Rationally Speaking #186: Tania Lombrozo on, “Why we evolved the urge to explain”

Julia: Welcome to Rationally Speaking, the podcast where we explore the borderlands between reason and nonsense. I'm your host, Julia Galef, and I'm here with today's guest, Professor Tania Lombrozo. Tania is an associate professor of psychology at the University of California Berkeley. She's also an affiliate in the philosophy department and the director of the Concepts and Cognition Lab. Tania, welcome to the show.

Tania: Thanks for having me, Julia.

Julia: So, Tania's research is situated at the intersection of philosophy and cognitive science, and asks questions both about how humans reason and how they theoretically, normatively, should reason.

It is all utter Julia-bait. So it was very hard to pick an area to focus on for today's episode.

But the aspect of your research, Tania, that I was hoping to focus on today is the research on explanation. People are constantly reaching for explanations about the world. Like, a tragedy happens and we want to explain, "Why did this happen?" Or we want to explain patterns that we see in the world, like why are some people successful and other people aren't.

So, one of the many things that Tania's work looks at it what kinds of explanations do we reach for. Like, for example, do we have a bias in favor of simple explanations over complicated explanations? And to the extent that we have this bias, is it justified? Does it make us more accurate? And if not, does it do other things for us? And so on.

So, those are all questions that we will get to this episode, I hope. But maybe to kick things off, Tania, you can just kind of situate the study of explanation in the broader project of cognitive science. Let's say we've solved all the questions we're interested in with respect to humans and explanation. What would that open up for us? Or what other mysteries would that solve for us about human psychology?

Tania: Right. Well, I think one way to motivate why we should care about explanation and what it can tell us, is to think about the ways in which explanations actually may be a little bit mysterious. So if you think about the sorts of things that we want science to do and the sorts of things that we want our everyday cognition to do, we want it to be able to support prediction and control. It's very clear why those things are valuable.

Julia: Control, like controlling the world around us?

Tania: That's right. We want to be able to intervene on the world to make particular outcomes that are desirable come about. We want to be able to intervene on the world to prevent things that we think are not desirable, and so on. And that I think is true for both science, the way we approach the world of scientists, but also just the way that the human mind works. We need to be able to predict and control the environment in various sorts of ways.

But we also are really, really motivated to *explain*. And it's a whole lot less obvious what explanation is doing for us. Why is it that we bother engaging in this activity of trying to make sense of something or trying to understand why it happened, of asking why and trying to find good answers to those why questions?

I think an answer to that that's attractive, and that I've advocated and many other people have as well, is that in some ways, something about seeking explanations and engaging in this process of trying to explain is instrumentally valuable. Because it's going to downstream help us predict and control better. If we can understand something, we're gonna be in a better position to make inferences related to that in the future.

So, I think if you think about explanation playing that very fundamental role in the very mechanisms that allow us to predict the world, the very mechanisms that allow us to know how to make effective interventions in the world to bring things about or to prevent outcomes, then it's really clear why it should be super important -- for understanding both the human mind, but also the scientific process. If we could understand explanation, we're gonna understand learning, we're gonna understand inference, we're gonna understand how the process of discovery works.

Julia: And the descriptive aspect of this -- looking at what are the heuristics that the human brain seems to be using when we reach for explanations, or come up with explanations -- can be useful for the purpose you're describing, because we might be able to discover ways in which we're systematically erring. Like, types of explanation that maybe aren't ideal or aren't ideal in the modern context that we are now using them in, even if they were optimal when they evolved.

Tania: Yeah. I think that's right. I think if we understood how explanation supports good learning and good inferences, we could presumably foster that better, but also by understanding the cases where we tend to get things wrong, then maybe we can apply good correctives.

Julia: Right.

Tania: So, for example, if we discover, and there's some work I've done for this, that we favor simple explanations more than we should, then that's a case where recognizing that explanatory preference can actually lead us to be better reasoners, by making us more wary of that kind of a preference when it shouldn't manifest.

Julia: Right. One thing that I've noticed personally about the usefulness of explanation is that it just helps you personally understand something so much better.

I used to teach these classes on reasoning and decision making and judgment. The concepts and the techniques that we were teaching were relatively complicated. And often talking to people after the classes, we would realize, "Oh, gee, they really didn't understand what we were saying." They *seemed* like they understood. They were nodding their heads. They *felt* like they understood. But in talking to them, we realized they really didn't.

So then we started adding in this paired tutoring session, where we had people explain the classes to each other. And this -- even though there was no real content in this class, it was just people explaining things to each other that we'd already taught them -- it became by far the most popular class that we taught. And in the feedback forms, people kept saying things like, "Wow, I didn't actually understand the classes until I tried to explain it to other people and I realized what was missing," and so on and so forth.

Tania: You reinvented a discovery from the cognitive science and education literature, which I think is really an important one. I think it fits a lot of people's experiences that when you try to teach something, you often come to understand it better or realize that there are gaps in what you thought you knew.

But actually, in the research literature, there was an influential paper in the late 80's where the researchers were trying to figure out what it was that made some students much more effective learners than others. In some of the early studies, what they looked at was the patterns in study behavior of students going through physics problems and physics textbooks. And they found that some of them were doing this thing which looked a whole lot like explaining to themselves. They were just doing it spontaneously, as they worked through the problems and thinking, "Okay, why does that step follow? Why is this the case?" And so on.

The students who were doing more of that self-explanation were also doing better on post-tests, showing that they had learned more from this training.

Julia: Interesting.

Tania: So, initially those studies were just correlational. They just found the people who are learning better-

Julia: Like, maybe smart people just do this, and also smart people do better on the tests?

Tania: Exactly. So, the next step was to look at that experimentally. So, what they would do is do these studies where they would give people some sort of a learning task. Like, some of the early ones involved students learning about the circulatory system and how the heart works. They would then randomly assign then to two conditions and have half of them explain to themselves as they studied this, and have the other half do something else.

There's now been many, many years of these studies, so the control conditions have varied. Sometimes they have them just matched for time, but don't give them other instructions. Sometimes they have them think aloud. Sometimes they have them go through the materials twice. There's lots of variance on what you compare the group-

Julia: The explanation condition to?

Tania: Exactly, the self-explanation condition to. But what you find in most cases is that those who were prompted to self-explain do significantly better on measures of learning afterwards than those in the control condition.

Julia: So interesting. You know who else has independently discovered this phenomenon? Computer programmers, who have -- there's this phenomenon that I learned about, after we started having these tutoring sessions, in the computer programming world called “rubber ducking.”

Whoever started this trend had a rubber duck sitting on his desk, I guess for sort of decoration, or companionship or whatever. The way this started was programmers would encounter a bug in their code. They would notice their code wasn't working, figure out there was a bug, couldn't find it, and then would try to explain to a fellow coder what was happening and where the program was failing. And before they could even ask for help, they would interrupt themselves and go, "Oh! I know what it is!" And then they would go solve the bug.

But this was somewhat inefficient because it requires another person to come over to your desk and sit there while you explain it to them. So, this coder started explaining the problems to his rubber duck sitting on his desk instead -- and found that that worked basically just as well. So, he would start talking aloud to his duck about the problems with his code and then he'd solve his problems.

Tania: Part of what I like -- I didn't know that example, and part of what I really like about it is it illustrates something that I think is really cool about this process, which is that usually the way that we think about learning working is that you've got some sort of new information from the external world. You observe something that you hadn't observed before, somebody corrects your mistake -- those are the canonical cases of learning.

