Rationally Speaking: Amanda Askell on “Pascal’s Wager and other low risks with high stakes”

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 is today's guest, Amanda Askell. Amanda is in her final year of her philosophy PhD at NYU, where her research focuses on all the ways in which incorporating infinities, into our decision theory and our moral theory, causes problems. Amanda, welcome to the show.

Amanda Askell: Thank you for having me.

Julia Galef: So, I've been meaning to invite you on the show for a while now, but the immediate impetus was, in a recent episode with Will MacAskill, we were talking about normative uncertainty, among other things, and we ventured briefly into the territory of Pascal's wager. Which very briefly, is the philosophical argument put forth by Blaise Pascal, in which he said, “You know, people should really try to believe in God. Because if you're wrong and there is no God, then, oh well, that's fine -- but if there is a God, then believing in him is going to be much better for you than not believing in him. And so the expected utility of belief is much higher.”

This is one of those arguments that I find there's this non-monotonic relationship between how much philosophical exposure people have and how seriously they take that argument. In that a lot of people without a ton of education are like, yeah, that makes sense. You should believe, because higher payoff.

And then a lot of people who are well-educated are very dismissive of Pascal's wager, and they're like, “Oh, that's ridiculous,” and they have one or two objections to it. Then philosophers are like, “Well, actually it's harder to dismiss than you might think.”

Which is ... I like things with that shape of people's belief in them.

Amanda Askell: Yeah, and I think I personally went through that with Pascal's wager. Where when I initially heard it, I was just like, this is absurd; of course that's wrong, gave a couple of standard objections to it -- and then, I think, did the appropriate thing, which is test those objections…

Julia Galef: Test them?

Amanda Askell: Yeah, I think it's important that if you're going to raise objections to an argument like that, that you want to be like, “Well, what's the best defense that I can make against these objections?” Then when I started to do that, it was like, “Actually I'm really not convinced by any of these objections. At the very least, we should be taking this argument quite seriously.”

Julia Galef: Yeah, interesting. In fact, to finish my thought earlier, Will's pick, the recommendation he gave at the end of the episode, was a post that you wrote about the most common objections to Pascal's wager and, in 100 words or less, why they fail or how they fail, which is great.

Maybe I'll just tell you what I ... My views on Pascal's wager have become a little more nuanced or uncertain the more I've thought about it. But I can tell you what I historically thought was the knock-down objection to Pascal's wager, which is: “Well, sure, we're currently thinking about this possibility of a god who will give us infinite utility in heaven if we believe in him and maybe give us infinite dis-utility in hell, if we disbelieve in him. And if we're just thinking about that god, then Pascal's wager makes sense -- but what if there are other gods? What if there's another god who will put you in an even better heaven or an even worse hell if you believe in him and disbelieve in the first god?”

You could posit any number of possible gods with different preferences about how you believe. So it isn't at all clear -- if you're going to start positing that, it's not at all clear what you should end up believing or disbelieving.

Amanda Askell: Yeah, so it's interesting. This is the “many Gods” objections to Pascal’s wager. I think that the one question I usually ask for people who raise this is:

Let’s just assume for the sake of argument that there is only one kind of heaven, because things get a bit more complicated when we think there are lots of different kinds of Heaven. So that's our infinite outcome.

Julia Galef: Mh-mmm

Amanda Askell: And then I say to you "Well, would you prefer a greater chance of getting that outcome, or a lower chance of getting that outcome?"

Julia Galef: Greater.

Amanda Askell: Yep. So now Pascal's wager starts to look a lot like a standard decision problem. Suppose you said to me "I have this great cup of coffee for you" and the way that you can get this cup of coffee is to give me two dollars.

I may be like, "oh, there are other ways I can get that coffee". I can steal it from you, I can do these other things. Standard decision theory says "well, if this is the desired outcome, you should do the thing that's most likely to lead to this desired outcome."

Then the question is, why isn't the same true with Pascal's wager? So if you say to me "Well there are lots of different Gods that you could believe in" and I say to you "Okay, is there one that is more likely to lead to this outcome that you'd want?" And if so, it seems kind of sensible to think that you should do whatever pleases that God today. If that's the case, then we've

already bought Pascal's wager at this stage we're just debating about which action to take. That's my worry, with the many Gods objection, is it comes after you've already bought the argument.

Julia Galef: So are you saying that… I think sort of implicit in my version of the many Gods argument was that we don't have any way of assigning a higher probability to one God than another hypothetical God.

