Julia: Welcome to *Rationally Speaking*, the podcast where we explore the borderlands between reason and nonsense.

I'm your host, Julia Galef, and with me today is our guest, Professor Kenny Easwaran, who is a professor of philosophy at Texas A&M University, where he specializes in sub-fields of philosophy, including epistemology, and mathematical logic, and the philosophy of math. Kenny, welcome to the show.

Kenny: Hi, welcome. How are you?

Julia: Good, thank you. Thanks so much for joining us.

Before we begin the episode, I just want to remind our listeners that there are full transcripts of every episode now posted on our website, rationallyspeakingpodcast.org, so if you prefer to consume your information and insight-dense podcasts by reading instead of listening, then just go download the transcript there.

Today I want to talk about a controversial paradox in philosophy called Newcomb's Problem, or Newcomb's Paradox. We'll discuss why it's important and how it's shaped the field of decision theory, and maybe what it has to say about other philosophical topics, like free will. Kenny, to start things off, why don't you just give some context for this paradox, you can explain it in your own words, and you can relate it to the work that you do.

Kenny: Yeah. Traditionally decision theory is based on an assumption that there are some parts of the world that you control, which we might call your actions, and there are some parts of the world that you don't control, which we might call the state of the world.

Traditional decision theory, as developed by Leonard Savage and other statisticians in the middle of the 20th century, assumes that these two things are independent, and that the outcome is the product of your action plus the state of the world. And suggests that the way you should decide what to do is by figuring out what's the probability of any given outcome -- given each of your actions -- and then do the one that has the highest expected value, which you can calculate mathematically.

Julia: So for example, if I were trying to decide between taking job A or job B, and I maybe have a more variable salary in job B, so a higher chance of getting no money this year, but a small chance of getting a ton of money, I might take job B because the expected value is higher. Something like that?

Kenny: Right. It depends on exactly what the probabilities are and what the possible values of the outcome are.

Then, this arises out of traditional study of gambling, in particular, is the way that this is initially set up. We think of the dice or the lottery ticket as the state of the world, and your choice of whether or not to play or how much to bet is the action. In that sort of setting, this makes a lot of sense.

Now, there's a lot of situations in the actual world where the things that we want to describe as the states of the world aren't clearly independent of your choice of action.

Here's a way this initially emerged: in the 1950s, the statistician R.A. Fisher, in his early life developed a lot of the important groundwork of statistics and biology that shaped much of the 20th century. But later in life he was actually working for the cigarette companies. He was arguing that the evidence we have, so far, at least in the 1950s, doesn't prove that smoking causes cancer. What he said is, "For all we know, there's just certain people that tend to like smoking and this is caused by some sort of biological feature, and it's just a coincidence that the people who tend to like smoking, the same trait that causes them to like smoking also tends to cause lung cancer," he said.

Julia: Feels like kind of a reach, doesn't it?

Kenny: Yes. Yes it does. His claim is -- if he was right, which I'm sure he's not -- then that would mean that your choice of whether or not to smoke would not directly affect the outcome of your decision, the outcome of getting cancer. He suggests, "Well, in that case you might as well just smoke if you like smoking, because if you like smoking you're probably going to get the cancer anyway. Might as well get a little bit of pleasure along the way."

In that case, it looks pretty silly. But we actually see this sort of thing arising all the time. Whenever someone talks about not wanting to confuse correlation with causation.

Julia: Right.

Kenny: I think there's some famous studies of school choice initiatives, for instance, where you find that the students who get enrolled in charter schools do better than students who stay in the public schools. But in many cases it turns out that the students who try to enroll in charter schools do just as well as the ones who actually succeed -- just because their parents are motivated, and that's actually the bigger causal factor.

Julia: Right. Another example of this that I like is, some hospitals have higher death rates, and people often assume that this means that the hospitals are worse, but in fact they're better hospitals and they're just taking on riskier cases of people who are more likely to die anyway, and the hospitals are increasing their chance of surviving. Still, the base rate of death is just higher, because those are tougher cases.

Kenny: Right. In all these cases, it looks like there are ways in which your action is related to the outcome that don't seem to go through the effect of your choice.

The way that philosophers often put this problem -- there's a slightly different version which is a bit more science fiction-y, which perhaps brings out some of the worries more clearly, and the way that this raises challenges to standard decision theory. This is the one that's traditionally called the Newcomb Problem.

Say that you are at a carnival and you walk into a booth and you find a strange game being played there. There's a mad scientist who offers you the option of taking just what's in an opaque box -- or, she says, you can take that box *plus* a bonus thousand dollars.