But in these cases where you learn by explaining to yourself, the only sort of new information you're getting is information that in some sense was already in your head.

Julia: Right.

Tania: And so what's nice about the rubber duck case is the rubber duck's not giving you any feedback. The rubber duck is not telling you when you're getting it right or when you're getting it wrong.

Julia: He's not even nodding, or looking confused, or anything!

Tania: Exactly. Nonetheless, there seem to be some benefits, at least in some cases, right? So, what's cool about that is I think the benefits are not coming from those standard learning mechanisms that we think about which derive from the feedback we get.

It's a phenomenon that in my lab we've been calling learning by thinking, in contrast to learning from observation, which I think is the more canonical case, where what you're observing is in some sense outside your head. It comes from another person. It comes from a textbook. It comes from the natural world.

But in cases like explaining to yourself, in cases like thought experiments in science, in cases like mathematical proofs as well, there's a sense in which it seems like you can learn something genuinely new *even though* you're not getting new input from the external world.

Julia: This was a big update for me a few years ago. I used to get into arguments, especially with people from the humanities, because I didn't see how you could learn from fiction. People say this about novels, about fiction all the time, like, "You learn about the world and learn about human nature and what it is to be human." I kept objecting, "But it's *fiction*. So whatever new things you think you're learning aren't actually justified, because they were just made up. They're not necessarily, they *may* be representative of the world, but you don't know, because the author wasn't optimizing for representativeness. He was optimizing for a good story."

And I eventually decided that I was at least partially wrong. I still think there's some bias that comes when you sort of form impressions about the world from fiction, as we intuitively do, but I still think I was significantly wrong because what fiction often does is it shines a light on things you already kind of knew, but hadn't been thinking about. So, the updates you're making about the world can in fact be justified, but you just hadn't put the pieces together that you already had, until the fiction highlighted them.

Tania: Yeah, I think that's right. I think that the toy example that I some times give people, to just have it be easier to think about, is the case of deductive reasoning. So, you might know P, and if P then Q, and just hadn't yet derived the conclusion that Q follows. You had those two pieces of information, but they were sort of isolated in the way that you represented them in your mind. And so from your armchair, maybe in the course of reading a novel, you put those together and you derived Q. And in that case, presumably, it is justified. It's justified because it's the deductive argument. It's just that the process that led you to that conclusion didn't involve a new piece that wasn't already in your head.

But I think there's harder cases to think about. I don't know if you've thought about this in the context of fiction, but a lot of scientific models are themselves, in some sense, fictional. We make lots of idealizing assumptions that we know are not true. We assume that something is a perfect vacuum. We assume that something has no friction. Or we assume that someone has infinite time. And whatever the assumptions are that we built into our model. So, in some sense we know that they are false, and yet there are cases in the scientific process where it seems like we use those idealized, fictionalized models to draw inferences that we hope to tell something about the real world.

So, I think there are these more subtle cases where maybe we can use this intermediate step as sort of a tool, this model that is in some sense not true or accurate, but it nonetheless helps us in drawing inferences.

Julia: Right. I guess that is a trickier or more complicated case, because it's not purporting to be data. It's a new framework that we didn't have before, that is useful, but not true, that is helping us interpret the data we do have. So, it's complicated.

But one thing I wanted to ask about “explanation for the purpose of understanding,” was why that works. It feels sort of intuitive, that of course you will understand something better if you explain it to people, but I could imagine that there exist different mechanisms for that.

Tania: And almost certainly there are multiple mechanisms that operate in parallel. So I'll tell you some of the ones that there's good evidence for.

Julia: Cool.

Tania: One thing that happens when you explain is that you appreciate you maybe didn't know as well as you thought you did. So, if there's a big gap in your understanding, then trying to explain it might bring that to light. So, in some sense, what you're learning is what you didn't know.

Julia: Right. And that's one thing we saw happening in our tutoring sessions.

Tania: Yeah. So, that might then allow you to ask the right questions, to seek the right information, and so on. And some of the time it might then allow you to draw the right inference. So, in the course of explaining to somebody, you realize that you thought you knew how you get from step two to three, but you get to that point in your explanation, you're like, "Actually, I don't really know how you get to step three from step two," but maybe having drawn attention to that, now you can actually figure it out on your own.

Julia: Right.

Tania: So, some times you can identify the gaps and some times you can fill the gaps. That seems to be one important thing you're doing.

Another thing that you're doing is you seem to be integrating whatever you're trying to explain with the prior beliefs that you already had. You're trying to make sense of it in light of what you already know. So, rather than just have the thing that you're learning be this isolated, new thing, it's now more usefully integrated with your prior beliefs.

Another part which seems to be important, and this is the part that my own research has focused on the most, is that I think one of the key things that happens when you try to explain is that you try to relate the thing that you're trying to explain to some broader pattern or regularity.

Part of what gives us the sense that you understand it, now that it's been explained, is that you say, "Oh, okay. Now I see how this fits into this more general explanatory pattern, this broader generalization. It's an instance of something that makes sense to me." And I think part of the process of doing that makes you move away from the particular idiosyncratic things about this example and to focus on the things that are actually generalizable, useful, more general purpose pieces of information. So, by virtue of doing that, you're now going to be in a better position to, for example, generalize to a new case.

So, this is one of the things that you see in the educational literature. You have people explain to themselves, for example, how to solve a particular word problem -- and then you have a test later where they have to apply the same mathematical principle, but it looks really different. So, maybe the first case involved how quickly you could pick berries of different types from bushes, and the second case involves building things in a factory. So, superficially, they're very different, but maybe the actual formula you need to apply to solve the problems is very similar.

And what you see is the people who explained are more likely to be able to generalize from these cases to the new cases. And I think part of that is because what you do when you explain is you don't just focus on the fact that it's about berries and bushes and so on. You relate it to some more general principle, that makes the principle a little bit more explicit, a little bit more accessible.

Julia: And that's not something that you would do just naturally, intuitively, unless you were prompted to explain what's happening?

Tania: What we know from the experimental literature is that it makes a difference when you do prompt people to explain. There's something that they're not all doing spontaneously. Now, that doesn't mean that there aren't other routes to getting there, but it does suggest that it's not something people always do naturally.

I can give you another example, which comes from these other studies with children. A lot of people are familiar with stories that involve a moral of the story. So, you read a story to a child, like Aesop's fables, and superficially it's all about, to give you one example from the literature, there's this cartoon about Clifford the Big Red Dog and there's a three legged dog. The three legged dog wants to play with him. And initially they exclude the three legged dog, and they come to realize by the end of the story that it's fine to play with the three legged dog, it's just another dog like they are.

So, clearly the moral of the story there is something about social exclusion. But if you ask children what they think the moral is, it's that you should be nice to three legged dogs.

Julia: Aww. That's adorable.

This is reminding me of the question that you posed at the beginning of the episode about why would explanation be necessary or why would it have evolved in the first place, above and beyond just prediction. Let's say there's an animal, and on the savannah, we probably have some instinctive sense of whether the animal's dangerous. And there's this sort of intuitive pattern-recognizing prediction algorithm that's producing this prediction, and maybe it's using things like, "Is the animal large? Does it seem confident or not?" Maybe large and confident animals are more likely to be predators than not.