There are some Gods that have have been proposed by humanity throughout its history. But we could certainly imagine other Gods, and we can't sort of rule out the fact that they don't exist. For some definition of God. So it's not like there's one God that I'm more confident exists, and we should just be optimizing for having that belief.

Amanda Askell: So I think it's interesting, because then you've got something like "Well, I have just of an equiprobable distribution over all of these Gods."

If you're quite responsive to evidence and you think that we can have evidence in favor of one God over another -- say the testimony of someone you really trust -- it's really easy to break that kind of equiprobability. Suppose you just read a really good article by scientists that you think is very credible and they say "oh, PS I have evidence that this god exists." You may not find that in any way compelling and it may make almost no difference to you-

Julia Galef: -- But that “almost” is doing lot of work.

Amanda Askell: Exactly. It's just going to break the symmetry. The other thing is, I think that actual people don't assign the same probability to different Gods because some of them just have more plausible properties than others. It's just that when we get to really small probabilities people start to-

Julia Galef: They start to seem the same.

Amanda Askell: Yeah. They start to look very similar.

Julia Galef: Well, let me actually revise my statement. I think the thing that made it feel like all a wash was not actually that I think all possible Gods have the same probability. But rather that you can always construct a hypothetical God such that even if it has a really low probability, has an even higher payoff for believing in him.

Amanda Askell: Yes.

Julia Galef: Or it.

Amanda Askell: This is where things get kind of complicated because this is kind of the different heavens point.

Julia Galef: Oh, right, yeah.

Amanda Askell: It is interesting. So the question might be, suppose that the Judeo-Christian God says "Hey, I'll give you–" let's assume that we add this to the bible or something, where it says "Heaven is a series of +1 utility days, and just to be clear that this is what Heaven is like."

And then another kind of religion comes along and they say "Hey we know that you think that we have, like, a quarter of the chance of being true compared to the Judeo-Christian religion -- but in our Heaven it's +5 utility days that you get, and for an infinite period of time."

Julia Galef: I'm seeing, like, a competition for market share of believers.

Amanda Askell: Yeah. So then it does become more complicated, because the natural thing to say -- and I think it can be tricky to get this out normally, but the natural thing to say -- is you care about both the probability of the reward and also how good the infinite reward is. Again, we're already at the point of accepting the wager.

It's easy to see this as a reductio, but it's actually not reductio it's more like "oh, that seems like a difficult decision problem." So many-

Julia Galef: By reductio, you mean it's easy to say "Well that invalidates the whole concept of Pascal's wager"?

Amanda Askell: Yep, we see this kind of difficulty and we're like "Oh, so therefore Pascal's wager doesn't work."

But it's unclear how we can make that stick. We can say, "Oh I have all of these credences across different Gods" -- and let's say that in the case that I just gave, you prefer something where a lower credence is true but offers you a higher heaven. We can imagine just constructing a decision theory that would yield that result. If that's the case then the problem that we're facing is one of just deciding what the best thing to do is.

I think what people are foreseeing is that counter-intuitive things are going to happen if we go down this road. When counter-intuitive things happen, I'm going to just reject whatever is this that generates in Pascal's wager.

So even if it's a fairly plausible decision rule -- that says you prefer higher probability of infinite outcomes to lower probability, prefer better infinities to less good infinities -- I'm going to ultimately reject one of those plausible looking premises. If it makes me chase infinities for my whole life, then that's the kind of reductio that other people are moving towards there. In and of itself it's not like an objection to the wager if that makes sense.

Julia Galef: Let me see if I understand what you're saying. Are you saying we run into the same problem for all sorts of things that don't involve Gods or infinite

utilities or small probabilities? Like you could say – well, not the small probabilities part, I think that's still important, but -- for example you could say "well, I want to go to a nice restaurant and the most plausible way to do that is to think of a restaurant I like and go there."

Amanda Askell: Yep.

Julia Galef: But like it's always possible that by doing that, there's a tiny probability that I'll run into an accident, maybe there's some tiny probability that if I sit here and wish to be at a nice restaurant it's possible there's a genie who just responds to wishes about nice restaurants on this particular day or something.

Amanda Askell: Mh-mmmm.

Julia Galef: So I can construct sort of arbitrarily low probability but higher value ... I don't know. That's kind of a mangled example because the restaurant itself isn't a high enough utility to justify not doing the plausible thing. But like I could imagine arbitrarily bad and low probability and arbitrarily good and low probability outcomes for anything, not just Gods.

