Rationally Speaking #154: Tom Griffiths on, “Why the brain might be rational after all”

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

Tom is a professor in the University of California, Berkeley Psychology and Cognitive Science Department, where he runs the Computational Cognitive Science Lab. Tom's research focuses in large part on understanding how the human brain forms judgments and makes decisions. And to what extent we can model those decision-making processes as following ideal statistical algorithms, the sort that we might program into, say, an artificial intelligence.

That's going to be the topic of today's episode. Tom, welcome to the show.

Tom: Thank you.

Julia: To start off, maybe you could give us a sense of why this is even a plausible hypothesis. Why would we expect a priori that the human brain might be using statistical algorithms?

Tom: I think that's a good question. To me, a lot of it comes down to the fact that, despite what we might read about our human failings, human beings are still the best example that we have of a system that can solve a variety of different kinds of problems, the problems that are the kinds of things that are at the cutting edge of AI and machine learning research. Things like learning languages, learning causal relationships, learning from small amounts of data, being able to do things like science, which requires us to engage with the world and then try and figure out the structure of that environment that we're in.

I think for me the key question is how it is that we're capable of making those kinds of inferences. And then using rational models as a starting points gives us a set of tools for engaging with those questions.

Julia: Something I realized that I'm confused about as I was preparing for this conversation is: I'm confused about whether the hypothesis here that you're investigating, or the claim that you're making, is that the brain actually uses statistical algorithms, for example Bayesian inference, which we'll get to later in the show… or whether the claim is that the brain can be described *as if it is using* Bayesian inference.

Just to clarify in case the distinction isn't clear to listeners: you might say that, for example, when people decide what music to listen to or what clothes to wear, that those decisions can be modeled as signaling decisions, in which people are trying to align themselves with one social class through their consumption choices and away from another class, but just because the decisions can be modeled that way doesn't mean that that's what's happening inside their mind at the time that they buy the Kanye album. They may just be thinking, "Oh, I'm going to enjoy this music." Still, the forces that conspired to shape what music they expect to enjoy or not can be influenced by signaling forces, essentially.

So with the question of whether the brain is using statistical algorithms or just making choices that closely match what statistical algorithms would produce, that seems like an open question to me.

Tom: Yeah, that's right. I think that's an important distinction to make when we're thinking about modeling cognition. In cognitive science, we talk about different levels of analysis, which is basically different kinds of questions that we can be asking and different kinds of answers we can be getting about how the mind works. One of the classic versions of this division of these questions into different levels of analysis is due to David Marr, who is a neuroscientist at MIT. He basically argued that you could distinguish three different levels at which we can ask questions about information processing systems in general, and minds in particular.

One is what he called the computational level, which is really about the kind of abstract problem that's being solved, specifying what it is that's the problem that the information processing system has to solve, and then what the ideal solution to that problem looks like, or, as he put it, the logic by which it's carried out.

The next level is what he called the algorithmic level, and that's really about what the algorithms are that implement or approximate those ideal solutions. In the context of cognitive science, that's a question which is about actually what's going on inside our heads. What are the cognitive processes and strategies that we're engaging in?

Then the third level is what he called the implementation level, and that's about brains and neurons, but for people it's really about how it is that those algorithms are executed in hardware.

Most of the work that I've done historically has been at that computational level, where we're saying, "Let's look at the problems that people face," including those problems that I talked about. How do you learn a part of a language? Or, how do you figure out causal relationships? How do you learn from small amounts of data? Those are all problems that we can formulate in certain mathematical terms and we can say, "Okay, here's the problem that we want to talk about." Then we can say, "Well, what are the solutions to those problems?" Those solutions end up being things like Bayesian inference, as you mentioned.

What that does is it says, "If this is the problem that's being solved, then here's what the solution to that problem looks like." That equips us to basically go out with a certain set of explanatory tools where we can be asking, "Okay, are there problems that give us solutions that end up looking like aspects of human behavior?" If there are, then we get some sort of hypotheses out of that about *why* it is people might be behaving in the ways that they are, but we don't have answers to the question of *how* it is they're behaving in the way they are, which I think gets at the second part of your distinction about what it is that's actually going on inside our heads.

