galef: Okay. I don't think that makes it better, though.

Julia Galef: Or does it? Do Brits call ...

Will MacAskill: I think so. I don't feel like that's a weird thing that I was saying, but I'm not…

Julia Galef: Really? Oh. I don't think between the two of us, we have enough sports knowledge or familiarity to have any idea about the answer to this question. Wait, so Brits would say ...

Will MacAskill: Yeah. Football team supporter, at least. Yeah.

Julia Galef: You would say, "Oh, there's a stadium full of cheering football supporters?"

Will MacAskill: Football team supporters, perhaps. Yeah.

Julia Galef: Huh. Okay. All right. I'll take your word for it.

Will MacAskill: Okay. The football fan model, going to revise now.

Julia Galef: Okay.

Will MacAskill: The football fan model, where you have a team that you usually support, and you just identify with that team. That's the way people think about moral theories often. They will say, "Oh, well, I'm a utilitarian," in the same way as they might say, "Yeah, I'm a Rangers fan," or, "I'm an Arsenal fan." It's like they have an allegiance to a particular moral view, but that's very different from having a certainty of belief in it.

Julia Galef: Right.

Will MacAskill: That means that when people think about practical moral issues, they tend to say, "Well, I'm a utilitarian, therefore I believe that you ought to kill one person to save five," or they will say, "Well, I'm some sort of natural law theorist, and therefore, I don't think that animals have any relevant moral status. It's not wrong to kill them, therefore, it's okay for me to eat the meat."

That's very different from how we ought to reason empirically, because they're not taking into account the possibility that they might be wrong about their moral value.

Julia Galef: Let's talk about what it means to be wrong about our moral values. If I were to express my empirical uncertainty, that can be cashed out in something meaningful, like ... There's different ways to do it, but I might say, "Well, I'm 50% confident this coin will come up heads." By that, I might mean that out of all the past coin flips of which I'm aware of, roughly 50% of them were heads.

Or I might say something like, "I put a 25% probability on Trump being impeached before his term is up." By that, I might mean that I would bet at those odds. I'd be willing to pay you $30 if he gets impeached if you pay me $10 if he doesn't. Is that right? Anyway, let's assume that's right.

Those statements of uncertainty can be cashed out in concrete meaning, but what would be the equivalent for moral uncertainty?

Will MacAskill: I think one equivalent for moral uncertainty is what you might think at the end of a very, very long period of reflection. Supposing you could have 100,000 years and your brain was souped up in all of these ways. You just didn't have any empirical failings --

Julia Galef: Cognitive biases?

Will MacAskill: Yeah. You were able to mess around with your cognitive architecture to remove your biases. You had an IQ of 1,000. You had all sorts of ways in which your brain was souped up. You could discuss indefinitely with friends. You didn't have any pressing needs that might distort your thinking as well, so you're in some sense a very idealized version of your own self.

What would you believe? What would you believe at the end of that process?

Julia Galef: It seems to me debatable whether that would still be *me* in any relevant sense.

But I'm also not confident that that matters. Maybe the question I really care about is not what I would value, but what would a person I would consider an ideal person, an ideal moral reasoner, what would that person value? Even if that person is so different from my current self that they wouldn't really be me. Does that seem right?

Will MacAskill: Yeah. Yeah. This idea is, yeah, exactly, called, yeah, the ideal reasoner view.

Julia Galef: Got it. Moral uncertainty would be my uncertainty about what this idealized version of myself would end up valuing, after all of these improvements to reasoning and to knowledge -- including having all the empirical facts?

Will MacAskill: Yeah. I think ...

Julia Galef: Or maybe just having the empirical facts I have right now ...

Will MacAskill: Yeah, I think that's right. For the purpose of these examples, because moral uncertainty is so hard alone, just as I did in my thesis, we can just assume you know all the relevant empirical facts, because we don't want to get into issues with that at the same time.

Julia Galef: Right.

Will MacAskill: That's one way of operationalizing it. It does depend on your metaethics though. It might be that you have a very simple subjectivism, according to which what you ought to do is just what you want now. In which case, moral uncertainty is just uncertainty about what you want now. I don't think that's a very plausible meta ethical view, but it would have an implication about what it means to be morally uncertain.