Amanda Askell: Yep. So I think the thought is that standard decision theory says nice things, like I should help my neighbors or I should help people in other countries. I should do what I can to create large but finite amounts of utility in the world. And we might be concerned that if we start to take infinities really seriously then I say, well, how do I get infinite utility in this world?

Suppose I think things like, “Well maybe I could invent a time machine or I could invent a machine that would let me travel much faster than the speed of light. Maybe I could have a taxi service where I could move people around in an infinite universe.”

Julia Galef: Right.

Amanda Askell: We can create kind of outlandish scenarios. Our worry is that the kind of actions we'd take if we became the chasers of infinity would be intuitively completely wrong. They wouldn't be things like helping people, they would be things like taking this huge risk for some like sci-fi scenario that involves helping infinitely many people in the universe. At that point we might feel warranted in saying "Hey, I don't quite know what went wrong along the way, but I know something went wrong, because that shouldn't be the conclusion."

I think I am actually really sympathetic to that. I think that Pascal's wager or this kind of Pascalian thinking is only really going to be plausible if it doesn't lead to those kind of absurd conclusions. If it ends up being that you should only take infinties into account if you happen to be in a scenario where it doesn’t change the ordering of what you would do.

But that's really rare. I'm less worried about that. I'm less worried about a theory that says “Care about infinities, but actually most of what you should do is what you're already doing,” than I am about a theory that says “Try and invent a time machine.”

Julia Galef: What kind of theory would say "You should care about infinities at all, but that shouldn't change most of what you’re doing"? Don't infinities, if you take them into account, just swamp all possible decisions?

Amanda Askell: They do swamp, and I think it some ways this feels like it would maybe be too lucky. This is a worry that i have.

Julia Galef: What do you mean by too lucky?

Amanda Askell: It would be too lucky if the world were such that the actions that create the most finite value -- I mean if we're in a finite world and we can only create finite value -- line up neatly and perfectly with the actions that create infinite value.

You may think something like being a good person, most religions believe that being a good person is just something that is much more likely to get you into heaven than not. So this is a fairly consistent action of, you're looking at the religious hypothesis that may be kind of infinite in expectation.

But luckily being a good person has this huge finite benefit to people around you. So it's like "Oh well that's fine, I'm okay with an outcome that says you've got much more reason to be a good person than you thought you did." I'm not super worried about that, because I was already happy with the theory that I should be a good person.

But we might just worry that it tells you to do things that are kind of strange. So in the case of religions, maybe for example it means that even if you think that these religions are just incredibly unlikely to be correct, that you should nonetheless devote yourself, and devote your life, to becoming a believer in that religion. Or you should raise your children to have that religion.

I think when we get to actions like that which may even be harmful, you can imagine a case were it would be very psychologically distressing for your children or something. That becomes a lot harder to accept, that you should take this finite hit in just some type of hope to get some benefits.

Julia Galef: Yeah. I think my true rejection of Pascal's wager type arguments is the thing that you were talking about. Where if you find yourself “chasing infinities,” as you put it -- doing things that intuitively seem crazy and not the best thing to do at all, just because your decision theory is taking into account infinities -- then something’s gone wrong somewhere.

But that, to me, feels like strong evidence that there *is* a solution to Pascal's wager -- but it isn't in itself a solution. We still have to figure out what went wrong.

Amanda Askell: Yeah, I think this is the thing that troubles me. Because you obviously have this sort of "Okay every one of these premises seemed plausible, but the conclusion just seems very implausible."

One kind of possible reaction is to say, “Actually the conclusion isn't as implausible as you thought,” because of the actions you take, like being a good person. Another one is just to say "I know something went wrong. Because that conclusion, the negation of that conclusion is more probably than the premises than I put into the argument.”

Julia Galef: Right.

Amanda Askell: When that's the case, I think it is important to know that you're left in this sort of difficult situation, because it's not enough to just say "Well something went wrong somewhere. Farewell, argument." The thing to say is "I know something went wrong somewhere, but I now need to identify what it was.

Because one possibility is actually i was just wrong about the conclusion being super implausible. Or I find some reason that that was incorrect, and I should actually take into account really small probabilities of extremely large amounts of value.

That's the kind of situation I'm in with Pascal's wager, and I think just with infinitive decision theory generally. It’s a point of tension.