The next question after that is to say, yeah, "What are the actual algorithms inside our heads?" We get a little bit of constraint on thinking about those algorithms from having perhaps identified some good hypotheses at that abstract level about what

the structure of the problems is.

Julia: Let's delve a little deeper... Maybe let's do a simple example, or-- I don't know exactly what counts as simple, but -- an example that seems to me to be likelier to be simple, like visual perception. How does the human brain recognize an object as being a cat instead of a table? That sort of thing. Or maybe language is simpler, I don't know. If you could just walk us through what would a theoretical ideal solution be to that kind of problem so that we can see ...

Tom: Those are both pretty hard problems.

Julia: All right! I realized that my intuitions might be off, there, as soon as I started that sentence. Why don't you pick a problem, like a typical problem the human brain might face, and you could describe what a theoretical solution might look like?

Tom: I'll just say, those problems, part of the reason why they're interesting and part of the reason why we study aspects of those problems is that they're things for which in many cases we don't actually know what the right rational solutions are. The reason why they're at the cutting edge of AI and machine learning is that those are things where we haven't necessarily got exactly the right formal tools for engaging with those problems.

We can characterize the structure of those problems in very abstract terms, where we say, "Well, what you get is some data," which would be an image or some utterances in a language. And then what you're entertaining are hypotheses about the contents of that image, basically trying to go from, say, its two-dimensional structure to a three-dimensional structure, and understanding what the objects are that appear there. Or going from those utterances to a grammar, or another kind of understanding of the nature of the language. Doing that is something which is, at the moment, beyond the capacities of the kinds of algorithms that we have available to us.

I'm going to focus on a simpler case, or at least a case where we can boil it down to a simple possibility, which is, say, learning a causal relationship. In the context of learning a causal relationship, the very simplest thing that you could think about is something like trying to infer a single causal relationship, so whether, say, a medicine causes people to recover from a disease. In the context of studying that, what we'd say is, "Let's think about what the data are that people might have to work with."

You could imagine that you were a doctor and you treated some people with this medicine and there are some other people that hadn't been treated with the medicine. What you get out of that is what a statistician would call contingency data, the frequency with which people recover whether or not they're treated with the medicine, as well as the frequencies with which they don't recover. Then what you have to do is evaluate whether there's an underlying causal relationship. If you're somebody living in the 21st century with the benefit of over 100 years of statistics investigating this kind of question, you can pretty quickly say, "Okay, well, I'll go off and do my chi-square test," or whatever it is that's your favorite statistical test.

What's interesting about this problem is that for thousands and thousands of years, human beings have made inferences about causal relationships without having access to statistics. In fact, the history of scientific medicine is one where even as recently as the 19th century people would have to figure out whether medicines worked or not relying only on their intuitions about causal relationships. If you were a doctor living in the 19th century and you wanted to convince somebody else that they should use this treatment that you'd come up with, what you basically end up telling them is the data, and then they have to make their own judgment as to the causal efficacy of that treatment just from those raw data. Then those causal intuitions are what determines whether that becomes standard medical practice.

The way that we approached that problem then at this abstract level, thinking about the computational problem, is to say, "What we're trying to do is induce whether this relationship exists." Then we can go to statistics or we can go to AI where people have started to think about causality. In practice, this is joint work with Josh Tenenbaum. What we did was go and look at AI where people were starting to use what are called “graphical models” as the basis for making causal inferences. Basically, what you can do is you can say, "Here are two different models that could describe the world, one in which a causal relationship exists, one in which one doesn't exist." Each of those models implies probability distribution over the data that we expect to observe, and then we can use Bayes' rule to evaluate those different hypotheses with respect to those data.