However, as you were walking into the tent, her machines were scanning your brain, your body, and her computers predicted what you were going to choose. If they predicted that you would choose just the box, then she'd put a million dollars in it. If they predicted that you would take the box plus the thousand dollars, then she put nothing in. While you're deliberating and trying to figure out what to do, you learn that her predictions in the past have been fairly reliable, so a majority of the people that she predicts will just take the box, just took the box and got the million dollars. A majority of the people that she predicted would take the thousand dollars, she predicted that right and they only got the thousand dollars and didn't get the bonus million.

In this case, if we tried to set this up as a traditional decision problem, what you might think is that, well the state of the world, there's either already a million dollars in the box or not, so it looks like your choice is: do I take the box or do I take the box plus the thousand? You just have to figure out, well, what's the probability that the million dollars are there, and whatever that probability is, you're better off taking both, taking the bonus thousand.

Julia: Right. Well, it almost seems like you *don't* have to figure out what probability there is of a million dollars being in the closed box, because no matter what it is, you might as well take both, otherwise you're leaving money on the table.

Kenny: That's right. That's right, and that's the sort of reasoning that philosophers call Causal Decision Theory. They say look at your actions, see what it's going to

cause, assuming the state of the world is already fixed and then just look at the effects of your action.

Meanwhile, there's also a group of philosophers called evidential decision theorists, who say, "Well, but look at this, the people who just take the box, they end up rich. They get a million dollars. The people who take both generally don't." What you should be doing is looking at: what's the probability that the million dollars will be there, given that you just take the box, versus what's the probability of the million dollars will be there given that you try to take both. They say you should take into account your choice as part of your evidence for what the value of the action is. They say you should just take the one box.

I think a lot of people find it very hard to decide what's the right response in this version of the problem, this Newcomb Problem.

Julia: Actually, in my experience people have a very easy time deciding what the right choice is, and they can't understand how anyone would choose the other thing. Except people are evenly divided between. I think the originator, was it Nozick who came up with this problem, or this puzzle initially?

Kenny: I know that Nozick is one of the early places where it's discussed, but the fact that it's named the Newcomb Problem and not the Nozick Problem, suggests that there's someone earlier named Newcomb, but I don't think Nozick explains who that is.

Julia: Right! … Although I heard there's another law that states that scientific discoveries are never named after the person who discovered them. I forget what the name of this law is, but it's not named after the person who discovered that law.

Anyway, the quote from Nozick, at least, about the Newcomb Problem is that, "To almost everyone it is perfectly clear and obvious what should be done. The difficulty is that these people seem to divide almost evenly on the problem, with large numbers thinking that the opposing half is just being silly."

Kenny: That's right. I think the slightly different versions of the problem, though, do get different intuitions. Like in the smoking case or in the charter school case. Once people understand what the causal structure of the situation is, once they understand it's just correlation, there's no causation involved, most people go with what the causal decision theorists advocate. They say, if Fisher was really right, if smoking does not cause lung cancer and it's just that smoking is correlated with lung cancer, then you might as well smoke. I think most people go along with that intuition. I think there's different versions of the problem that push things much farther the other direction.

I think another classic problem that has some of the same structure is the Prisoner's Dilemma. For those of you who aren't familiar with that necessarily, though a lot of people probably are: Imagine that you and your twin have just robbed a bank and the police have caught you. They don't have enough evidence to convict you of robbing the bank, but they do have enough evidence to convict you of a minor crime on the side. Then they give you this offer, they separate you and the twin and they give you this offer: if you will testify against your twin, we'll drop this minor charge and, with your testimony, we'll be able to convict your twin. Now, of course, they're making the same offer to your twin. If you testify you get a year knocked off your sentence no matter what, but if your twin testifies then you get five years added to your sentence. What you'd really like is for your twin to not testify and for you to testify.

Julia: This is presuming you don't actually care about the well-being of your twin.

Kenny: That's right.

Julia: An important stipulation in the thought experiment that probably doesn't apply in real life.

Kenny: Yes, that's very important for the stipulation of this one.

In all of these situations we have the same sort of setup. There's some feature of the world which we are unaware of -- whether or not you have this biological lesion that causes a desire for smoking and cancer, or whether or not the million dollars is already in the box, or whether or not your twin is testifying. Then there's a choice that you have, and one of your options is better than the other regardless of which state of the world is actual, but one of the states of the world is highly correlated with one of the choices that you could make. This just undermines the setup that Savage assumes for decision theory initially, which is that the states of the world are independent of your action.

Julia: I agree that it feels like there's a difference between the Newcomb's Problem and the Prisoner's Dilemma. In the case of the Prisoner's Dilemma it feels like the state of the world is being determined concurrently with my decision, that if I decide to testify against my twin then sort of simultaneously my twin is deciding to testify against me. It feels like there's this tight linkage between what I decide in Prisoner's Dilemma and what my twin decides.