Julia Galef: For you internally?

Amanda Askell: Yeah. I think for me internally, is “What has gone wrong, and am I going to have to give up something really fundamental about what I think decision theory should be about, or ethics should be about?” Yeah.

Julia Galef: One other classic solution that seems plausible to me is just to say "Well, this idea that we should be maximizing our expected utility -- maybe that's the problem, there's something wrong with that idea."

I find that rejection of expected utility maximization pretty plausible in other cases. Like I think it's fine to be risk averse for high stakes one-shot deals. If I'm looking at a decision that affects my whole life, I'd much rather have a sure bet of a pretty good life than a very small chance of an amazing life and a large chance of a bad life. So I'm happy to reject expected utility maximization there. Couldn't we also therefore say that "Maybe I just don't care infinitely much about infinite good outcomes”?

Amanda Askell: Yeah, and I think this is somewhat a case that can split the kind of ethical theories from the decision theory. Because I don't find it that plausible that people have bounded utility functions, personally.

So over the course of my entire life it's not the case that I can get infinite benefits. There's just some kind of upper bound on utility. We find that plausible in the case of, like, dollars, because there's only so much I can do with dollars. So maybe we just think if you give me infinitely many dollars, at some point these things just become useless to me and so I stop valuing them.

But in the case of actual pleasure I'm like, "No, my life could potentially go on forever and I get to keep having great days." I'm like -- that just has unbounded value.

But some people are going to reject that, some people are going to say, "No, in the personal case we can have or even do have bounded utility functions."

Julia Galef: Its tough, actually… I worry maybe -- in response to my own point -- that when I look at an expanse of infinite happy days, and I say, "Well, I don't care about that infinitely much", I only say that because the end is far away. But if I were close to the end of my finite days then I would continue to care about getting more happy days.

Sort of like how people say "Oh, I don't care about living to 80 versus 90." But when you're 79, you care a lot about living to 80 versus 90.

Amanda Askell: Yeah. From a kind of personal note, the thing that often strikes me is that this feels -- and maybe this is where you start to transition into the sort of ethical realm, because it feels a little immoral.

Julia Galef: To your future self?

Amanda Askell: Yeah.

Julia Galef: Like you're wronging your future self.

Amanda Askell: Yeah, I often feel like we really wrong our future selves, in ways that to me seem quite unjustified.

I've thought about this in the context of a completely different topic, which is promise-making, where you make promises on behalf of your future self. So people sometimes talk about how you can ruin the life of your future self by doing things like failing to invest in things like a pension.

But suppose I promise to dedicate a year of my life to the fire department in my local area when I'm 40. And I do that when I'm 18. Should you be allowed to make that kind of promise? Technically you're of age… but it feels like the

reason we should be kind of suspicious of that is that people just don't care as much as they should about their future selves.

Julia Galef: Ah, I see.

Amanda Askell: We kind of think that there's nothing unacceptable that you cannot do to your future self, because its you. But I'm not sure that's true, because the 18 year old just isn't caring enough about the person who to them seems like a very different person.

I do worry that when we're like "Yeah, I'll take this great finite life" we're just trading off our future selves for our present selves, in a mean way.

Julia Galef: Right.

Amanda Askell: To me this is a stronger argument when it comes to other people. So in a case where we might think that personally we have bounded utility over the course of our life --

Julia Galef: Just to clarify for listeners, and to make sure I'm right, “bounded” would mean as you increase the number of happy days that you're offering me, from one hundred to a thousand to a million to a trillion and so on… the value that I put on those things does go up, but it doesn't go up proportionally. So the amount of extra value that I place on each additional increase starts to go down.

Amanda Askell: And importantly that tends towards a kind of upper bound, so some finite amount. So any given lifetime value can only be finite. So sometimes when we do temporal discounting and we think of our future selves, the idea is that we're just giving less and less value to each day.

Julia Galef: Oh, the farther away it is in time.

Amanda Askell: Yep. So if you give me a dollar today, if you give me a dollar tomorrow that's slightly less good, and so on and so forth until… if I could look at the amount of value for all of my dollars for all of my days it’s always finite.

Even if we think this is true of like, your prudential or self-interested value, to me there's something odd about it in respect to moral value.

Julia Galef: So when it’s other people experiencing those millions or trillions of happy days simultaneously, in parallel -- not in sequence, in one person?