But it's not clear why we would need to be able to explain what are the factors that are causing that animal to seem dangerous to us. Why couldn't we just have evolved fear of animals that have whatever properties our learning algorithm in our brain has determined are associated with danger?

Tania: Yeah, so if you've got the prediction, why do you need the explanation?

Julia: Yeah.

Sorry, the reason that I brought this up now is all these benefits of explaining something to someone else are making me think that maybe there was this adaptation that happened when humans started evolving language, and the social complexity of human tribes started to go up. Where individuals were able to exploit the social environment they were in, either for the tribe's benefit or for their own benefit.

The context for this suggestion is something I'm sure you've heard of -- and I did an episode on, but for those listeners who aren't familiar with it -- the Argumentative Theory of Reason. It's a theory that says our capacity for conscious, deliberate reasoning evolved *not* to help us figure out what's true about the world, but to help us justify our beliefs to our fellow tribes-people. So, basically the point being, reason didn't evolve to help us make decisions, it evolved to help us win arguments.

So, maybe there's something similar that's true of explanation. Where, to take the direct analogy from the Argumentative Theory of Reason, maybe we're fine just using prediction if we were on our own, but in order to convince the rest of our tribe that we should be afraid of this animal, we have to be able to explain why we got the prediction that we did.

I could think of other, less direct analogies as well. Like maybe I end up better at avoiding dangerous animals if I try to explain to my fellow tribes-people why these animals seem dangerous to me. And that's only a benefit to me. Well, I guess it could also benefit my tribes-people, but there would be this separate mechanism as well.

Tania: Yeah, I think there's two parts to your question and I think both are really interesting. So, one is what is explanation at all doing?

Julia: What's it adding above prediction.

Tania: Right. And then, why would it be this explicit, verbal, conscious process?

Julia: Right.

Tania: So, let me say something about the first part, first. One reason why I think explanation might be particularly powerful, even if you ultimately only care about prediction, is because it might be that by trying to explain something, you figure out how it is that you should generalize what you know to new cases.

So, this is much easier to think about in a concrete case. I'll give an example that actually comes from an experiment that we did.

So, the cover story for our participants in this study is that you are the assistant to the director of a museum and you're just collecting all sorts of data about the museum. You just have these giant sheets and sheets of different observations that have been made about people who come to the museum, and what they do and so on. And one of the things that you notice is this really strong correlation between having visited the portrait gallery, for a particular visitor, and having made an optional donation on the way out of the museum. So, there's this correlation there. There's evidence that there's some kind of a relationship here.

Julia: Right.

Tania: Now, from just what I've told you there, if you had to try to predict what might be the case in other contexts that are similar, it doesn't seem like you have a lot to go on. Like, for example, would you predict that people who go to the sculpture garden are also likely to make an optional donation? Would you predict that people who go to the portrait gallery are also likely to give money to somebody asking for money outside the museum?

Julia: It's really hard to-

Tania: Right. You just don't know a lot. Now I'm going to give you an explanation for this relationship. And the explanation is that when you are surrounded by portraits of watchful others, that triggers mechanisms that we have, that have to do with social obligations and our reputation and so on.

And by having activated those kinds of mechanisms, where people are now thinking about themselves as social creatures with social obligations, who are being observed by others and so on, they're more likely to engage in this pro-social behavior of having given money.

Julia: So, knowing that, assuming that's true, then I could say in response to your question about the sculpture garden, I could say, "Well, if the sculpture garden has sculptures of *people*, then yes. I would expect the same effect. But if not, then no."

Tania: Exactly. So, what did the explanation get you there? Well, one of the things that it did by having a sort of mechanism explanation that relates those two pieces of information, is it tells you something about how you should generalize that relationship from that case to other cases.

Julia: It seems like this kind of transfer learning is something that our brains do intuitively, it's like built into our -- I keep wanting to say “machine learning” algorithms, but it's not, it's human learning algorithms!

Tania: Brain learning algorithms.

Julia: Brain learning algorithms, yeah. Like if a baby pushes a block off a table and it falls to the floor, the baby can generalize from that, apparently, to expect that if he pushes a cup off the table, it will also fall to the floor. So, we do have some capacity to do that.

Tania: There's no question that we generalize in all sorts of ways. But the thought is that maybe the process we go through of trying to generate explanations is one of the tools in our generalization toolbox. And maybe it's one that's especially useful in cases where the type of generalization that you need to make relies on causal mechanisms or complicated underlying principles.

Because a lot of the cases where we know that all nonhuman animals can do a fair amount of generalization. They learn that this reddish thing is good to eat; this slightly different reddish thing is also good to eat.

Julia: And they're presumably not using explicit, verbal-

Tania: That's right.

Julia: Well, definitely not verbal.

Tania: That's right. So, the claim is definitely not that explanation is necessary for *all* kinds of generalization, but rather that it just-

Julia: Gives us extra power?

Tania: For humans, it seems to be a particularly powerful way to engage in certain kinds of generalization.

Julia: Got it.

Tania: So, that's the part of your question, but then why would it be this explicit verbal process? And there I do think that what you're pointing to about it being this social communicative activity might be a really important part of that.

Julia: Right. And I would expect that we would see different effects, and we should expect different mechanisms, in cases where the explanation someone's giving is something that they're figuring out on the spot, versus cases where someone already believes something and they're trying to explain to someone else why they believe it. Has anyone looked at that?

Tania: There is a distinction between “why something is the case” versus “why you *believe* something to be the case.” Those are different kinds of explanations.

But I think what you're pointing at is more a case where someone is coming to understand, themselves, on the fly, as they explain -- versus the case where maybe I'm the teacher, I already understand the material well, but I'm just thinking about the best way to present it to you, given certain assumptions about what you are like as a learner. Is that the contrast you're imagining?

Julia: Yeah. I guess I'm imagining that we should expect explanations to be more accurate, or to track more with accurate models of the world, in cases where we are *figuring it out*. Where we're trying to figure out what's happening, to the extent that we can give an explanation.

Whereas in other cases, where I already believe that the animal's dangerous and I'm trying to explain to my fellow tribes-people why that's true, maybe there we shouldn't really expect the explanations to be accurate, so much as just *compelling*. This is basically the Argumentative Theory of Reason.

Those just seem like two different kinds of explanations that we should expect to work differently, because they're serving different purposes, is what I'm saying.

Tania: To the extent that different explanations do have different purposes, I agree that you might expect them to be more or less accurate, or more or less persuasive, and so on. I don't know of any research that gives us clear boundaries for where those should be, that these are the types of explanations that would have these properties and these would be different.

One thing we do think is not too surprising, is that if what you're explaining is wrong, then it's not typically going to be so beneficial.

So, if you're going through a math problem that you've solved yourself and are explaining out loud why that was the solution -- if you got it right, then you might get some extra benefit from having gone through the process of explaining why that's right. That's gonna reinforce the correct principle and so on. But if you got it *wrong*, and you go through the process of explaining how you got there, without recognizing that your answer's wrong, then the explanation might just be reinforcing the mistake that you started out with.

So, there's a way in which explanation might actually be extra beneficial when you're on the right track and you're getting things mostly right, but in cases where you're getting things wrong, it might actually help reinforce or entrench those misconceptions or false beliefs and so on. So, that's one of the cases where I think explanation can actually be dangerous or go wrong, or lead us astray.