That part of the problem was something where there were existing tools in AI for formulating it. What we get from studying human cognition is actually insight into the implicit assumptions that guide people's inferences about causal relationships. In work with Josh and then in subsequent work with one of my former students, Saiwing Yeung, we've done a series of experiments where we've established that what people expect about causal relationships is that if you say, "A causes B," that means that you assume that A occurring is going to increase the probability of B occurring. That might seem like a trivial observation, but it's actually something which is not assumed in almost all statistical notions of what you do when you look for a causal relationship and most of the notions that are used in AI.

Another important thing is that if you say, "A causes B," people assume that A causes B most of the time, that if A occurs it's very likely that B will occur, so that it's a near deterministic relationship. It turns out that to model human judgments accurately, you need to have those two assumptions, that causes increase the probability of their effects, and that they are near deterministic.

When you do that, you get a very nice model that does a great job of predicting people's judgments of whether causal relationships exist or not, and something which you can then translate over to AI machine learning, where if you want to build an AI system which induces causal relationships in a way that's going to make sense to people, you probably want to have it make similar kinds of assumptions, because the things that it finds otherwise aren't going to be the kinds of things that people think about as causal relationships.

Julia: Is the takeaway there that human cognition, the inferences that our brains automatically make about causal relationships, that those do basically follow the

process of Bayesian inference, I guess? We're comparing the distribution of data we would expect given a causal relationship, to the actual distribution of data that we've seen, that our brains are following that process -- but subject to certain assumptions about the world that an AI might not necessarily have?

Tom: Yes. There are two paths to that. One is we get a pretty good model of people's inferences by assuming that they're doing something like Bayesian inference over these causal graphical models, but with some important deviations which I'll mention in a moment.

Then the other is, yeah, that in order to make that work, we have to make these additional assumptions which aren't present and weren't present in the existing mathematical methods that had been used for answering these question in AI. We get value out of that by considering the human case.

For something like causality, causality is a weird and interesting thing, because nobody has actually ever observed a causal relationship. Causal relationships are things that, they don't exist in the world. As Hume pointed out, you've never really got good evidence for a causal relationship being something that actually exists. It's more an expectation which we're imposing on the world around us. It's interesting to ask the question of whether there's a notion of causality independent of the human construal of causality. I think a lot of statistical arguments that you can have about, okay, what's a good way of characterizing a causal relationship or not, really end up being psychological arguments about the intuitions that we have about the nature of causality.

If you come back to those important deviations, one thing that we observe is that people are not as sensitive as they should be to sample size. This is consistent with a lot of other psychological research. The place where our Bayesian models definitely deviate from people's judgments is that, as you increase the sample size, the Bayesian model's going to say, "Okay, now you're increasing the amount of evidence," and it's going to say so at a much faster rate than people do. People act like the evidence is increasing at a much lower rate than it should, as you increase sample size.

There are ways of thinking about that. One is that when we make these models we make very strong independence assumptions, that observations are independent of one another. That means that you're going to get strongly accumulating evidence, whereas if you think that you're not getting data that's as good as that and some of those observations are dependent in some way, then the rate at which you're going to adjust your beliefs should be slower. This is a general observation, and it's certainly something which we see in our data.

Julia: One topic that comes up a lot in the skeptic, pro-critical thinking, pro-science circles is, as you alluded earlier, the fact that humans have been trying to make these inferences about causality -- especially in, say, medicine, what kinds of cures will fix what kinds of diseases -- for a long time. And a lot of the traditional folk wisdom/folk cures that have been around for hundreds of thousands of years don't actually work.

One of the main insights that the skeptics have been trying to point at is that the reason that these false beliefs have persisted for so long, that the stars determine our destinies, for example, is that the human brain is not very good at reasoning about causality, and in more and deeper ways than just being insensitive to sample size. That, for example, we can quickly form a hypothesis maybe based on only a few data points, where someone was born on a certain day and turned out to have a certain personality, but then we're very selective in what pieces of supporting evidence we notice or dismiss.

We will notice the times that eating birch bark was followed by someone recovering from their sickness, but we won't notice or remember the times when that relationship didn't hold-- because that isn't consistent with our theory, et cetera, et cetera, and so the theory just becomes more and more entrenched over time, even though there was little or no evidence to support it in the first place.