Amanda Askell: I mean that's possible. It's strange because in the case where there could be infinitely many people, or where lives go on for infinite periods of time -- You have to discount across space time in that case. And then it's like "Well why do I care?"

I think it was Parfit who said something like “Pain in the future feels just as bad as pain today does.” So there seems to be something already a bit odd about our prudential theory.

Julia Galef: You care less about pain the future.

Amanda Askell: Yes. But there's something extra odd about us caring about this -- if we're just disinterested, and we're just caring about people and what's the value to them, why then would I care less about these future people?

Julia Galef: Right, than these present people. So the effect of infinity on ethical theories is the other big categories I wanted to get into, but before we leave Pascal's wager I just wanted to ask what’s your current best guess about what we should do in cases of Pascal? I assume you haven't chosen to believe in God because of this theory, or that you wouldn't choose to believe in God if you could push a button and make yourself believe in God?

Amanda Askell: Oh I've wondered that before.

Julia Galef: Oh okay.

Amanda Askell: That's a tough one, because I feel like for most people it's difficult to control your beliefs and to not make them merely responsive to the evidence.

Julia Galef: Yeah.

Amanda Askell: I do also think that at the moment I would say the reason I would not press the button -- the case would be made very difficult if you said to me "You can never press the button again, you get your one chance." Then I might actually break out in a sweat, and it would actually be kind of tough.

Whereas the reasons against it are just kind of uncertain, like I've mentioned before. Where all of this just feels like something might have gone wrong, but the things you can come up with, like the many Gods objection, or bounded utility functions -- they don't quite convince me that somethings gone wrong.

Julia Galef: You’re pretty sure there is something that has gone wrong, but the fact that you haven't been able to put your finger on it gives you pause.

Amanda Askell: Yeah. Basically at the moment I just stopped working with the impossibility theorem. I know everything I can have, and I don't quite know which thing I should get rid of.

Julia Galef: Right. At least one of these things must be wrong.

Amanda Askell: Yeah, that does mean that, from a kind of visceral point of view, even the original religious version -- the things you consider in Pascal's don't have to be religious. But I find it kind of psychologically compelling sometimes. I

remember first reading about or understanding the concept of hell, and I was just like -- this is actually extremely psychologically distressing, that this is even a plausible state.

Julia Galef: I agree. I first read about it when I was -- I mean I grew up in a non-religious family, but I remember playing on the playground with another 7 year old who told me about hell.

Amanda Askell: Yep.

Julia Galef: And I couldn't sleep for like a week. I couldn't understand how other people went on with their lives, even believing this was a thing. Not even if they thought *they* were going to hell, but just knowing that *anyone* could go to hell, like -- how could you just live normally?

Amanda Askell: Yeah.

Julia Galef: I was just appalled.

Amanda Askell: Yeah. And I think it means that I have more sympathy for people who evangelize because if you literally think other people are going to hell if they don't believe this then it seems like almost immoral for you to not evangelize to people.

Julia Galef: I know. Yeah. I agree.

Amanda Askell: Yeah, I think those considerations do give me pause, and I'm like, well, even if I didn't believe it was evidentially based and I could predict that it wouldn't have a huge influence on my life as a whole… it would put me in a very difficult situation.

Julia Galef: There's actually one other Pascal's wager related thing that I wanted to ask you, which is: the reason it came up in the episode with Will is because he was talking about giving some weight to moral theories that you disagree with but you think *might* be true. Like you put a ten percent chance on them being true or something.

Amanda Askell: Yep.

Julia Galef: And that should affect your decision making. You should act in ways that wouldn't be catastrophic, on the ten percent chance that that moral theory was true. And a number of commenters afterwards said, that's basically Pascal's wager.

And the response to that -- I can't remember if Will gave it or not, but the response I would give to that – is that a ten percent chance is substantial. Like, I make decisions all the time on a ten percent chance that something -- I take precautions even though I think there's way less than a ten percent chance that I'm going to get robbed.

Amanda Askell: Yes.

Julia Galef: I still take a precaution because it's not completely negligible.

Amanda Askell: Yeah.

Julia Galef: So I guess I'm wondering… this kind of question comes up when people are talking about working to reduce global catastrophic risk, or existential risk. An accusation is that this is Pascal's wager -- you can just say the stakes are so high, if you're talking about risks from technology or natural disasters that could end humanity, that even if that probability is incredibly low, that forces you to devote all your energy and resources to it. And then something's gone wrong there.