Julia: I know you've looked into several ways in which explanation can have a downside. What are some of the others?

Tania: I think one of the ones that I've found most interesting in my research comes from the idea that one of the things that you're doing when you're looking for an explanation is looking for a *good* explanation. You want a satisfying explanation. You don't just want any old explanation, you want a really good one.

And humans are really picky about what it is that makes something a good explanation. We like explanations to be beautiful. We like them to be elegant. And when you try to cache that out, usually that means they have to be simple in some sense. And saying what “simple” means is not at all simple! Typically we want them to be broad, and so on.

So, there's these characteristics that we're looking for in a good or satisfying explanation --

Julia: “Broad” in the sense that this explanation doesn't just explain the case at hand, but it also helps us understand lots of other cases?

Tania: Exactly. For example, if you want to be able to explain why your friend got angry on this particular occasion, ideally you want that explanation to be the same one that you appeal to in similar sorts of situations where-

Julia: Like, people in general get angry when you kick their chair out from under them.

Tania: Exactly. Something that is very idiosyncratic and just applies to this particular case might be a little less satisfying.

So, we're typically looking for simplicity in breadth. So, what I've shown in some of my work is that when there is something simple and broad to be learned or discovered, then engaging in this active process of trying to find an explanation can be beneficial. It might lead you to look beyond the obvious to find this more subtle underlying pattern.

But what about a case where you're trying to look for an explanation and there just is no satisfying explanation? The world just is messy and complicated, and lots of real cases are --

Julia: Where there is some order, but it isn't an easy, simple to explain order?

Tania: I think both of those might be cases where explanation is harmful. So, one could be the case where it's actually genuinely random. And another one is a case where maybe there is some sort of underlying pattern or regularity, but it accounts for 78% of cases and the other cases seem totally unsystematic. So, in those kinds of cases, I don't think explanation's always going to be beneficial, because in some ways what you have to do is step back from the idea that you're going to find a really elegant, beautiful explanation and accept some amount of messiness.

Julia: Right.

Tania: So, our data suggests that in those cases, looking for an explanation can actually some times impair learning.

Julia: Huh. This reminds me a little bit of what I've read about the phenomenon of verbal overshadowing, in which -- I'm gonna try to explain it and you can correct me after I try -- your ability to accurately recall what an experience was like, or what a person looked like, say, is impaired, if after having had that experience or after having seen that person, you were asked to give a verbal description.

Like, the process of generating the verbal description seems like it kind of replaces, or overshadows, some of your intuitive memory. And you're less likely to be able to recognize the person again after having given that verbal description of them. Does that seem like a similar phenomenon?

Tania: There's one part of it that I think is similar. First, let me say the way in which I think it's dissimilar.

Julia: Okay.

Tania: In all of our experiments, we compare participants who are prompted to explain, to participants who do something else which is also verbal. So, they might be thinking aloud. They might be writing down their thoughts. They might be describing. They are also doing something verbal.

And so, at least for our experiments, and for a lot of the other studies that have been done in the literature, we know that the effects that we're attributing to explanation have to do with something about explaining, and not just thinking about language.

Julia: Not just speaking.

Tania: That's right. So, that's, I think, an important dissimilarity from the verbal overshadowing.

Julia: Yeah, that makes sense.

Tania: But I think there is an important similarity there. Where I think one way to understand what's going on in a lot of the verbal overshadowing cases is that you have this very rich perceptual experience of something, like what the person looked like, or what the wine tasted like is another case where you get these effects of verbalization.

So, you have this very rich perceptual experience -- and language just does not supply you with fine-grained enough categories to try to represent it very well. So, when you try to sort of shoehorn it into your language, you're losing a lot of the richness that you had initially.

Julia: And probably adding some bias or error, as well, because you have to pick a word that not only is not *perfectly* capturing it, but it's kind of wrong. It’s just the best you could do.

Tania: That's right. So, it might be both imperfect *and* too coarse-grained.

So, then you see the downstream consequences of having tried to take this very rich aspect of your experience and sort of changed it to a representation that's coarser and perhaps not well-aligned to it.

So, I think what's similar about the explanation case is that there's perhaps the data or the world, and it has certain kinds of characteristics or structure to it. And when you're trying to explain, if what you're doing is looking for a good explanation, a really satisfying explanation, you're sort of trying to shoehorn it into something which has the pattern of a good explanation. A good explanation is simple and broad, and so on.

So, to the extent that there really is something like a simple, broad regularity or pattern there to discover, that might be a pretty effective process. But to the extent it's *not*, you're gonna be slightly distorting the data to try to make it fit better into that pattern.

Julia: Right, but in order for the process of explanation to make us *worse* off, it would have to be the case that we did have some ability to intuitively understand whatever order there was in that messy pattern, and be able to make somewhat accurate predictions.

Tania: That's right. So, to make this concrete, one of the things we do in one of our studies: we have one version of an experiment where people basically have to learn which individuals are the individuals who are very likely to give to charity, versus not very likely to give to charity. And there are in fact just 12 people that they have to learn this about. Six of them tend to give to charity and six of them don't. So, they just have to learn, for those 12, which are the ones that give to charity and which ones don't?

Now, you could just memorize the features of those 12. You could just remember, "Okay, Julia, who has brown hair. She's one of the people who gives to charity. And Bob, who has blond hair, doesn't." You could just memorize these idiosyncratic features. So, you don't really have an explanation for what it is that makes somebody give to charity or not.

Julia: But you can predict.

Tania: For that set of 12. You can classify 12 perfectly.

Now, we've also built in some regularities in the way that we do this. We also give them information about people's ages. We also give them information about whether people have personality characteristics that are more introverted or extroverted.

The way these predict giving to charity differs across people, but I'll just give you one example. So, for one participant, they might get data that's consistent with a claim that the younger people who are more extroverted are the ones who tend to give to charity. So, there's a pattern there that they could extract. Now, if that pattern is perfect, so that you could use that to classify these 12, then the explainers do just as well as people in a control condition.

But now imagine a case where-

Julia: Just as well? They don't do better?

Tania: They do non-significantly better, in that case. We do find numerically they do better, but when you analyze it statistically, it's not significant.

But now suppose you have a case where, of the 12 that you're studying, for 10 of the 12, it's true that the younger people give more to charity than the older people, but there's these two exceptions. It's a case like that, where the people who are prompted to explain as they learn, actually do worse than the people who are not prompted to explain.

So, if you imagine what it's like to be a participant in this task, you see this particular person who you've seen before, but maybe you don't remember now if they were associated with giving to charity or not. You see that this person is young. You think, "Okay. I'm gonna guess they do give to charity." And you get the feedback, "No. This person doesn't give to charity."

If you're in the explain condition, you now have to take a few moments to explain why you think it is that this person doesn't give to charity. And if you're in the control condition, you take a few moments to write out your thoughts as you study that this person doesn't give to charity and so on.

Julia: Got it.

Tania: And what we seem to be finding is that the explainers are just really reluctant to give up on there being some sort of a generalizable pattern that applies to everything. So, they keep on and perseverate in looking for a pattern. They think, "Okay, if it's not about age, maybe it's about where they're from," and they try to find a pattern there. And if it's not about where they're from, maybe it has to do with their college major. They try to find all these patterns on the information they have.