I guess the thing that you're describing, where the brain uses these basically optimal rules of statistical inference, with some caveats, seems like it doesn't fully explain just how often and how deeply wrong humans have been about causality over the years.

Tom: I think that's an interesting point. I think there are a couple of things to say about it. One is it's certainly true ... The kinds of things that you're describing are cases where what we're doing is failing to recognize that a relationship doesn't exist. Statistically, that's a much harder problem that recognizing that a relationship does exist.

Julia: Can you elaborate on that?

Tom: If you think about the structure of that problem, you can never get strong evidence that a relationship doesn't exist. ...Let's contrast the case of getting evidence that a relationship does exist. You can see something which deviates so much from your expectations about what chance would look like that you can get arbitrarily strong amounts of evidence that a relationship does exist. Any evidence that a relationship doesn't exist requires ... Basically, the problem is that any evidence that you got that it doesn't exist, that something didn't happen, could still possibly have happened under the hypothesis that it did exist.

Julia: Because there was some confounding factor?

Tom: Yeah, some confounding thing, or just that the set of events which are the events where nothing happens are still contained within the set of events where something happens, if that makes sense.

I have a paper about this. This is with Joseph Williams, who is a student here at Berkeley. What we looked at was people's randomness judgments. People have a bad reputation for making judgments about randomness. The argument is we're just bad at reasoning about what looks like chance.

In our paper what we did was make this observation, that part of what makes deciding that something is random difficult is that the events that provide evidence

for randomness can still happen even if you're talking about a more structured process. If you're flipping a coin and that coin has a bias of 0.9 in favor of heads, you can still get a sequence which is tails/heads/tails/heads/tails. That's not necessarily a particularly unlikely sequence, but if you're flipping a coin at random…

As a consequence, that event can still occur under the hypothesis of non-random stuff, and so it becomes hard for you to get strong evidence in favor of randomness, whereas you can get strong evidence in favor of non-randomness because you can get like long strings of heads or whatever…

Julia: To go back to the causal inference context, I guess it's not clear to me that the asymmetry exists. For example, I'm thinking of all of these many, many social science papers that purport to show a causal relationship between, say, I don't know, class size and educational outcome, that kind of thing. No matter how much they try to control, no matter how many confounding factors they try to control for, there's still this very strong possibility that's hard to get rid of, this strong suspicion that there might be some other factor that they haven't taken into account or controlled for that's in play that means the relationship isn't causal.

Like maybe the richer neighborhoods have both lower class sizes and also better educational outcomes, and that's why you see the correlation between the class size and the educational outcomes, but that doesn't mean that if you decreased class size that you'll improve educational outcomes.

It seems to me that in practice it's actually quite hard to be confident in a causal relationship just based on correlations. Although maybe you're talking about not just observational data. Are you talking about intervening, like with an RCT?

Tom: In most of our experiments we use interventions just to isolate individual causes that people have been reasoning about. I agree that there's a much more complex problem when you're trying to think about multiple possible causes, and that's actually not a case that we've looked at in detail in terms of our work. I agree in those cases. Basically, the thing that I would expect is that people should be good at detecting that there's *something* going on, although, just like statisticians, they're going to find it difficult to figure out what thing is actually going on.

Julia: I see, okay.

Tom: I guess the point that I was making about it being harder to recognize that there's not a relationship is it comes down to what a statistician would talk about as being *nested hypotheses*. Like the hypothesis that there's no relationship at all is nested within the hypothesis that there's some kind of relationship, because it's a special case. A coin that produces heads 50% of the time is a special case of the set of biased coins, it's just the special case of being the one that's actually fair. That's something which can make it hard to recognize that you're actually seeing data which is coming from a fair coin, because you can never get strongly diagnostic evidence for it.

If you think that people's minds are things that glom onto what seems like

diagnostic evidence, then you'd predict as a consequence of that that you are going to end up seeing people producing false alarms, seeing relationships where there are no such relationships. What we actually showed in our paper was that when you equate the difficulty, the statistical difficulty of that randomness judgment task with other kinds of decision problems that don't have those nested hypotheses, people do just as poorly on those other decision problems as they were doing on the randomness task.