You might have a very robust form of moral realism, according to which your uncertainty really is just uncertainty about what the world is like, such that it's possible that, even after this very long period of reflection, you would still be radically wrong.

Julia Galef: Got it. Okay.

Will MacAskill: There, obviously, you can't explain it in frequentist terms, like flipping a coin, where on average a fair coin will come up heads 50% of the time. But nor can you do that for something like, "What's the thousandth digit of pi?" Where naturally you would have 10% credence in each of the digits, but you can't explain that, at least not obviously, in frequentist terms, because either it is the number two or it isn't, and it's just that I don't know.

Julia Galef: Right. Okay. So what's the answer? How do you ... Well, maybe the right thing to ask first is -- we've been talking about simple, binary cases, like, "Is it wrong to kill the chicken or isn't it?" How many different moral theories could one take into account in an uncertainty calculation?

Will MacAskill: Yeah. I think ultimately the answer is, "A lot." In the first few years of research in this topic, people focused a lot on very simplified or idealized cases. One is this meat-eating case, where the argument was that because of the asymmetry of the stakes, that it's much worse to eat meat if meat is wrong than it is to not eat meat if eating meat is permissible. You ought to in general refrain from eating meat.

Julia Galef: I can see someone complaining that that's kind of like Pascal's Wager. That, "Well, maybe God is unlikely, but if he does exist, then we're way better off believing in him, so might as well play it safe."

Will MacAskill: Yeah.

Julia Galef: In the animal case, that seems less absurd, but I still am uneasy about establishing a system of moral decision-making based on that principle.

Will MacAskill: People often bring up Pascal's Wager when I mention this argument, but I think the analogy isn't great. I think it's more like, "Should you wear a seatbelt when you go driving?" Most people think, "Yes." They don't say, "Whoa. You're just saying ... There's a one in a thousand chance that I get into a car crash, or one in a hundred thousand chance, and I don't like wearing a seatbelt."

Julia Galef: The difference, the reason that it could be a good analogy I think is that it's so hard to pin down what the actual probability is of some very burdensome moral theory, that, if you follow this policy of saying, "Well, the consequences if that moral theory were true are so great that I should just follow it," that, because it's so subjective, there are going to end up being a ton of actually very bad or very implausible moral theories that you're going to end up kowtowing to.

Will MacAskill: Yeah. I think there is a question about, "Well, what credences, what degree of belief ought you to have in these different moral views?" I think at least as far as the idealized examples go, it seems like, "Well, what degree of belief ought you to have that it's wrong to eat animals?" -- again, if you imagine you had a thousand years to think about these issues, what's the chance that you

would come to believe this? It seems strange if you were putting it less than 1%.

Julia Galef: Yeah. Okay. In the animals case, it's not ...

Will MacAskill: Yeah.

Julia Galef: It's well within the range of plausibility that we should consider animals to have moral status, and therefore, it's not like Pascal's Wager.

Will MacAskill: Yeah. Although, I think I'll come back in a way where there might be an analogy in just a moment.

Julia Galef: Okay.

Will MacAskill: Before then, can think about certain other idealized cases as well. Animals is one. A second is our obligations to the very poor, to distant strangers. Peter Singer, for a long time, has argued that it's just as wrong to refrain from donating $3,500 ...

Julia Galef: Which is the amount needed to save a life with an anti-malarial bed net?

Will MacAskill: That's right, or to do the amount of good equivalent to saving a life. I think it's $7,500 to strictly-speaking save a life.

Julia Galef: Ah, okay. Cost disease!

Will MacAskill: Cost disease.

Julia Galef: Sorry. Go on.

Will MacAskill: It's just as bad to do that as it is to walk past a child drowning in front of you, which we would see as clearly, very gravely, morally wrong.

The question then is, well, supposing you don't believe Peter Singer's argument. You think there is some morally-relevant difference, that distance or psychological salience really does make a morally-relevant difference.