Whereas the people in our other conditions are more willing to at some point either not search in the first place, or abandon the search for some sort of perfect regularity and just settle for, "Okay. This person's one of the ones who gives to charity, this person's one of the ones who doesn't give to charity, and I don't have a broad, underlying principle to explain it, but I can tell you reliably for these 12 people who does what."

Julia: This does surprise me, these results surprise me a little bit. Because I wouldn't expect there to be perfect relationships in the world at all -- so, if having imperfect order in the world causes explanation to fail us, then I wouldn't have expected explanation to be evolutionarily useful at all.

Tania: That's right. And pretty much every regularity in the world involves exceptions, if only because there's noise and so on.

So, we've done a series of follow-up studies where we've tried to figure out to what extent there's really something special about there being a perfect, exception-less pattern, or if it's just that sort of accounting for more cases, accounting for more of the variance is good enough. And so far the data actually suggests that people really like the perfect, exception-less case, which surprises me.

Julia: Maybe it's like a super stimulus. It's like so much better than what we evolved to crave, but ...

Tania: That's right. But for the same reasons that you suggest, I find it surprising. For example, in psychological research, you never account for 100% of the variance. That's preposterous. That would never happen in a study. If you can account for *most* of the cases, that's phenomenal for most scientific research.

So, I think there's a few things that could be going on. To some extent, it's an open empirical question. We don't know all the details yet. But the other thing is that I think we're really, really good at explaining away exceptions, so we can preserve the sense that there really is a perfect regularity despite that.

So, if you're somebody who believes that people who have a particular horoscope have a particular personality characteristic, but then you encounter an exception, what do you do? You could say, "Oh, I guess I was wrong about that association." Or you could come up with some ways in why that person's an exception. And you sort of shield your view about there being this perfect relationship from that potential counter-example. And I think people do that a lot.

Julia: Oh, I totally agree they do that a lot, but that still makes us inaccurate, though, right? So, if explanation is supposed to help us understand and control the world around us better, then if we're making these excuses that aren't actually justified, it doesn't seem like we're better off.

Unless the purpose of explanation is to be compelling to our fellow tribes-people, which the Argumentative Theory of Reason might imply. In which case maybe we are fine as long as we can rationalize why this exception makes sense.

Tania: Yeah, there's another argument one can give -- and I, on alternate days of the week I find it very compelling. Today I'll say it's compelling.

The idea is basically, it has to do with something that comes up in curve fitting, or when you're trying to build a model to fit data, which is that you don't want to over-fit the noise. So, if you have some data points and you're trying to figure out what's the best way to characterize these data points, you want to capture the underlying signal. You don't want to fit the noise. So, you could think about what explanation is doing in cases like this as being a kind of what's called a “regularization” sort of process. It's trying to prevent over-fitting.

Now, part of the reason I only find this compelling on alternate days of the week is because I think, almost certainly, to the extent we do it, it's probably more than we should. I think it interacts badly with other kinds of cognitive biases that do lead to erroneous thinking.

So, for example, I'm guessing at some point in your podcast you've covered things like motivated reasoning, confirmation bias. These are processes that lead us to interpret things in the way that's most favorable to what we want to believe or what we already believe.

So, I can imagine explanation interacting really poorly with those kinds of processes where it might give you extra tools in some way for holding onto beliefs that you want to believe. So, I think it's definitely not always going to be beneficial if it is playing this role of allowing you to sort of ignore some of the noise and favor the broad generalizations.

On the other hand, there might be cases where it's actually useful in preventing you from over-fitting noise and focusing too much-

Julia: On focusing too much on any single data point.

Tania: Yeah. That's right.

Julia: Liking your models too much.

Tania: Yeah. Sort of focusing on what's the single big thing going on here that allows me to predict most things, rather than sort of getting in the weeds of the small variations which maybe don't generalize much from case to case.

Julia: Right. That makes sense.

Well, since we've started talking about potential justifications for prioritizing simple theories over complicated theories, I have a couple ideas that I want to run by you. They're not my ideas, but I'm confused about how convinced I should be by them.

One of them comes from a philosopher at Carnegie Mellon named Kevin Kelly. He's basically been working on trying to see if there's a formal way that we can state Occam's Razor -- which is, you know, "prioritize simple theories" -- such that it is actually provably true.

One way that he's cashed it out is in saying that, "Look, it's not that simple theories or simple models are more likely to be true. It's rather that having a policy of reaching for simple theories -- like the simplest theory you can find that still explains the data -- having that policy over time allows you to reach the true theory more quickly, or more efficiently." Assuming you're, of course, discarding those simple theories as new data invalidates them, and then reaching for the next most simple theory that explains your current data.

So, it's basically, I think he calls it Occam's Efficiency Theorem, which is an interesting take on ... before reading about that, I hadn't considered that there are different ways of striving for truth. One is, "I want to be as accurate as possible *right now*". Another is, "I want to maximize my ability to *converge on accuracy* quickly or efficiently". And I think there are others, as well.

But what do you -- it seems like you are familiar with Kevin Kelly's work. What do you think about it?

Tania: I am. I think it's really cool. I think the general idea, which you just highlighted, which is that maybe we should favor simple explanations not because the world is simple, but because it's a good policy for getting at the world, I think that's a really interesting and powerful idea.

For Kelly to prove that, he has to come up with a very precise way of articulating what simplicity means. And the way he does so, I think, may or may not map on to what we intuitively mean by simplicity when we talk about explanations in sort of everyday cases. I think that's sort of an open question.

Julia: Yeah. I agree. I kind of glossed over that quite a bit. But it still feels like there's a colloquial version of that that's kind of true ... what's that expression? "Strong beliefs held weakly"? Or "Strong view weakly held"? It's a mantra that's like, "You should, instead of having vague, uncertain views about things, you should just have bold views that you know are likely to be wrong, but at least if you state them clearly enough that the world can disprove them for you clearly, you're going to become more accurate over time."

Tania: Actually, the philosopher Popper said something along those lines. He argued for simplicity, again, not on the grounds that the world is likely to be simple, but that a simple theory is easier to falsify. So, what you want to be doing to make scientific progress is articulating theories where it's clear how you would go about testing them, which for him meant trying to falsify them. Then it's going to be good to be formulating those kinds of theories.

So, in general, I really like this sort of approach. I've actually argued for a position with a philosophy co-author that we call "explaining for the best inference", and that's supposed to contrast with what's called, "inference to the best explanation". So, the idea behind "inference to the best explanation", it's a formal term that was introduced in the philosophy literature, but I think it's made its way outside of philosophy and I think it's probably reasonably well known, but it's basically the idea that if there's one hypothesis that explains the data better than any other hypothesis, then you should infer that that's the hypothesis that's true. So, it's a pretty intuitive sort of idea.

Julia: Hard to argue with.

Tania: Yeah. It's a pretty intuitive idea. Although a lot depends on how you flesh out what it means to be the best explanation. Exactly.

Julia: Is it the model that makes that data the least surprising, like that's what a Bayesian might do?

Tania: Right, so the attempts that there are to give a formal articulation of what people do when they evaluate explanations suggest that that's not quite what they're doing. They're doing something more like figuring out how much evidence the observations provide for some hypothesis. Rather than figuring out the posterior probability of that given hypothesis.

Julia: Right. So, they're sort of ignoring the base rate?

Tania: I'm basing this on about three studies that have been done, so this is not a huge literature. But those three studies suggest that when people are making explanation judgments, they are less sensitive to the base rates or priors than if they were calculating [posterior] probabilities.