Whereas in Newcomb's Problem, it feels much more like the state of the world is already determined. There is already a million dollars in that box or there isn't. And my decision can't change that, because that would be like reverse causality, going backwards in time.

It's a little weird, because we've stipulated that the mad scientist is really good at predicting what I'm going to do, it's almost like there is reverse causality -- because if I were to decide, "Well, I'll just take the box," then that means that probably the scientist decided not to put money in the box. It's not really ... My choice can't really determine whether there's money in the box because the money already happened or it didn't.

Kenny: That's right. I think there's a few related issues here all around the idea of causality.

One reason why I, for a long time, hated this issue was because I thought it's all about causality, and I think that's just this messy thing that probably isn't really real in the world anyway in some sense. And that if we could just avoid that we could just go back to Savage's decision theory and everything would be nice.

I think it's not just the temporal order that's relevant here. So we get this temporal order that's like the Prisoner's Dilemma, where the state of the world is being decided simultaneously, or even after your decision, just in a classic gambling case. We're about to flip a coin and I'm deciding, should I bet on heads or should I bet on tails. Yet, of course, my decision of whether to bet on heads or tails isn't going to affect the outcome of the coin.

Similarly, in the Prisoner's Dilemma, my twin is in a separate room from me. He doesn't know what I'm doing, and my choice isn't going to cause him to do anything different. He's already who he is, he's his independent person, my action can't affect him any more. Yet, there's still this suspicion that maybe somehow my action's got to be strongly correlated with his in a way that's different from what's going on in the Newcomb Problem.

Julia: Another aspect that we didn't touch on with the Newcomb Problem so far, is this notion of a perfect predictor, or a really good predictor. In some versions of Newcomb's Problem it's stated that the mad scientist -- well, sometimes it's a mad scientist, sometimes it's a super intelligent alien, sometimes it's an artificial intelligence that's just got a really, really good prediction algorithm -- and it's stated that this predictor is 100% accurate. That, whatever you do, it knew that you were going to do that.

Some people, I think including me when I first heard this version of the problem, think that the reason that this seems like a paradox is just because it's assuming this impossible thing. That there's no way to actually perfectly predict with 100% certainty ahead of time what someone's going to do. Therein lies the ... That's where the shell is hiding in this shell game, or whatever.

Kenny: I think this is actually not that essential. I think ...

Julia: Yeah, that's what's interesting. Go on.

Kenny: If the predictor's only 60% accurate, it's still very likely that the people who take the one box end up better than the people who are taking both. I heard, actually, that the philosopher Dave Chalmers did this once at a party. He had invited a bunch of philosophers, and he knew them all, and he offered them all a version of the Newcomb Problem, of course not with a million dollars, smaller prizes. He apparently got 60 or 70% of the predictions accurately.

Julia: I actually had a philosopher do this to me on a date once. I two-boxed. I mean, I took both the money and the empty box. Sorry, the opaque box. I just gave it away! It was empty because he predicted ahead of time, knowing what he knew about me, that I would take both. And he was right.

Kenny: Yes. That's right, and I think that it doesn't rely on the predictor being 100% accurate. Though I think if the predictor's 100% accurate, that will strongly push people towards one-boxing. And I think it raises this deeper worry, that gripped me for a long time, that there's something fishy about this as a decision problem, that it should be something that we somehow exclude from the action of what we're considering.

I've been thinking about this stuff lately, and I'm working on a paper discussing these issues. The thing that I've been noticing is that we have these different problems that are structurally similar to each other, and I think philosophers have assumed that whatever your theory of rationality says about the Newcomb Problem, it should say the same thing, or the parallel thing, in the smoking case, or in the Prisoner's Dilemma.

What I'm interested in is maybe there's actually important subtle differences between these problems, based on something like this causal structure. What I've been doing is I've been thinking about these things in terms of trying to understand, what's the full causal structure of what's going on here?

I think, actually, there's a related puzzle that some philosophers have thought about, on the theory of action and intention, that I think actually has enough similarity that I think can shine a bit of light here. This is the Toxin Puzzle, from Gregory Kafka.

The way this one works is that a billionaire is coming up to you with a very strange offer. He places before you a vial of toxin that, if you drink it, will make you painfully ill for a day, but will not threaten your life or have any lasting effects. What he says is, "I'll pay you a million dollars tomorrow morning if, at midnight tonight, you intend to drink the toxin tomorrow afternoon." He emphasizes you don't actually need to drink the toxin, in fact the money will

already be in your bank account hours before the time for drinking it arrives, if you succeed in having this intention.