The question again is, "Well, how bad is it to walk past a child drowning in front of you?" Secondly, "Just how unlikely do you think that is?" Again ...

Julia Galef: How unlikely do you think it is that he's correct that those are morally equivalent?

Will MacAskill: That Peter Singer is correct that those are morally equivalent.

Julia Galef: Right.

Will MacAskill: Again, it seems like it would be very overconfident, given the difficulty of ethics, given how much change there is in terms of generations, as subsequent generations [experience] moral change, given the obvious biases we potentially have in this area, it would be overconfident, again, to have lower than say 10% degree of belief that Peter is actually correct about this.

In which case, again, it's like, "Okay. I'm going to walk past this shallow pond in order to protect my nice suit, because I'm not sure it's someone out there. It might be. It might not be." But you think it's still worth the cost to investigate, because of the chance that you're going to save a child.

Julia Galef: Right.

Will MacAskill: Some people even go further, and argue that there's just not a distinction between acts and omissions at all, and therefore, that failing ...

Julia Galef: Between acting and failing to act?

Will MacAskill: Between acting and failing to act.

Therefore, that failing to save the child in Sub-Saharan Africa, who you could save with $7,500, is actually just as morally wrong as going over there and killing someone.

That's a stronger view. Maybe you have somewhat lower credence in that. But again, that means you've got potential to be doing something very seriously morally wrong.

This isn't all idle speculation. Think back hundreds of years. Think of the way that people treated women, or homosexuals, or people of other races. Think of people who kept slaves. Aristotle spent his entire life trying to think about ethics and how to live a good life, and he didn't even ... He never thought that slavery might be immoral, and so he kept slaves, despite his best intentions.

We should actually be very open to the idea that we're doing things that are perhaps very seriously morally wrong, if we don't want to be engaged in those practices, then we should be thinking, "What are the ways that we should be morally wrong?" Then ...

Julia Galef: What are the ways that we *could* be morally wrong?

Will MacAskill: Sorry. What are the ways we could be acting that are gravely morally wrong, in the way that we look at slavery now?

There's one final example that I think has caused, as a psychological matter, caused a lot of animosity towards the idea. Which is the application of this to abortion, where some people have made the similar style of argument,

which is that ... Let's consider, say, a 20-week abortion. You might think, "Yeah, I think probably the case that a fetus at 20 weeks is not a person." Therefore, it's permissible for someone to kill that fetus on grounds of not wanting to have a child at that particular time in life.

We'll put aside cases of rape, or where it seriously compromised the health of the mother or the child, where it's just a case where it's not a great time for the parents to have that child.

You might think, "Well, we shouldn't be that confident." There are lots of views of personhood. There's that it starts much earlier than 20 weeks, in which case ... If it's the case that those views of personhood are correct, well, then you're doing something as wrong as if you were killing a newborn baby. Which everyone -- well, with some exceptions, such as Peter Singer, but almost everyone -- agrees is a very serious moral wrong.

If you think, well, even if it's only one in a ten chance that you're killing a newborn baby, that risk enough would be sufficient to mean you shouldn't do that, even if it was at significant cost to yourself.

Julia Galef: Mm-hmm. Yeah. You could even push it farther back to the point where the fetus is really just a cluster of cells, and there's really no empirical uncertainty about whether it's at all conscious, or whether it can feel pain -- but there's still significant disagreement, moral disagreement, over whether it has moral status, whether it's wrong to kill that cluster of cells, basically.

Will MacAskill: That's right.

Julia Galef: Although I would be less confident in those moral theories than I am at 20 weeks.

Will MacAskill: That's right. I think it would of course be reasonable to be less confident.

Julia Galef: I think so.

Will MacAskill: Yeah. There is this Catholic tradition that thinks that personhood begins at conception. This takes me to the way in which I do think this is similar to Pascal's Wager, but it's not for the reasons you might think. In my view, the reason why, when you hear Pascal's Wager, you think, "Something's up here," is that it's a very ... It's kind of like trying to do surgery with a hydrogen bomb or something. This very powerful, new intellectual weapon that you've taken, and then you're trying to just do very one small thing with it.