There's something just inherently statistically difficult about that kind of inference, where you're recognizing that there's no underlying relationship between things.

Julia: Is the upshot, then, that the brain is basically doing good statistical inference under the constraint of having limited time and resources – like, it's doing the best it can, essentially?

Tom: Yeah.

Julia: It's not systematically biased, just pressed for time, I guess?

Tom: That's right. One of the big directions that we're focused on in our current research is trying to really find a satisfactory way for formalizing that notion. The way that I've expressed this in terms of those levels of analysis we were talking about earlier is, when we were operating at that computational level, we have this useful principle of optimality, this idea that people are doing a pretty good job with solving these problems, and then that narrows down the space of hypotheses that we have to explore to those ones that constitute good solutions to the structure of those problems.

When you push down to the algorithmic level and you start to say, "Okay, fine, we've got those ideal solutions, but we know that people don't always act in ways which are consistent with those." The way that I think about it is those models are really useful because they give us a starting point for investigating some of these questions and we hope that they capture some of the variation in the data, but they also give us a guide to: What are the things that are actually deviations that we should be investigating?

The value of rational models is partly that they highlight the things that people are doing that might be different from those rational models, and then that gives us a place to look in terms of trying to understand what the actual mechanisms that are involved that are implementing those solutions or whatever it is that the actual cognitive processes are that people are using.

This is very similar, by the way, to the way that Kahneman and Tversky originally laid out the heuristics and biases research program. They said, "Probabilistic inference is something which is hard, so we assume people are going to be doing this via heuristics, and then the biases are the clues as to what those heuristics are."

Julia: Right. Just in case it isn't clear to all listeners, heuristic is a fast, efficient guideline for making a decision that gets you pretty good results in many cases, but isn't necessarily perfect.

Tom: Yeah, that's right. I think what's happened is that people have tended to focus more on the biases as failures of rationality, rather than as clues to underlying cognitive processes. That's what we've start to do recently, is say, "Let's take that principle of optimality that we're using at the computational level and take it down a level to the level of algorithms and ask the question: What makes a good algorithm? What makes a good heuristic?"

Heuristics… are normally defined pretty much heuristically. They're defined as pretty good solutions. We can be more precise than that. We can say a good heuristic is one that really hits the sweet spot in terms of a trade off between how much time or computational effort you're putting into solving a problem, and how much error you're making, how much you're deviating from the ideal solutions.

We can formalize that precisely. And then that gives us a guide for beginning to explore some of the cognitive processes that people might engage in. We can ask the question of whether a bias that we see in people's behavior is actually something which we can understand as a consequence of operating under limited resources.

The other thing that I normally say about these kind of false alarms of causal learning is that I think our evaluation of people as being bad at learning causal relationships partly comes from thinking about modern adult humans, and in most of my research I'm kind of thinking about premodern, preadult humans.

As a modern adult human, you actually don't have a whole lot of causal relationships to learn, right? Most of the things that you need to know in order to operate in the world around you, you pretty much got pinned down, and as a consequence those abilities that we have to figure out causal relationships end up sort of firing falsely on all of these kinds of things, like pseudoscience and so on.

If you think about what's going on with, say, a kid who comes into the world not knowing anything about the causal structure of the environment that they're placed in, and then within 4 or 5 years has basically bootstrapped all of these kinds of causal relationships which are fundamental to understanding how the world works, the trade-off between getting these false alarms and actually being able to figure out that structure ... It leans in the direction of, "Yeah, you really need to figure this stuff out as fast as possible, and don't worry if you sort of mess up a couple of those kinds of relationships."

The way that I think about a lot of human cognition is that ... It's really the learning capacities that we have are really designed for those first few years of our lives, and then end up being somewhat useful subsequently. A lot of the magic happens in that initial period. Equivalently, if you think about people living in a world where science has done less of the work of figuring out where the causal relationships are, there would be a greater dependence on quickly figuring out those relationships that you need to know in order to survive, and changing the direction of that trade-off.