Julia: Presumably he has advanced enough neuroscience and brain scans that he can actually tell whether you're intending, sincerely, to drink the toxin.

Kenny: Exactly, that's right. In fact, he's going to leave before you ever have a chance to drink it. Now, what Kafka is raising this problem for is, he suggests: Well, think about your situation tonight. I start thinking, "Let me plan to drink that toxin tomorrow," but I know that once I wake up in the morning I can check my bank account, I'll see if the money's there or not, and at that point I feel like I have no more reason to drink the toxin. So I know that, in some sense, no matter how sincerely I try to intend tonight, I know that tomorrow I'm just not going to do it, at least I suspect that.

Julia: I will have no reason to do it.

Kenny: Yes. That's right. Though even, I think some people might claim there's controversy around that, but it seems to me, and it seemed to Kafka, that you'll have no reason to do it tomorrow, and you can predict that now. Kafka's saying that intention is just not compatible with a full belief that you're just not going to do it.

Julia: Right. Is this related to the Parfit's Hitchhiker Problem?

Kenny: Yes, that's right.

Julia: I'll briefly state it for our listeners who haven't heard it. The idea is that you're stranded out in the desert somewhere, you're out of water so you're going to die soon if someone doesn't rescue you. Then, luckily, a car happens to be passing by and the driver is -- you can stipulate that he's really, really good at reading people's minds, not literally ...

Kenny: Or intentions.

Julia: Yeah, intentions, just from listening to their body language and expressions. He says, "Well, I'll drive you to town, if you can give me $100 once we get to town," because obviously you don't have any money on you now, you're out in the desert without a wallet. You wish that you could commit to giving him $100 when you get to town, but you realize that once he'S brought you to town, he's got nothing on you. You don't have to give him the $100. So you're not going to have any reason.

This is assuming both of you are perfectly selfish agents, which is often a stipulation of these problems people have trouble with, but assuming that…

You realize that you really, really wish that you could credibly commit to paying that $100 once he's brought you to town, but you can't do that. Even if you say yes now, the driver can tell that you know that you won't actually have a reason to do it. And that therefore you don't truly intend to pay him off. And so you're left stranded in the desert. It's sort of too bad that this Causal Decision Theory seems to leave people in the hitchhiker's position with no way to actually save their lives.

Kenny: That's right. I think all of these puzzles have the same sort of payoff structure. There's one action that, regardless of what the state of the world is, guarantees you a slightly better outcome. In this case, it's not drinking the toxin, or not paying him when you get into town. The state of the world, which is whether you get the million dollars from the billionaire, or whether you get the ride back to town, is strongly correlated with what you're going to do, and that's a much bigger effect than the other one. It's not this direct causation that your action has on your outcome, but it's somehow an indirect correlation with a much better outcome.

The question is, what can rationality say about this? If there's an action that's guaranteed to, regardless of how the world actually is, make you slightly better off, but it will be much better for you if you hadn't made that decision, what's rationality going to say?

What I say is that the causal decision theories have something right. They say we should be looking at the causal structure. We shouldn't just be treating your action as a piece of evidence that you get. Obviously, in all these situations, if someone told you, “You're going to take the one box,” or “You're going to give him the money when you get to town” -- those would be pieces of good news for you.

And I think even the causal decision theorist says, if you discover that you are going to make one of these choices that the causal decision theorist says is the bad one, the causal decision theorist says, that's still good news. They say the evidential decision theorist is just looking at your actions as a special type of news you get about the world, and somehow ignoring the fact that what you're doing when you're acting is something different from just learning about the world, you're actually *making* the world a certain way.

The causal decision theorist says, well, rationality for action consists in making the world do the good thing for you, even if that might be bad news to learn that you are doing the thing that will be good for you. I say that because these

different cases are different enough that many people actually get different intuitions on them. I think many philosophers have tried to develop versions of Evidential Decision Theory, or Causal Decision Theory, that can explain why we should cooperate in the Prisoner's Dilemma, but smoke in the Smoking Lesion, and so on. They try to develop complicated theories.

But I say we should just look more closely at what the causal structure is. I say, in the Smoking Lesion case, what's going on is there's this fundamental thing, which is your genetics, and the genetics has an effect on the outcome. It has an effect on whether or not you get cancer. Then it also has an effect on what sort of person you are. What sort of person you are determines what your psychological state is at the beginning of this game. And then that determines, in some sense, what decision you do, and that decision then sometimes determines your act of smoking, which also gives rise to some features of the outcome.

Whereas in the case of, say, the Newcomb Problem, your genetics don't even matter, and even your psychological character, in general, isn't the big thing. What's important is *what your psychological state was at the moment that you walked into the tent*, because that's when you get scanned and that's the thing that has an effect.