Julia Galef: Uh-huh…

Will MacAskill: In Pascal's case, why does Pascal's Wager not immediately work? He says, "Okay. You ought to go to heaven. You ought to go to church because of the,

let's say, it's one in a million chance of going to heaven, which is infinitely good ..."

Julia Galef: Or avoiding hell.

Will MacAskill: Or avoiding hell. You might wonder, "Okay. Why should I go to church rather than flip a coin, and if it's heads, I go to church, if it's tails, I don't?"

Julia Galef: I never wondered that.

Will MacAskill: That is the same ... You never wondered that?

Julia Galef: About Pascal's Wager? No. Go on.

Will MacAskill: Well, so Pascal's argument is that going to heaven is infinitely good. We'll put hell to the side.

Julia Galef: Okay.

Will MacAskill: Going to heaven is infinitely good, so any chance of going to heaven has infinite expected value, because we said, "Look at the probability and multiply by the value. Even if the probability is as low as one in a million, the value is infinite, therefore, the expected value is also infinite." That's going to outweigh or be greater than any merely finite amount of good you might have by not going to church, and instead, having a nice brunch on a Sunday morning.

Julia Galef: Right.

Will MacAskill: The issue is that any probability is infinitely good. Any probability ...

Julia Galef: Times infinitely good is infinite expected value.

Will MacAskill: Times infinity is infinite expected value. As far as this argument goes, there's nothing to choose between going to church and flipping a coin and going to church if it's heads.

Julia Galef: Ah, yes.

Will MacAskill: Similarly, me drinking a beer, that has some chance of getting me into heaven, and that, therefore, also has infinite expected value.

Julia Galef: I guess I did sort of think of this counterargument, in the form of, "Well, what if there's another god who will bring you to heaven if you don't go to the church for the first god?"

Will MacAskill: That's right.

Julia Galef: You can just construct arbitrarily -- there's different levels of infinity, so you can construct even better heavens with other required actions in order to get into them, and it quickly becomes ... I don't know. Maybe some people have figured out some way to systematize all that, but it seemed uncomputable for ...

Will MacAskill: Yeah. There are some people trying to do this, including a friend of mine. Amanda Askell is actually working on this, where ...

Julia Galef: Oh, cool. I didn't know she was working on that.

Will MacAskill: Yeah. It's a key part of her PhD. She has this wonderful blog post, which is *10 Responses to Objections to Pascal's Wager in 100 Words or Less*.

Julia Galef: Oh, awesome. That's so great. There's not enough metaethical clickbait out there.

Will MacAskill: Yeah. Exactly. It's a niche audience, but ...

Julia Galef: We'll link to that on the podcast website.

Will MacAskill: Yeah. I actually think that Pascal's Wager is a much better argument than people give it credit for. I do think the response I gave is a very serious, in fact I think a devastating challenge for the argument as stated, but I don't think it's yet a reason to reject Pascal's Wager entirely. I think it's just a way in which it's showing that dealing with infinities is commonly something that our contemporary decision theory is not able to do.

Julia Galef: Right. In the context of moral decision-making, I could imagine maybe versions of utilitarianism having trouble with outcomes that we stipulate have infinite goodness or infinite badness. But it also seems to me like it might come up for moral theories, moral views of the world that just insist that there's *no* exchange rate. Like, "You just can't lie. Lying is just always wrong. There's no amount of good that you could do in the world with a lie that would make the lie okay." Is that like saying that lying has infinite badness?

Will MacAskill: Yeah. There's another analogy to Pascal's Wager that I'll come back to in a little moment, but one case is that very similar problems do seem to arise when you start thinking about absolutist moral views.

An absolutist moral view thinks not that there's something that's good and just has increasing value, say, saving lives, but there's some things that are never outweighed. So killing is always wrong, no matter what circumstances. Even if a hundred trillion lives were at stake, it would still be wrong to kill.

Then you might think, "Well, okay. It's a billion lives at stake, the end of the world, but I could kill one person to save those billion lives." Most people

would think on reflection, "Okay, yeah. You ought to do that." The absolutist view would say, "No. It's wrong."