Other people come back and say well it's not that low, I don't think this is a Pascal's wager situation. So I'm wondering, does it make sense to have a cut-off, where you're like: if a risk is above one percent according to my best guess, than it's worth taking into account. And it it's below one percent you should take it as zero, or something?

Amanda Askell: I think this is what people naturally do, but it strikes me as quite irrational.

Julia Galef: It's kind of kluge-y.

Amanda Askell: Yeah, and I think that we tend to round down to zero in a way that doesn't make sense.

So I can give my kind of thoughts on the relation between those two things. On the one hand it's worth noting that Pascal's wager does involve infinities. So the strange things about infinities is that all you need to get them off the ground, and doing all the work they do, is like a [tiny] probability and that's super easy to get. Nonzero probabilities are things that don't even have to be consistent with the laws of physics because you don't have certainty in the world of physics. So it can start to invoke magic, and that's enough to get the nonzero probability.

In cases where it's just really large finite amounts of value, I think that we're kind of inclined to round down too quickly. My suspicion, and maybe this isn't correct, is that the thing that we don't like is especially when were uncertain about our estimates.

So sometimes I want to be like, "Well, imagine that there was just a button, and if that button is pressed everyone going to die instantly who currently exists." It's extremely unlikely -- say that the buttons only going to be pressed if someone falls on it and it's just very unlikely, it's in a quite safe place. And I'm like, "I've actually found a way to create a barrier around the button, to make it even less likely that it can be pressed."

Julia Galef: That feels good.

Amanda Askell: Yeah, that feels good -- but it's already a small probability. I'm only making it a little bit less. But when the value isn't high enough and the probabilities are concrete, we're kind of happy with that idea of multiplying in that case.

Julia Galef: So maybe the rule should not be just, "If it's less than one percent, round it down to zero." Maybe the rule should be, "If it's less than one percent *and uncertain*." So it's, like, I'm very unsure how to estimate this risk, but if you ask me to pick a midpoint of my distribution it would be one percent or something.

Amanda Askell: Yeah.

Julia Galef: Those would be the things that may be worth rounding down. If there was like a natural disaster that happened once every million years or something, but it was a reliable thing -- every year there was a one in a million chance -- and we could reduce it, that then that would be worth it.

Amanda Askell: Yeah, and I recently started to think about an area that I'm calling the "moral value of information." I don't think that we should embrace it as a decision rule but more as a debunking and exclusion, in these cases.

So I think that when we're really uncertain about what the actual probability is, ask: "What would we do in a natural environment, as humans, when we're uncertain about that?"

Julia Galef: The natural environment?

Amanda Askell: Oh sorry, I'm just thinking like, in everyday life.

Julia Galef: Ha, so the natural environment is the world. As opposed to your environment of abstraction.

Amanda Askell: Yeah, the world where our decision procedures have evolved. Because sometimes it can be good to kind of try to imagine that world.

Julia Galef: Yeah.

Amanda Askell: In case we're worried about the rationality of our procedures. So sometimes I think about this as: You have no idea really about what probability is assigned to a given outcome. So say you're in a completely new environment, and you just don't know what the predators in your environment are like, and you don't know what probability to assign to there being tigers in your local area.

One thing you probably *shouldn't* do is act either as if there are definitely tigers or as if there are no tigers. The thing that you're going to end up doing is try to get more information about your local environment. You're going to want to constrain those probabilities, because when you're uncertain it's extremely valuable to get information about what your environment is like,

when you currently don't have very robust probability estimates. Rather than just act as if it's a really known robust probability, [like] maybe there's a ten percent chance that there's tigers, so let's act as if there is a ten percent chance, and go hunting as we would if that were the case.

Instead, you should try and figure out more. Because by sending out someone to find out what your environment is like, you're going to get a lot more value than anything else you can do.

I think that in cases like global catastrophic risks, maybe people are like worried about this being kind of a waste of time, and in part just like, "Well, I see these things" and I'm like, "You can't give me a good number on that, you can't give me a good probability estimate there." And we're inclined to throw up our hands and do nothing or kind of wait. That makes sense if what you want to do is wait to get more information.

But then you want to say to people, Right, but two things I guess: One is that a lot of the people who are working on global catastrophic risks are actually just trying to get more information about them. They kind of agree that information is super valuable here.

The other is that sometimes there's this very large cost to delay. Like, sometimes you can't delay and wait for more information to come in. You have to kind of act under the assumption that something might go kind of badly wrong.