In the Toxin Puzzle, it's different still. It's not what your psychological state was at the beginning of the puzzle. It's: What's the actual decision you make tonight at midnight? In all these cases, it's not the act that's causing this other thing, it's something that is part of the determination of your act.

Julia: So, we've been analyzing the problem from the wrong decision point, or from the wrong choice point, or something?

Kenny: Maybe, or at least there's many different choice points. I think what's going on is that the act itself, the decision to do the act, the psychological state that you're in at a given time, and your character as a person -- these are all things that are, to varying degrees, in our control. Though, also, to varying degrees, they're not in our control.

I think, when we analyze the problem at a different level, we get a different answer. What I think is correct is: If you can, right now, try to shape yourself into the sort of person who, whenever you're faced with a Newcomb Problem, will just take the box. That's a good move.

Even the causal decision theorists will say that. Because whatever actions you're doing right now to make yourself into a one-boxer for future Newcomb

Problems, that's going to actually cause future predictors to predict you'll one-box, and therefore give you the million dollars.

Julia: It sort of feels to me like some moral intuitions, or moral conventions, in human society, have developed as a hack to make hitchhiker's dilemmas, and prisoner's dilemmas, and Newcomb's problems work out well. We have this notion of following through on what you promise you'll do, or sort of doing unto others as you would want others to do unto you, even if your actions don't cause other people to treat you better. If living up to those moral standards is more important to you than getting the million dollars -- or even living, if you're the hitchhiker stranded in the desert [*JULIA's NOTE: I misspoke here. Following through on your promise doesn't cost you your life in the Hitchhiker's problem*] -- then that makes you the kind of person who would just take the one box, or who would give the $100 to the person who saved you, even if you have no selfish reason to do so at that point.

Kenny: That's right. Although, as many people have observed, many of these social features that tend to reinforce this sort of behavior depend on the fact that we have continued interactions with each other. This is what game theorists often talk about as the *Iterated* Prisoner's Dilemma -- that if I know I'm going to be playing Prisoner's Dilemma with the same person multiple times, over and over, then that gives me a strong incentive to cooperate. So that my action now, of cooperation, can cause better results later on.

Kafka, in his paper about the Toxin Puzzle, sort of anticipates this. And he notes that one way you can make yourself drink the toxin is if you make a side bet with your friend where you arrange, "If I don't drink the toxin, you should steal my money and give it away." If you can arrange for that, that's going to help you guarantee to drink the toxin. Kafka says, the billionaire has anticipated this and written into the contract that you're not allowed to make any side bets, you have to make this intention without any side bets. Without any of these sort of social sanctions that could help you stick with that intention.

Julia: The iteration of these games makes them less interesting in one sense, because it sort of transforms them all into Causal Decision Theory problems. Because your actions in this game can cause the state of the world in the future games, which dominates the overall outcome for you. That's a little less interesting, in a sense.

Kenny: Right, it's less interesting for this particular issue about what rationality is.

I think one assumption that's gone into a lot of this is that people assume that there should be one right answer. That either rationality should say that the right action is one-boxing or two-boxing or whatever, and then rationality will tell us

in a more general sense, "Well, then a person with rational character is the sort of person who does the rational act," or vice versa. Maybe that the right level at which to analyze rationality is at the level of character formation, and then the rational act is just that which the rational character would carry out.

I see some analogies here to -- there's been over the centuries, a lot of discussion in moral philosophy about whether the right way to analyze morality is at the level of the consequences of your action, or at the level of the action, or at the level of virtue or character.

Julia: You might have already intended to refer to this, but another level that I'm personally sympathetic to is at the level of the rule. Following a certain behavioral rule might be good in general or in the long run. According to this take on utilitarianism, or moral philosophy, you should follow the rule even in cases where, in this specific case, it gives a worse outcome.

Kenny: That's right, because it's a rule, that if followed generally would result in better outcomes.

Julia: Right, exactly. And there's some setup in which, if you try to sort of game the rule and abandon it in the cases when it gives a worse outcome, you are thereby dooming yourself to worse outcomes overall.

Kenny: Yes, that's right. I think thinking about it actually in the contrast between rule utilitarianism versus act utilitarianism is actually the most useful way to think of the parallel here, because we're still analyzing all these things based on what is the outcome that you get -- as opposed to many of these moral theorists who think that actions or virtues are the right level at which to think. They say that virtues are even primary to the outcome. That what makes an outcome good is that it's the sort of outcome that a virtuous person would bring about for something. They actually even deny the analysis through outcome.

I think here, in decision theory, most people are committed to some sort of outcome-based analysis, some sort of consequentialism as they call it in moral theory. The question is, still, at what level does this consequentialism analyze you? Is it at the level of the action? At the level of the decision? At the level of the psychological state at a time? At the level of your virtuous character? Which one of these consequences we are most interested in analyzing.