But it seems like a reasonably intuitive idea. And the shift that I've argued for, this is with co-author Daniel Wilkenfeld, is from "inference to the best explanation" to "explaining for the best inference." Where the idea, like the Kelly example, is that maybe the way to think about it is not that the reason we favor particular explanations is because those explanations are the most likely to be true, but rather the reason we have certain explanatory practices is because those explanatory practices are, downstream, most likely to lead us to true or useful inferences.

Julia: Right. Nice.

Tania: So, what that allows in are cases like that, where maybe if you imagine sort of a long process of learning or scientific inquiry, maybe the process of engaging in explanation doesn't get you to the true thing immediately, but maybe it gets you to formulate the kinds of hypotheses and do the kinds of experiments and collect the kinds of data that downstream are gonna have beneficial consequences.

Julia: Right. Maybe now's a good point at which to talk about “explanatory vices.”

So, we've been talking about -- we didn't use this phrase, but we've been talking about explanatory *virtues*, features of explanations that make them more desirable. Explanations that we want to reach for more, explanations that are more likely to lead us to the truth, etc. But you've also written about explanatory *vices*, which are features of explanations that make them more desirable but aren't actually more useful or true.

So, one example of an explanatory vice that I like is: random, scientific, irrelevant gibberish makes people apparently more likely to consider something a good explanation. What's happening there?

Tania: There's a few explanations for what could be going on in a case like that. The uncharitable one is to say that when people are given scientific gibberish, their ability to engage in good, critical reasoning just kind of goes out the window. So, they're doing something pretty dumb in those cases.

Julia: They're like, "Well, it seems science-y, and science-y things are more accurate than unscience-y things"?

Tania: See, I think even the way you just described it there is not the most uncharitable --

Julia: Did I just inadvertently steel man it?

Tania: You made it slightly more charitable, I think, actually, because that's not a crazy inference, right?

Julia: That's true, I guess. Science-y things are more likely to be accurate than unscience-y things.

Tania: On average.

Julia: I retroactively endorse that.

Tania: So, I think the least charitable is something like, "In the face of science, your ability to engage in real thinking shuts down".

Julia: I see. Okay.

Tania: And then there's the increasingly more charitable ones. I'll give you the other extreme, the more charitable interpretation, which is: You just gave me an explanation. I'm really good at evaluating explanations. I realize that there's a gap in the explanation you gave me. You didn't really give me a good causal mechanism, or you didn't really give me a good generalization as part of my explanation. But you also said the scientific jargon-y thing. And I didn't totally understand that, but I'm going to assume that because you're an authority, an expert on this topic, that thing I didn't entirely understand, that's what fills the gap and makes it a good explanation.

Julia: Right.

Tania: So, I'm gonna basically defer to the authority that I think you have, and assume that, "When the doctor said that, the doctor's an expert on this stuff. I didn't totally understand it, but presumably when the doctor said, 'Blah blah amygdala, blah blah,' that actually was part of the explanation."

So, in that case, I think what people are doing is basically applying a heuristic that's reasonable in lots of cases, but that isn't perfect, and it's that if I have reason to think that you're an authority in this domain, that you know what you're talking about, then even though I didn't understand it perfectly, yeah, I'm gonna count that as good enough.

Julia: I guess it seems like a little bit of question substitution. In that people were asked, "Is this a good explanation?" And they're answering instead, "Is this probably correct?"

Tania: Yeah. I think that could be true. That's right. We know that in lots of cases, people are doing something more sophisticated than just, "Is this true?" for an explanation. So, for example, if I ask you, "Why does inflation occur under these circumstances?" And you say, "Because there are trees." Yes, there are trees. I know that's true, but people are not going to accept anything true as an explanation. So, they're doing something a few notches more sophisticated.

But I think the idea is that you've asked them, "How good is this explanation?" And what they're answering is, "How good do I think it would be *to somebody who could understand it*?"

Julia: Right.

Tania: They are answering a different question, but it's probably a reasonable one. It's a reasonable substitution in lots of everyday cases. So, I think I can understand why people would function well enough a lot of the time doing that, and I think that is the more charitable way to understand that, this phenomenon.

It also is something that can go wrong in lots of real world cases. For example, people have looked at the effects of neuro-scientific jargon in the case of an expert testimony in a legal case. And there it does suggest that people are often very swayed by neuro-scientific-y sounding explanations for people's behavior, in a way that -- we could argue about whether that's good or bad for the legal system, but it has an influence there.

Julia: Certainly seems like it *can* be bad, I don't know if it's *on net* bad.

Tania: That's right. So, that's a case where it seems like it might have important real world consequences whether people are able to engage in good, critical evaluation of explanations when they involve neuro-scientific jargon.

Julia: Right. It also seems like there's an aspect of science-y explanations that we find persuasive above and beyond the authoritativeness of it. So, I'm thinking of all these pop science articles that say like, "We have the scientific explanation for love. It's that this region of the brain activates," or something. And that, I guess, kind of feels like you've *explained* love. Like you have this physical mechanism that causes the subjective experience of love.

But it doesn't really explain anything. In that we don't really know what causes that region of the brain to activate. We don't know, sort of zooming out, we don't know why love would have evolved. Maybe we have other explanations for it, but I'm just saying knowing which part of the brain is active when love is being experienced in a human brain, in a human mind, is not really that much of an explanation. But I think a lot of people feel like because there's a mechanism, that counts as a good explanation. And that doesn't just seem like appealing to authority, that seems like a deeper, more philosophical confusion.

Tania: Yeah, I think you're right. I think there's a lot going on in cases like that and we've tried to actually do a few different studies that relate to that.

Julia: Oh, cool.

Tania: I'll tell you about one of them. So, what motivated one of the studies was the sense, not so much that when you get the brain region, that that seems like an explanation, but that when you give a category or a name for a phenomenon, that seems to provide an explanation.

Julia: That's right. Feynman wrote about this.

Tania: Did he?

Julia: Yeah, I think so, that people will often feel like ... you give kids the name of a bird and now the kid understands all there is to know about the bird. Like, "Oh, that's a robin," or something, "Now I get it." And he instead wanted to focus on -- I hope this is correct and I'm not mixing him up with someone else, but -- he was instead arguing for explaining about the bird's behavior, or how it differs from other birds, instead of just giving it names.

Tania: Okay. Interesting. So, maybe we should have called our phenomenon The Feynman Effect. We called it the dormitive virtue effect.

Julia: Also catchy.

Tania: Because of this famous passage from Moliere, where the question is why some particular drug makes you sleepy? Because it has a dormitive virtue. What's a dormitive virtue? It makes you sleepy.

Julia: Oh, right!

Tania: Exactly. So, what all of these explanations that we're talking about I think have in common, is the idea that you're pointed to something -- whether it's a brain region, a category, a label -- and it seems like that's doing some explanatory work for people that maybe it shouldn't.

So, what we did in our study is we gave people little vignettes where we described a person's behavior and then we explained it in a way that either did or did not include a category. For example, this person engages in theft, they steal objects from a store that's not particularly valuable. Why did they do this? Because they have a tendency to steal objects from a store -- or because they have "dypathopy, a tendency to steal objects from a store".

Julia: Oh, I bet that second one was even more ...

Tania: The second one was much more satisfying. We just added this made-up name for a condition, and that made it much more satisfying.