I don't know how convincing this is, as a debunking explanation as to why people are inclined to just be like "Oh, pretend it's zero," but I think what they are really saying sometimes is like, "Let's just wait and see." I suspect that you can give a kind of answer to that, like, "No, the way to get more information here, and to make these probabilities more concrete, is to actually work in this area." So we'd actually have more reason to work in this area than you might think.

Julia Galef: I wonder if part of the intuition behind rejecting worries about small but important risks is that it's the kind of thing where you're likely to get scammed, in some sense. Not that people are deliberately deceiving you, but that this is the kind of argument that it's really easy to rope people in with. And so it's not necessarily appropriate to use standard decision procedures -- we should have more cautious, or more skeptical procedures, when people are telling us things that are commonly scams.

Amanda Askell: Yeah, and also for which we can't get immediate feedback. So if I give you a small probability of a really high reward – like, take a standard lottery case. I buy a lottery ticket and I lose every time. This is consistent with everyone losing the lottery and no one is getting a payout.

So you might worry that in this kind of case someone can always say to you, if they're doing this "small probability of gaining" venture, that "This product

is almost certainty going to fail. But if it succeeds, we're going to be bigger than Google." And if they say that, then it's totally consistent with that they fail every time.

You might think that this means you just kind of have to use your intuition or having to analyze their proposal.

Julia Galef: So maybe it's like: The incentives to exaggerate, or even unconsciously deceive other people, are greater when it's not going to become quickly apparent to the viewer that you're exaggerating or deceiving them.

Amanda Askell: If it's consistent with everything that I say, that you will likely never get a payout from me, then maybe you're kind of worried, because this does give me a strong incentive to say "Oh yeah, when the payoff comes it will be really great."

Julia Galef: Right. Yeah.

Amanda Askell: This is a little like the kind of "Pascal's mugging" cases. Where someone comes along and says, "I know you think it's really unlikely that I'm going to give you a great payoff, but how unlikely do you think it is?" Well, I think it's maybe one in a thousand, or one in ten thousand. And they're like, "Oh, I'll offer you ten times that. So I'll give you ten thousand dollars, if you just give me a dollar today." And then you're like, uh, hang on…

Julia Galef: Right.

Amanda Askell: I think one thing that's important to mention in relation to that, though, is that this stuff is kind of complicated.

So when someone makes you that offer -- this is just bolstering the Pascal mugging case -- they say to me, "How likely is it that you think that I will keep a promise to you?" And let's say I say, like, one in a thousand. Then they say, "Well if you give me a dollar today I'll give you five thousand dollars tomorrow, so in expectation you're like five dollars better off. So you should do it." And they can kind of increase the offer depending on how skeptical I am of their claim.

But there is a complex set of factors at work here. As the offer that someone makes me becomes higher, the chance that I think that they can actually provide it gets lower. If someone is offering to completely change the universe for me, then I'm like, I really just don't think you can do that.

Julia Galef: Right.

Amanda Askell: And they also have a larger incentive to deceive.

Julia Galef: That's true. But that really only feels like it applies in cases where it's an individual person promising that they have resources that they can give you.

If the claims are more about the state of the universe, or what's going to happen in the universe, then it's less clear to me that large stakes are less probable. Well -- maybe they are…

Amanda Askell: No, sometimes i like to distinguish between, I call them, Pascal's muggings and Pascal's trades.

Where with the standard Pascal's mugging case, my credence that they're going to give me the reward is closely tied to the reward that they are offering me. For those reasons I gave. It's not by magic, it's connected to the outcome, but they have a higher incentive to deceive. And I just don't think that they can actually provide it, and so that drops over fairly quickly.

It's bit like the difference between opening up a book, and it's got a voucher for a free metal bookmark in it, and you're like "awesome." [Versus] you open up your book and it's got a voucher for a million dollars, and you just throw it in the trash.

But there are lots of cases where you feel that this offer being made to you completely makes sense. So there is a sense in which, if the mugger comes up to me and is like "Hey, what's your credence I'm going to keep a promise? If you give me a dollar today I'm going to give you five thousand tomorrow," and I'm like, that really seems like you just made that number up.

[Then they say,] "Here's all the evidence that I'm actually a multimillionaire. Plus I'm like super lost and I really need this dollar right now." And they just start to pile on the evidence. Then suddenly it could easily reach the threshold.