Julia: Right. I like that because it addresses a common objection that I hear to one-boxing in the Newcomb's Problem, which is, "Well, look, this just happens to be a weirdly constructed case in which the mad scientist or the artificial intelligence, the predictor, has chosen to reward people for being irrational." That you could construct similarly artificial cases where you get a million dollars for truly

believing that the sky is green -- and that's just, "Too bad for rationality in this particular case," but that doesn't mean that rationality is wrong or flawed in some way. It just means that in a few constructed, contrived cases, by stipulation, you're going to be worse off by being rational.

Which… I feel the pull of that argument. But at the same time, there does seem to be something wrong with a rationality that leaves you worse off in a systematic and predictable way.

Kenny: That's right. I see many people associated with artificial intelligence and related issues -- perhaps many of the listeners of this podcast, who have been interested in versions of decision theory that advocate one-boxing, and advocate giving the money in the Parfit's Hitchhiker case, and so on -- On these grounds, saying that a theory of rationality ought to be the one that gives you the best outcomes overall, and not out of evidential grounds. Not by saying we should ignore the causal structure of the action.

I think what they're doing is, I think they're still thinking that rationality should act at one of these levels, and that then, once you've determined what a rational character is, then we can understand what a rational act is on the basis of a rational act is just one that rational characters would do.

Now, what I'm thinking is that perhaps there's actually just a deep tragedy in the notion of rationality. Perhaps there just is a notion of rational action, and a notion of rational character, and they disagree with each other. That the rational character is being the sort of person that would one-box, but the rational action is two-boxing, and it's just a shame that the rational, virtuous character doesn't give rise to the rational action. I think that this is a thought that we might be led to by thinking about rationality in terms of what are the effects of these various types of intervention that we can have.

Julia: Still, isn't there some way to... Doesn't the rational character trump the rational act, in some sense?

Kenny: I don't know. I think what any of these one-boxers will tell you is that the right thing to be is to be a one-boxer, but you really wish that by accident you'll two-box even though you are, deep down, a one-boxer. You'd still get the thousand dollar bonus, and so you'd really like the action you take to be the one that goes against your character, even though you're not that sort of person.

Julia: It reminds me of a sort of modern fairy tale -- I can't remember where I read it. The hero in the fairy tale is presented with this opportunity to sacrifice his own life in order to save a person, or the world, I'm not sure. And he takes the step to sacrifice his life, but then is disturbed when it turns out that he does, in fact,

have to lose his life. He's used to the stories in which, when you do the heroic, noble thing, you're rewarded for it by, "Oh yay, actually you're saved after all. You just had to prove that you were in fact willing to sacrifice your life." That's not actually how it turned out.

Kenny: That's right, and I think perhaps rationality just has that sort of tragedy to it, that there's one notion of rational character, there's one notion of rational decision, there's one notion of rational action -- and they don't all necessarily line up. If you think that they should line up, then the question is: Well, which one is the most fundamental thing to have? Is it the notion of rational action, or the notion of rational character?

I think if people go for what they call Updateless Decision Theory, or Timeless Decision Theory, they say we should go as far up the causal chain as we can. It's the character that determines everything else. And once we understand what rational character is, then we understand what everything else is.

Julia: You alluded to Updateless Decision Theory and Timeless Decision Theory. I don't know how brief you can be, but can you attempt to summarize those alternatives to Causal and Evidential Decision Theory?

Kenny: Unfortunately, I don't think I understand them in enough detail to do that. I think that the approximation that I mentioned so far is useful: You should always act in the way that you wish you would have initially committed yourself to act. That's approximately what's going on. That when the driver brings you back to town, you should get the $100 for him because that's what you, yourself, back in the desert, would have wished. Even though now you don't. Similarly, you should take the one box in the Newcomb case, because that's what you would have wished yourself to be committed to before you entered the tent.

In the Smoking Lesion case, they say it's not like deciding not to smoke is going to have changed your DNA or something like that. Somehow, that one, the causal chain begins outside anything that you have ... Well, not in your control, is the right way that they would want to put it, but it's the way that I think of it. That somehow I don't determine my DNA, I only enter the picture once we get to the point of my character or something like that.

Julia: The thought experiment that makes this kind of algorithm seem absurd, to me, is something called the Counterfactual Mugging. Where someone comes up to you and says, "Well, I decided that I would flip a coin and if it came up heads, I would give you a million dollars, and if it came up tails, you would give me $100. And I'm just letting you know, the coin came up tails, so you owe me $100."

Kenny: Also I think it's important to set up that, "I decided to do this only because I know you're the sort of person that would give me the $100."