But then the interesting question is, why? What is it that's going on when you give people that label, that makes them think it's more satisfying? And what our subsequent study suggested is that when you give a label, people seem to assume that there is some underlying cause. Something about the person that causes them to have this behavior.

So, in some ways, what the label is doing is serving as a placeholder for a causal mechanism, in the same way that you might think labeling a brain region does. And in lots of cases, it's reasonable that if you give a causal mechanism, that does explain why something was the case. So, I think what's going on here is people are basically accepting an indication that there is a causal mechanism for the explanation itself.

Julia: But even that is sort of a leap or an assumption on their part. That it's not just ... like, if you have the flu, there is a virus in your body that's causing your symptoms that we call the flu. But if you have Narcissistic Personality Disorder, that's just a name that we've given to the set of symptoms you're displaying. We don't necessarily know that there's some underlying thing in your body or your brain that is causing those symptoms to exist.

Tania: That's right. People tend to think --

Julia: People tend to think it's like the flu, yeah.

Tania: And to the extent they find it explanatory, it seems to be because they assume it is like that.

Julia: Interesting.

Tania: I should say, just to be clear, it is not in fact like that. And most clinicians do not think it is in fact like that.

Julia: No, yeah, I get that.

Tania: But our studies with lay people who have no special expertise related to mental illness do suggest -- you do get some variation in the extent to which people think of it more like the flu, where there's a single underlying cause, or maybe a set of underlying causes. And to the extent you do think about it like that, you're more likely to find an explanation about "Why does that person experience persistent sadness? Because they have depression," they're more likely to find an explanation like that satisfying, to the extent that they are actually thinking about depression as something that involves some underlying cause that is responsible for the persisting sadness.

Julia: Interesting.

Tania: As opposed to thinking about it as just, "That's what depression means, it means to have these symptoms".

Julia: You're so good at steel-manning these apparently irrational tendencies, I'm gonna throw one at you that seems especially hard to steel-man.

What about the finding that people, when you ask someone if you can cut in line in front of them at the copy machine, and your explanation is, "Can I cut in line in front of you, *because I need to make copies*?" They're much more likely to let you cut, than if you just say, "Can I cut in line in front of you?"

And you have literally given them no additional information. Of course you need to make copies, that's why you're in line at the copy machine. And yet, somehow, that's compelling. What's going on there?

Tania: I think I'm probably not going to be able to give you a rational explanation in that case. But I do think… here's some attempts.

So, one thing you're doing is giving people something *in the form of* an explanation. It still has some of the same structural properties as a good explanation. And it could be just for dumb associative reasons, we get some amount of satisfaction from that.

The other thing which I think could be going on in that case is particular, is you've at least acknowledged the need for an explanation. Like, if I just cut in line without offering you something which even looks explanation-like, I've violated a sort of social norm about the conditions under which it'd be acceptable to do that. But if I at least say -- it's more like saying, "Can I cut in line? I have a good reason." I'm not telling you what the reason is, but I'm at least acknowledging that I need some reason.

Julia: It's acknowledging that you *deserve* a reason.

Tania: Yeah.

Julia: Which might even be less about the information you're conveying, more about playing the social game right.

Tania: Right. So, that could be going on in that case. But I should say there's other cases of similarly, or almost equally, minimal explanations that also seem to make people more satisfied. And I'm not sure that explanation works for all of them.

Julia: This has just become a joke line among me and my friends. Every time we ask for a favor, we'll just add, like, "…because I want to make copies".

I guess the last explanatory advice that I want to ask you about before we have to close is these teleological explanations, which you've written about, that I think are so fascinating. Where you'll ask a child, "Why does the sun provide light?" And they might say something like, "So that plants can grow".

It's this weird, kind of confusing -- it's almost like confusing the effect with the cause. The sun has this beneficial cause, therefore that must be the reason. Sorry, the sun has this beneficial *effect*, that must be the cause of the phenomenon of the sun. What is going on there?

Tania: There's a few different explanations. I think the first thing to say is that there are some domains where that's actually a completely reasonable way to explain things. So, if you were trying to explain why a pen has a little clip and you say, "It's so that you can put it in your pocket." In that case-

Julia: Right. That's someone designed it that way.

Tania: Exactly. So, in that case, it's totally fine. So, we say, "so that you can put it in your pocket", and that's kind of a shorthand that we all understand for something like, "When it was designed, that part was put there with the intention that you'd be able to put it in your pocket." So, that case seems like straightforward or intentional behavior.

So, "Why did you go to the cupboard?" "So that I could get the chocolate." I'm giving the effect of my action, but it makes perfect sense that what's going on is I had the intention to get the chocolate, and the intention caused me to act a particular way.

So, there's lots of cases where it looks like what we're doing is explaining something by appeal to its effect, but in fact you can give a totally reasonable-

Julia: Someone anticipated those effects, and that's the cause.

Tania: Exactly. The cases that get trickier are… evolutionary cases get tricky for lots of reasons, which maybe we don't want to get into. But let's think about cases like a child saying, "Why are there mountains? There's mountains for climbing." Okay, how do you make sense of that case?

Julia: That's so cute.

Tania: So, one possibility is that they have basically taken this mode of reasoning that makes a whole lot of sense when you're reasoning about human artifacts like pens, or human behavior that's intentional, and they just overgeneralize it from that case. Maybe just because it's familiar, so they do it that way.

The strongest version of that says they've basically assumed the world is designed. So, the strongest version of that says, "Basically, people are creationists deep down, and they're operating as if things are designed, and the reason they find this mode of explanation compelling is because it makes sense if you really assume a designer".

Julia: I mean, if this phenomenon is only occurring in children, or is strongest in children, maybe children just don't know that mountains weren't designed. But if adults are reaching for teleological explanations, despite presumably knowing that the mountains weren't designed -- I guess if you were doing the experiment on non-creationist adults -- then it's confusing.

Tania: So, I'll tell you what the literature suggests there. The literature suggests that if you take non-creationist adults, most of them will not accept that sort of explanation. But if you then make them respond really quickly, the kinds of errors that they make are the errors of saying that that type of explanation is right, rather than the error of saying other types of explanations are wrong.

Julia: That's so interesting.

Tania: So, the person who's done most of this research is someone named Deborah Kelemen. She and her colleagues have argued that maybe this kind of explanation is sort of our cognitive default. It's like our primitive, preferred form of explanation. And that's why you see it in children, that's why you see it in adults under speeded conditions.

She and I have a paper from many years ago when I was a graduate student where we looked patients with Alzheimer's disease [who] show the same kind of pattern. So, it seems like one way to make sense of that is that that's sort of our default, and then if you have appropriate education, you have appropriate beliefs about what actually led to mountains and so on-

Julia: Or just the cognitive resources to perform an override.

Tania: That's right. Exactly. Then you can override it. But in the absence of that knowledge, and that ability to do the override, you're gonna see it emerge. So, that's one view. And Deborah Kelemen's been the most strong advocate of that type of view.

There's another view which I think I'm actually, on most days, more sympathetic to -- and you will accuse me of again having a cheery, positive view of people as not being so irrational…

Julia: No, I think that's a great practice! I endorse the tendency to try to steel-man apparent irrationality.

Tania: That's right. I am a glass half full sort of person about human rationality, I don't disagree about the half empty, but I tend to focus on half full. So, here's another way to make sense of it. If you look around the natural world, and the man-made world, human made world, there are some things that have a very, very good fit between some structure and some function.