So even if I think it's quite unlikely, someone could reach it. The optimist in me about this case says there is nothing going wrong. Your intuition is totally correct, in the case of being mugged or scammed. But the thing that was making you be mugged or scammed wasn't just a low probability of some really high value outcome.

It was just that the probability kept getting lower and it was never an expectation or thing that you should do. You always had enough of a credence that you were being scammed that this was never a valuable trade for you. So it wasn't the structure of it, just something generating the underlying doubts that made it rational for me to not accept the mugging, but may make it rational to accept some cases.

Julia Galef: So let me see if I understand. So is the rule you're proposing something like, you should take into account the probability that you're being scammed -- and that, in some cases, will be related to the amount of payoff the person is promising.

Amanda Askell: Yes.

Julia Galef: But not in all cases. In addition to that, you should be taking into account the evidence about whether this is real.

Amanda Askell: Yeah.

Julia Galef: The probability doesn't just depend on the payoff structure, it also depends on evidence about how plausible this is.

Amanda Askell: So it's essentially to say that standard utility theory just does fine with these cases. So in a case where a stranger just comes up to you and offers you this ludicrously high payoff, or in a case where you get a voucher in your book that says a million dollar is yours, I'm just like, "This bookstore owner probably doesn't have a million dollars. They have no incentive to give this to me. And there's almost certainly some kind of catch here. Or it's just lying to get my attention." So the probability that you assign to getting a million dollars might be so low that it's not even worth doing anything other than throwing it in the trash.

But that's consistent with standard expected utility theory.

So one rule might be to reject expected utility theory, and say "Hey, just assign these things something like probability zero, or treat them as if they were probability zero." Another thing to say -- and the thing that I'm more sympathic to -- is "No, expected utility theory will tell you not to accept the offer of Pascal's mugger, because your credence that they're going to give you the reward diminishes more steeply than the reward they are offering."

This doesn't mean that all cases of small probability and high volume outcomes are ones that you should ignore. Just take a case where I say to you, "Hey, would you like to take a ticket to a lottery that has a non-zero chance, has a .005 percent chance of getting you twenty million dollars?" If you just rounded that down to zero you'd be like, "Well, I'm indifferent between having this and not," but you're not indifferent. You'd probably rather have the ticket.

Julia Galef: Oh, but that's kind of support for the rule that I came up with a few minutes ago. Where I was like, If it's a very low probability of a very high outcome, *but* it's very certain, very well defined -- then you should take it into account. But otherwise not.

Amanda Askell: Yeah, and I think again this is just a reaction to ambiguity, or uncertainty about probability, that's interesting. Like, we almost *want* to ignore them. But I don't see any kind of rational reason to do that.

Suppose I said, "I really don't know what chance this ticket has of getting you a million dollars, and it's really low, could be anywhere from zero to .005 percent." It still it seems the dominant strategy is for you to take the ticket. It becomes more complex when we have trade-offs to make. But we seem

really averse to treating these cases where we're really uncertain about the probabilities, to taking an expectation and just acting on it, in those cases.

I think ultimately we can give an explanation for that in terms of the fact that the rational thing to do in those situations is usually to seek out more information. But sometimes you just cannot. Sometimes the cost of trying to wait to get more information is just too high. If I'm just like, "You can only take the ticket now or never. Take it and you have to pay me a cent." Maybe we will have some type of trade off, maybe a cent is actually too high in that case. But take any arbitrary amount.

Julia Galef: Sure.

Amanda Askell: I'm inclined to just think, hey, we can explain why people don't like acting on those things, but-

Julia Galef: - But it's not clear that they're right.

Amanda Askell: Yeah.

Julia Galef: Well, before we wrap up I wanted to invite you to recommend a pick for the episode -- that's a book or blog or a journal article, something that influenced your thinking in some way. What would your pick be?

Amanda Askell: So Vallentyne and Lauwers have this paper called *Infinite Value [Ed: Utilitarianism]: More is Always Better*.

Julia Galef: Do they have a thesis they are defending, or is it just a survey of different problems?

Amanda Askell: In that paper they're kind of really looking at these principles, these kind of extensions of Pareto, so principles if you take Pruital to be pretty core and you take certain other principles to be pretty core in this area, then what sort of rules do you end up endorsing? And I like the sort of train of thought in that paper.

Julia Galef: Cool. Excellent. Well, Amanda thank you so much for coming on this show, it's been a pleasure.

Amanda Askell: Cool thank you.

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