Perhaps the way to represent that is that it's infinitely wrong. If it's infinitely wrong, then no matter how low a probability you have in that view, then you would get the conclusion that you ought not to do it under moral uncertainty, if you're maximizing expected value. Or so it seems.

It wouldn't just apply to killing, which is one of the more plausible ... I still think very implausible, but one of the more plausible examples for an absolutist constraint.

It also would apply to, say, lying, or extramarital sex. There's many things where you might imagine some moral view, and it's a moral view you might have very low confidence in, saying, "There's an absolute constraint against this action." If we represent that as saying there's infinite negative value to that action, then you should seemingly just not do it no matter ...

Julia Galef: At any cost, yeah.

Will MacAskill: Yeah, at any cost. I think there are very many issues when you start trying to take moral uncertainty into account in the same way as you do empirical uncertainty. And so my research was firstly making the case for thinking why you ought to, and then just going through all the many problems and trying to work through them.

In this case, I think one thing we can do is a bit of a partners-in-crime argument, where we can say, even under empirical uncertainty, as long as you think there's some chance of gaining an infinite amount of value, it seems like expected utility theory says, "You ought to do that." That suggests that when it comes to very, very small probabilities of huge amounts of value, or even infinite amounts of value, something is going wrong with expected utility theory.

So this is just a bug in our current best theory. Maybe it happens a little bit more often in moral uncertainty, but it's just the same sort of bug. We know we need to iron this out in some way. We don't know exactly how.

That's why it's definitely still a problem, something we want to address, but not necessarily one that's very distinctive for moral uncertainty.

That's one approach. But there is another approach as well. So far, we've been assuming that different moral views are comparable, in the sense that you can say how much is at stake on one moral view compared to another. Where, say, if I'm killing one person to save five, there's a meaningful answer to the question of, "How much is at stake?" for the utilitarian, who says, "You ought to kill the one to save five" versus the non-consequentialist, just think of that as an anti-utilitarian, somebody who just rejects utilitarianism or

similar theories, when they say, "You ought not to kill one to save five," there's a question of: for which theory is there more at stake?

It's not at all obvious that we ought to be able to make comparisons between theories. Perhaps it's possible for very similar theories. Perhaps I got one form of utilitarianism that only cares about humans, another theory that cares about humans and non-human animals.

But when you've got these very different moral theories, like absolutism and utilitarianism, perhaps it's just not possible to say that actually one is more high stakes than another.

Julia Galef: High stakes here would be like, for the utilitarian, that "kill one life to save five lives" is not ... There's some utility created there, net utility created, but in the grand scheme of the world, it's not a huge deal. For the anti-utilitarian, if you kill the one person to save five, it's a huge deal, and you've just created, committed, a huge moral wrong.

Will MacAskill: That's exactly right.

Julia Galef: In that, if you felt there was a comparability, then you could say, "Well, we should defer to the anti-utilitarian because the stakes are higher for their view."

Will MacAskill: That's exactly right. Earlier, we were saying that, for the animal welfare view, eating meat seems higher stakes than for the non-animal welfare view.

There is another response, which is that you could think actually, there's just the idea that one of these is much higher stakes than another, that's a kind of illusion. These are such different moral views that, instead, you should not think that we are able to compare the stakes across different moral views.

Julia Galef: What are the units on stakes?

Will MacAskill: I think the natural unit ... I just coin a term and say it's choice-worthiness.

Julia Galef: Choice-worthiness.

Will MacAskill: I think the natural unit is wrongness, degree of wrongness. Where we fairly naturally say that lying is a bit wrong, but punching someone, that's more wrong -- and murdering someone, that's much more wrong again.

I think we do naturally think in terms of quantities of wrongness, even though we might not ever use that term. Because you also might think that the difference in wrongness between punching someone and killing someone is a larger difference than the difference between telling one sort of a lie and a slightly worse lie.

Julia Galef: Right.