Julia: Right. That you're the sort of person that would follow the terms of the bet that I came up with. You also have to assume this person is honest and not just making up this dumb story to get you to give them $100.

According to the Updateless Decision Theory approach, you should just pay the $100 because that is the algorithm you would have wished that you would follow. Because otherwise the person would never have been willing to flip the coin and risk the outcome of giving you a million dollars. But that just seems so crazy.

Kenny: That's right. I don't get on board with their view. And to me, I just think that the right response here is to say that there's some sort of tragedy in rationality. That there's multiple levels of rational analysis, and there is one level of rational analysis at which being the sort of person who would take part in this, is perhaps a sort of character virtue to have. But when you're actually faced with this, the right thing to do is not to pay the money. I don't know if there's any way to square these two conflicting notions of rationality.

I think that it's important for all of us that there is some notion of causation involved here. There is some notion of self-determination. That for all these cases, if we're not talking about me deciding to take the one box or two box, but if we're talking about just that, if you have one sort of muscle spasm you get a million dollars, if you have a different sort of muscle spasm you get a thousand dollars -- there's no question of rationality at that point.

Julia: Then it's just billiard balls bumping into each other.

Kenny: That's right. Then, I think, this is what's always made me uneasy about the puzzle to begin with: That in all these cases we're accepting that there's some way in which the causal structure of the universe allows for our decisions to be predictable. Once we allow for our decisions to be predictable, I think this raises worries about whether our decisions are really decisions in the relevant sense overall.

It seems to me that when you're in the Newcomb Problem, it sort of destroys this illusion that we have that we're acting freely in the world. Once we destroy this illusion of acting freely, the notion of rationality stops making as much sense as we might have thought initially.

Julia: Right. In Newcomb's Problem, it feels like you have unusually less free will. Because the causal structure of the situation is such that your choice has already been determined in some sense, by the mad scientist.

Kenny: In this case it's determined. The mad scientist has determined it epistemically. But there's some feature of the world that she's reading that causally determines it in some way.

Julia: It's causally determined both your choice and the prediction of your choice.

Kenny: That's right.

Julia: Yet, even though it's starker in Newcomb's Problem than in general, it's not fundamentally different than how the universe works in non-Newcomb cases.

Kenny: That's right. I think one way to think about this is, there's this way of understanding causation that's become popular over the last few decades, due partially to the work of Judea Pearl and the computer science department at UCLA, and partly to the work of Scheines, Glymour, and Spirtes, who are in -- I believe they're all in the philosophy department at Carnegie Mellon, but they may be in statistics as well.

And that is trying to understand causation through what they call these causal graphs. They say if you consider all the possible things that might have effects on each other, then we can draw an arrow from anything to the things that it directly affects. Then they say, well, we can fill in these arrows by doing enough controlled experiments on the world, we can fill in the probabilities behind all these arrows. And we can understand how one of these variables, as we might call it, contributes causally to another, by changing the probabilities of these outcomes.

The only way, they say, that we can understand these probabilities, is when we can do controlled experiments. When we can sort of break the causal structure and intervene on some things. This is what scientists are trying to do when they do controlled experiments. They say, "If you want to know if smoking causes cancer, well, the first thing you can do is look at smokers and look at whether they have cancer and look at non-smokers and look at whether they have cancer." But then you're still susceptible to the issues that Fisher was worrying about.

What you should actually do if you wanted to figure out whether smoking causes cancer, is not observe smokers and observe non-smokers, but take a bunch of people, break whatever causes would have made them smoke or made them not

smoke, and you either force some people to smoke or force some people not to smoke.

Obviously this experiment would never get ethical approval, but if you can do that -- if you can break the causal arrows coming in, and just intervene on this variable and force some people to be smokers and force others to not be smokers, and then look at the probabilities -- then we can understand what are the downstream effects of smoking.

In some sense, these causal graphs only make sense to the extent that we can break certain arrows, intervene on certain variables and observe downstream effects.

Then, I think, in all these Newcomb type problems, it looks like there's several different levels at which one might imagine intervening. You can intervene on your act. You can say, imagine a person who's just like you, who had the same character as you, going into the Newcomb puzzle. Now imagine that we're able to, from the outside, break the effect of that psychology and just force this person to take the one box or take the two boxes. In this case, forcing them to take the two boxes, regardless of what sort of person they were like, will make them better off. So that's a sense in which two-boxing is the rational action.

Whereas if we're intervening at the level of choosing what the character of this person is before they even go into the tent, then at that level the thing that leaves them better off is breaking any effects of their history, and making them the sort of person who's a one-boxer at this point. If we can imagine having this sort of radical intervention, then we can see, at different levels, different things are rational.