For example, my glasses are very well suited to fit the top of my nose. There's a very good fit between the shape of my glasses and the shape of my nose, and ears, and so on. Now, that's probably not coincidental. So, how do we make sense of the fact that there's this very good fit between this structure and this function? Well, maybe let's assume that there was some kind of a process of design or co-evolution or something like that, such that you end up with this good fit. So, a good fit between structure and function is gonna be a really good cue that something is the sort of thing that supports one of these two logical explanations. And in fact, with the case of my glasses, it's true that I can say, "Why are my glasses shaped like this? So that they fit on my nose in this particular way."

But that cue is not perfect. You can get it wrong. You could get it wrong if you said, "Why do we have noses? So that we can hold up our glasses," right?

Julia: Right.

Tania: Same cue about the fit between the noses and the glasses, but there you got it wrong. So, the thought is there's this really good cue that a teleological explanation is warranted, so you can sort of defeasibly make an inference that that kind of explanation is good.

Julia: Defeasibly? Defensibly?

Tania: Defeasibly meaning, you can make an inference, but that it can be defeated by additional information.

Julia: Oh, I see. Okay.

Tania: So, it's sort of like a cue that this is the case, but it's not a perfect cue. There can be exceptions. So, when I say a good fit between the structure and the function, like my nose and my glasses, I think, "Okay, a teleological explanation is good here, but I know extra stuff, and in this case I know extra stuff that tells me it'd be a mistake to explain that I have a nose so that it fits glasses."

So, if that's true, then what's going on is that what you see in kids, and what you see in adults under speed conditions, is that they don't have the time, or the resources, or the cognitive abilities to factor in other information besides the single strong cue about the structure/function fit. Part of the reason, I think, that that might be compelling is because even under speeded conditions, adults do not accept terrible teleological explanations. I can't think of a good example-

Julia: Terrible, meaning cases where there isn't actually a good fit?

Tania: Cases, yeah, where there's not a good fit, or where something-

Julia: Like, "Why are there mountains? Because the sun is yellow," or something.

Tania: Yes. That's right. They will not do it under those conditions. Or, "Why are there elephants? For holding down papers in the wind."

Julia: Got it, yeah -- that's actually teleological, mine wasn't.

Tania: Elephants *could* actually hold down papers in the wind. They're big and they're heavy. But that's not a good explanation for why are there elephants.

Julia: Not the typical use of an elephant.

Tania: That's right. So, I think the more charitable way to make sense of people's behavior in this case is that if you don't have the relevant prior beliefs, if you don't have the cognitive resources to take other sources of information into account, then you're just gonna go with structure/function fit and you're gonna infer a teleological explanation as warranted.

Julia: And that that may in fact be a rational, bounded heuristic -- that given limited time and resources, this is the best explanation we can get.

Tania: Yeah, I don't know that I want to go all the way to say that it's optimal. But I guess I'd be willing to say it's a cue that is often right in lots of cases. Optimal is going a little bit stronger, by saying, "and you could do better with other cues". I'm not sure I want to go there.

Julia: I guess I was just trying to define bounded rationality there, but I wasn't clear.

Tania: Fair enough.

Julia: Excellent. Well, this is probably a good place to stop, if we have to stop. Before we close, I want to invite you to give the Rationally Speaking pick of the episode, which is some book or article or blog that has influenced your thinking in some way, like maybe it changed your mind about something or got you interested in the field you're currently in. So, what would your pick be?

Tania: Can I cheat and give you two?

Julia: You can. People have cheated in the past and given me two, or even three, yes.

Tania: All right. Well, in terms of what got me into the field, I had a sort of circuitous path into cognitive science. And part of what started it was that I read a book by Steven Pinker called *The Language Instinct*, which I actually read as a high school student, and I immediately found it incredibly fascinating, because it approached the topic of language, which I was interested in to the extent that a high schooler who likes English and language can be interested in this.

But it did so really formally and rigorously. It sort of seemed like it gave me a glimpse of what it might look like to do something like a rigorous science of the mind, and I thought that was really fascinating. In reading this book when I was, I think I must have been 16 or 17, he kept mentioning this Noam Chomsky guy, so I thought, "This sounds like someone kind of important that I should have heard of."

So, I went to the bookstore and I bought a book by Noam Chomsky -- and my criteria for choosing was that it was the shortest book. For those who don't know Noam Chomsky, he's done very influential work in linguistics, but also a lot of work in politics and political theory and so on. So, this could have gone very wrong.

Julia: Yeah, I was wondering how this was gonna turn out.

Tania: But I happened to pick a book-

Julia: There's a fork in your life path right there, with the Noam Chomsky book you pick.

Tania: That's right. I went with the language side, so that was good for me. And so I read this book by Chomsky where he kept talking about the cognitive revolution. That sounded kind of interesting and important, so that just sort of led me on this path of reading things related to cognitive science. So, for me, that was an influential book in sort of giving me a glimpse early on of what an interdisciplinary science of the mind might look like, why it might be interesting, and so on. So, I would definitely recommend that today still, although in a lot of ways, the field of linguistics and the study of language has changed since that book came out.

The other thing that I really would recommend to listeners of this podcast is a resource that I find extremely useful, which is the Stanford Encyclopedia of Philosophy. And part of the reason I really think this is phenomenal is because it is a publicly available resource, so anyone can go now online and find the Stanford Encyclopedia of Philosophy, but what it provides are very clear, peer-reviewed summaries of topics in philosophy.

So, you sort of get some of the best of Wikipedia, but some of the best of the peer-reviewed process in academia. Because these are publicly accessible, but for the most part, written by leading scholars in the field. They have entries on topics like, I recently read the updated entry on what makes something pseudoscience. So, talking about the philosophy literature on how you demarcate science from pseudoscience. But you'll also find lots of things about historical topics and philosophy, contemporary topics and philosophy of mind and philosophy of science, and so on.

When I need to point students to sort of the first thing to read, in getting oriented to some topic that's been addressed in philosophy, that's usually that place that I point them to.

Julia: Nice. Is it also open, like a crowd-sourced project the way Wikipedia is?

Tania: No. In that way, it is not like Wikipedia.

Julia: Maybe that's why it has such high, rigorous quality.

Tania: That's right. It goes through a peer review process, which is similar to the peer review process that you would see for standard journal publication in academia, but it is publicly available.

Julia: Excellent. Cool. Well, we'll link to *The Language Instinct* and should we link to the Chomsky book as well, or do you only want to recommend The Language Instinct?

Tania: You know, I'd have to figure out which one, I have it somewhere on one of these shelves, but I'd have to figure out which Chomsky book it was.

Julia: Okay. Well, we can always add it later.

Tania: All right. Sounds good.

Julia: And we'll link to the Stanford Encyclopedia of Philosophy as well as to your website and to your blog. Actually, I didn't mention in my introduction, but Tania also blogs for NPR and Psychology Today, right?

Tania: That's right. The NPR blog is called *13.7: Cosmos and Culture* and it's a blog about the intersection of science and culture, broadly construed. So, I and four other academics blog most days of the week. And I have blogged for Psychology Today as well.

Julia: Okay. Excellent. We'll link to all of that. Tania, it's been such a pleasure having you on the show. Thank you so much for joining us.

Tania: Thank you for inviting me.

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