Will MacAskill: Again, we seem to be able to make sense of the idea, and that's the unit, in this case. Then the question is, yeah, is the unit of wrongness the same unit? Can you make sense of it? Or is it like saying the difference in temperature between 20 degrees Celsius and 22 degrees Celsius is the same size as the difference in length between a two centimeter object and four centimeter object?

Julia Galef: Right, just a category error.

Will MacAskill: That's just a category error. That's just meaningless.

Julia Galef: Yeah. Right.

Will MacAskill: One of the things I do is actually explore ... If that's the case, you just can't apply expected utility theory. It's just like trying to push a square peg into a round hole. It's just not the right formal system for that. One of the things I do in my PhD, and in this book that I may well one day finish ...

Julia Galef: Hey, you're one for one.

Will MacAskill: One for one.

Julia Galef: Every book you've started, you've finished, so far.

Will MacAskill: That's true. Well, no, I'm 50%, because I ...

Julia Galef: 50%?

Will MacAskill: Well, I've started this moral uncertainty book, but I haven't finished it.

Julia Galef: Oh. Well, okay. Fine.

Will MacAskill: Anyway, I am hoping to finish it in the next few weeks. I develop a formal system that would allow you to make what seem like more reasonable decisions even in light of theories that are very, very different, and therefore, incomparable. Even in light of theories that don't even give you quantities of wrongness.

Perhaps this absolutist theory just says, "No. It's not that there's degrees of wrongness. You can't make sense of the idea that murder is very, very wrong. It's just that there's two categories of actions. There's right actions and there's wrong actions."

We don't need to go into the technical details of the proposal, but the key insight is to think of the different moral views in which you have some

degree of belief as kind of like a parliament, and you can take a vote among those different moral views.

Because when we go the voting booth, we don't compare intensities of people's preferences -- although, I actually think you should…

Julia Galef: Separate episode.

Will MacAskill: Separate episode, I can just rant about current voting systems. I can do that for hours. Even in countries like Australia, they rank candidates. You say, "This is the best. This is my second favorite. This is the third favorite." That's still not comparing preference intensities, it's just giving an ordinal ranking, but then you can put that into a voting system that then spits out, "This is the best candidate."

In the same way, I can use a weighted voting system, where let's say I have three different moral views. One is this absolutist view. The second is a utilitarian view. The third is very different again, let's say it's a virtue ethical view.

Let's suppose I just have 1% credence in the absolutist view, in which case, I would give that a 1% weighted vote. Perhaps ...

Julia Galef: As if, in your parliament of 100 people that are inside your head, there's one person who thinks that absolutism is correct.

Will MacAskill: That's right. Exactly. Maybe I have 59% credence in utilitarianism, so it would get 59 votes. Then the final view, virtue ethical view, would get 40 votes.

Then it turns out that despite current voting systems, there's this huge amount of research that's been done on different voting systems, their mathematical properties, and which are more and less desirable.

All of the ones you have heard of through voting systems that are actually used are among the worst in terms of their mathematical properties. “First past the post” is about as bad a voting system as you can get.

A single transferrable vote, or the alternative vote, which is what Australia uses, is probably second-worst out of the commonly discussed ones. It's actually so bad that, in some circumstances ... If I was going to vote for, let's use the US example ... if I was initially planning to vote Republican and put Democrat second, and then I changed my mind, and switched from Democrats second to Democrats first -- that can cause the Democrats to lose. It violates a property called monotonicity, where having one person increase their level of preferredness for a candidate can make that candidate do worse.

Julia Galef: Wow. Okay. Separate episode.

Will MacAskill: As I said, I could really talk on this for hours.

Julia Galef: I believe you.

Will MacAskill: Turns out, you can just import all of this great work that has been done by voting theorists and social choice theorists from economics into this domain.

In cases where you can't straightforwardly apply expected utility theory, you can treat the different moral views kind of like, yeah, kind of like people in your head. You give them some greater weight depending on how plausible you find the view. Then you take a vote amongst all of them.

That has the implication that you can still make better-or-worse choices even when you can't make comparisons of strength of wrongness across these different moral views.