Julia: Right, yeah. That disambiguation really helps resolve some of the uncomfortable feeling of reverse-causality that you get from the original formulation of the problem.

We're almost out of time, but I want to make sure that we talk, at least briefly, about implications of this field of decision theory. In one sense you could say, “Well, in the real world,” as you were saying Kenny, “any kind of Hitchhiker or Prisoner's Dilemma type problems that arise are, in practice, repeated games. So they turn into causal problems. Or even if they're not repeated games, in practice society has developed these patches for the problems, in the form of moral standards that we feel compelled to follow.”

So maybe in practice, it doesn't really matter what the rational choice turns out to be -- because we've sort of solved the problem in practice.

Kenny: There's also the second sort of games that we get where we have an observational study, and there's correlation but there's no causation. Here, most people just say, do the causally good thing, ignore the correlation. Once we understand that it really is just a correlation that isn't affected by your action.

Julia: Right. So the question is, do you think that are ... What do you think is the motivation for figuring out how to think about these kinds of problems?

Kenny: For a long time, what I thought was: we should ignore these problems. They're pseudo-problems. Because they only arise in these cases where we have to be confronted with the lack of control that we have in the world, that the world is really deterministic and our decisions are just part of the world, and there's no shoulds about it, there's just what will happen.

Nowadays, I'm thinking that perhaps the better thing to think is that in showing us how rationality is tied into questions of causation, and free will, and determinism.

I've always been happy with the idea that free will, in the sense that we really care about it, is compatible with the world being a deterministic place. And the question is just, "Are my actions being determined by the states that I identify with?" That is: my character, who I am as a person. As opposed to when my actions are caused by the kidnapper with a gun to my head, or are caused by the electrical stimulation of my muscles by an outside psychologist. There's multiple different kinds of causation, but the causation that matters to me as an agent is the kind that goes through me as an agent.

And there's still sense to be made out of that. There's questions about what that means for the notion of practical rationality here. I think that's really what these puzzles are pushing us towards: that the notion of rationality and the notion of the ways in which free will is compatible with determinism are going to have to ... There's going to be some complex interaction here. And maybe rationality falls into multiple different types of rationality -- one for actions, and another for psychological states, and another for character formation.

Julia: Right, excellent. Well, we are actually over time now, because I lost track of time because I find this topic so interesting. I will have to wrap up this part of the podcast -- and we'll move on to the Rationally Speaking Pick.

[musical interlude[

Julia: Welcome back. Every episode we invite our guest to introduce the Rationally Speaking Pick of the Episode. It's a book, or website, or something else that tickles his or her rational fancy. Kenny, what's your pick for today's episode?

Kenny: My pick is the book *Thinking Fast and Slow* by Daniel Kahneman.

Julia: Yeah, that hasn't been a pick yet, surprisingly.

Kenny: That's really surprising to me.

Julia: I could be wrong, but I don't think so.

Kenny: What this book is all about is the different ways that our mind works. And the title is a reference to the fact that there's one set of things that we do when we just work out of habit, and another set of things that we do when we consciously reason about the world. Each of these has their own sort of understanding of the world and their own rationality.

I think his book also gets at some fundamental problems in understanding what it is we even want. He talks about the contrast between the experiencing self and the remembering self. He says you can go on a vacation and go on this hike through the forest to get to this mountain peak, and for hours and hours you're fighting mosquitoes and your muscles are sore and you're suffering. Then when you get home, you think, "That was the most amazing thing, because I struggled through this and I got to this amazing view."

On the one hand, if we say, “Our goal is to maximize the great experiences that we have,” you spent hours, and hours, and hours on that hike suffering. And over the rest of your life, you're probably only going to spend an hour or two total reminiscing about this. Yet, somehow we think the remembering self is the one that gets to decide what our next vacation is going to be.

We, as psychologists, philosophers, decision theorists, are put in this difficult position: what is the thing that we really care about? Do we care about people leading good lives moment by moment, or do we care about what people say they care about? There's many other, I think, really interesting dilemmas that arise from this book.

Julia: There absolutely are. And I think all of the parents, or potential parents, in the audience might see some interesting implications of the hike example. I'll just leave it at that.

Kenny: Yes.

Julia: All right, we're all out of time. Kenny, thank you so much for joining us. I thought that was a very interesting discussion.

Kenny: Thanks for having me on.

Julia: I encourage our listeners to check out Kenny's research, which we'll link to his page on the site. As well as to his pick, *Thinking Fast and Slow*, by Daniel Kahneman. And to download transcripts of the episode if you so choose.

All right, this concludes another episode of *Rationally Speaking*. Join us next time for more explorations on the borderlands between reason and nonsense.