It also seems to have the effect that you won't get swamped by these low-probability but very-high-stakes theories, because what that swamping relied on was those very low-probability theories saying, "This is just hugely important, infinitely more important." That's not something they can say in this model.

Julia Galef: This is very cool. I'm wondering if it might be -- and I'm sure you've thought about this, but as I'm thinking about it for the first time, I'm wondering if it might be even better to ... Okay. Let's say I assign 59% credence to utilitarianism being correct. If I follow this voting system as I'm understanding it, that means that if there are a hundred situations in which I have to choose how to act, well, 59% is the plurality I guess -- well, and the majority, but even if it were just the plurality, it would win every time, so 100 out of 100 cases I would do the utilitarian thing. What if instead I just randomized? Yeah.

Will MacAskill: Okay. Perfect. That's actually not the case.

Julia Galef: What's not the case?

Will MacAskill: That you would win 100 times.

Julia Galef: Oh.

Will MacAskill: That would be the case if you were using first past the post.

Julia Galef: Oh, so I was misunderstanding your proposal.

Will MacAskill: All the best voting systems have ranked voting. Let's say, if I have five options available to me, you would look at each theory and say, "Well, how

do you order them?" The utilitarian would say, "Save ten lives, that's best. Save five lives, second good," and so on. Would give you not just, "This is my favorite, and these ones aren't," instead would give you a ranking. Some theories would only ever rank things into two categories. I said that perhaps the absolutist just says there are two categories, right and wrong.

Julia Galef: Oh, and the overall voting still works. Interesting.

Will MacAskill: The overall voting still works. It would just have many tied options.

Julia Galef: I see.

Will MacAskill: That would actually mean that, sure, if you had 59% credence in utilitarianism, it would win often, but it wouldn't always win. In fact, it can be the case that, because you're ranking these different actions, you can choose an action that is not the most preferred according to any moral view. Perhaps it's just second best according to all the moral views in which you've got some degree of belief, even though it's top, like most-preferred, according to none.

Julia Galef: Uh-huh. Interesting. We're technically overtime, but if I were to ask one more question about this, it would be:

I'm assuming this is not the consensus view among philosophers who've thought about ethics. Well, by “this,” I'm referring partly to your specific take on how to deal with moral uncertainty, but also to the slightly broader circle of the claim that we should be figuring out what to do about moral uncertainty at all.

Is the reason that philosophers don't all agree that this an important issue that demands a resolution, is that just because it's not very well-known? Or is there some other objection that philosophers have to the kind of reasoning you've been laying out?

Will MacAskill: Yeah. The main alternative view among philosophers, which is one that I struggle to have sympathy for -- it doesn't keep me up at night -- is the view that under moral uncertainty, you ought to do what's actually right.

Julia Galef: Wait. I'm sorry. That's begging the question.

Will MacAskill: Well, yeah. People who reason like you and I, Julia, tend to not find this a compelling view.

Julia Galef: Well, I mean, we don't *know* what's actually right.

Will MacAskill: Well, we don't know, but yeah… I would say, what you ought to do is ...

Julia Galef: You can't possibly be presenting this view charitably.

Will MacAskill: No, this is absolutely the view. In the same way that you might think, in epistemology as well, that what I "ought" to do is believe the truth and disbelieve the false. I *ought* to do that whatever my evidence is like. Actually, very few people have this view, but you might have that view.

Julia Galef: Okay.

Will MacAskill: On this view, yeah, what you ought to do, even if ... Should Aristotle have kept slaves, for example? I think there is a plausible sense in which we want to say, "No. He shouldn't have done. He acted wrongly in keeping slaves," even though, given what he knew, given his evidence perhaps, it was very lightly.

Julia Galef: Well, then I feel like ...

Will MacAskill: He thought it was extremely unlikely that it was wrong to keep slaves.

Julia Galef: Well, I feel like philosophers, if they're using the word "ought" that way, they're answering a different question than the question of, "What should I, a person, with imperfect reasoning and moral knowledge, do?"

Will MacAskill: Terrific. Yeah. That's my view as well. There's just two different senses of "ought" going on.

Then you have this problem of, well, how many senses of "ought" are there? Because I talked about what you ought to do when you're unsure morally what you ought to do, but you shouldn't be *that* sure of my view…

Julia Galef: Oh, no. Oh, no.

Will MacAskill: You shouldn't be ... I was defending this "maximize expected utility" view modified with voting theory a little bit, but that's a new view. You shouldn't be super confident in that. Is there another sense of ought?

Julia Galef: That seems correct, sorry!

Will MacAskill: It seems correct. Yeah. Ought you then to take into account your own certainty about what you rationally ought to do under moral uncertainty? Then I'm sure you're going to see where this is going to go, because you shouldn't be unsure about [crosstalk 00:50:15] that into account, and you're led in an infinite regress.

Julia Galef: Do you have an answer to that that can fit in the very short remaining time we have?

Will MacAskill: Honestly, I don't, actually. I think ...

Julia Galef: That's kind of exciting.

Will MacAskill: Yeah. I think there's a few plausible things you can say. I've tried to do work arguing that you should just go all the way up this regress, and what you ought to do is what you ought to do in the limit of the end of the hierarchy. I think at some point, when it comes to rationality, you do have to say, "This is just what's correct," even though it's not accessible to you, even though there's no way you could have known that this is correct.

Julia Galef: Yeah.

Will MacAskill: There just has to be some objective element.

Julia Galef: Yeah. The infinite regress doesn't seem like a unique problem. There are infinite regresses in all types of knowledge or decision-making. I find them somewhat troublesome -- like, I don't know, with respect to induction or something.

I find them somewhat troubling, but I consider solving infinite regress is this other, separate problem, that it would be nice to work on, but it doesn't ... The fact that it appears in lots of these other problems we're trying to solve shouldn't cause us to instantly give up on our solutions to those problems.

Will MacAskill: That's right. The reason this alternative view doesn't keep me up at night is because I am really trying to use my life to do as much good as I can. I'm really thinking about what are the biggest global priorities for the world. In order to answer that, we don't, despite two-and-half thousand years of Western philosophy, and even more years of non-Western philosophy, we don't have the answers to these moral questions yet.

But we need to act. We need to make decisions in light of our current state of moral ignorance. The question is what's the best decision? If someone says, "Oh, well, it's just, you should just do what's right," I'm just like, "You're just not being very helpful here."

Julia Galef: Yeah. Yes. I feel like the people taking that approach can't possibly have been part of real discussions about decision-making, or they would have heard themselves. They would have listened to themselves and been like, "Oh. I sound like a tool."

Will MacAskill: ...I was going to namedrop some of the people who hold that view, but maybe I won't do that now.

Julia Galef: Yeah, maybe let's not, actually! You could have done that earlier, and then I could have changed the way I phrased that, but we're down this path now.

Anyway, before we close, Will, I always invite my guest to introduce the Rationally Speaking Pick of the Episode, which, often, the prompt I give is just what's a book, or article, or a website, or something that has influenced you? For you, I'll give you a more specific prompt. I would like to hear a

book, or website, or article, or something that you disagree with, at least significantly, but that you still respect.

Will MacAskill: Yeah. The answer that I'll give to this is a book called Anarchy, State, and Utopia by Robert Nozick.

Julia Galef: Oh, Nozick. Yeah.

Will MacAskill: It is an imperfect book in a lot of ways, but what the book does is really lay out the philosophical grounding for libertarianism, and what's known as right libertarianism as a theory of justice, which says that all there is to justice is that, "Did you acquire your property in a way that was just? You didn't take it from anyone else," and, "Did you transfer it in a way that was just?"

I'm not a libertarian. I think it's incorrect. But I think it's very clear, and I think there are many very powerful arguments in that, and I think it's a moral view that at least we need to reckon with and take seriously.

Julia Galef: Very apropos.

Will MacAskill: Okay. Well, thank you so much for having me on.

Julia Galef: It's been a pleasure, Will. Thanks. I'm glad we finally did this.

Will MacAskill: Yeah. Me too.

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