Julia: This again comes back to the theme of our brain's built-in algorithms being pretty darned optimized for the most important things in our ancestral environment, and to the extent that the modern environment diverges from the ancestral

environment in important ways and to the extent that the stakes have gone up, for example, due to modern technology and an inter-connected world, it might make sense to try to patch those algorithms, but that doesn't imply that the algorithms were poorly optimized in the first place.

Tom: Yeah, but it's also important to think about ... I think we're self-centered in terms of thinking about what we do as adults as being the most important parts of our lives. In thinking about the structure of the problem we have to solve, recognizing that a big part of that problem is in childhood is also a really useful way of characterizing exactly what computational capacities that we need.

Julia: There's this ongoing debate in the heuristics and biases field and related fields. I'll simplify here, but between, on the one hand, the traditional Kahneman and Tversky model of biases as the ways that human reasoning deviates from ideal reasoning, systematic mistakes that we make, and then on the other side of the debate are people, like for example Gigerenzer, who argue, "No, no, no, the human brain isn't really biased. We're not really irrational. These are actually optimal solutions to the problems that the brain evolved to face and to problems that we have limited time and processing power to deal with, so it's not really appropriate to call the brain irrational, it's just optimized for particular problems and under particular constraints."

It sounds like your research is pointing towards the second of those positions, but I guess it's not clear to me what the tension actually is with Kahneman and Tversky in what you've said so far.

Tom: Importantly, I think, we were using pieces of both of those ideas. I don't think there's necessarily a significant tension with the Kahneman and Tversky perspective.

Here's one way of characterizing this. Gigerenzer's argument has focused on one particular idea which comes from statistics, which is called the bias-variance trade off. The basic idea of this principle is that you don't necessarily want to use the most complex model when you're trying to solve a problem. You don't necessarily want to use the most complex algorithm.

If you're trying to build a predictive model, including more predictors into the model can be something which makes the model actually *worse*, provided you are doing something like trying to minimize the errors that you're making in accounting for the data that you've seen so far. The problem is that, as your model gets more complicated, it can overfit the data. It can end up producing predictions which are driven by noise that appears in the data that you're seeing, because it's got such a greater expressive capacity.

The idea is, by having a simpler model, you're not going to get into that problem of ending up doing a good job of modeling the noise, and as a consequence you're going to end up making better predictions and potentially doing a better job of solving those problems.

Gigerenzer's argument is that some of these heuristics, which you can think about

as strategies that end up being perhaps simpler than other kinds of cognitive strategies you can engage in, they're going to work better than a more complex strategy -- precisely because of the bias-variance trade off, precisely because they take us in that direction of minimizing the amount that we're going to be overfitting the data.

The reason why it's called the bias-variance trade off is that, as you go in that direction, you add bias to your model. You're going to be able to do a less good job of fitting data sets in general, but you're reducing variance -- you're reducing the amount which the answers you're going to get are going to vary around depending on the particular data that you see. Those two things are things that are both bad for making predictions, and so the idea is you want to find the point which is the right trade off between those two kinds of errors.

Julia: Right. To visualize this trade off, you could imagine, and correct me if you think this is the wrong metaphor, but you could imagine shooting darts at a dartboard. If you're an unbiased darts player but have a lot of variance, then you're going to be basically aiming at the center. Some of your darts will be wildly off to the left, some will be wildly off to the right, but on average you're aiming at the center. Still, there's tons of error there.

Or, you could be a biased dart player but with little variance, and so all of your darts basically hit the same spot one inch away from the center or something. You're not aiming directly at the center, but still there's less error overall in how far your darts are away from the target.

Tom: Yeah, that's right.

Julia: Great.

Tom: What's interesting about that is that you basically get this one explanatory dimension where it says making things simpler is going to be good, but it doesn't necessarily explain why you get all the way to the very, very simple kinds of strategies that Gigerenzer tends to advocate. Because basically what the bias-variance trade off tells you is that you don't want to use the most complex thing, but you probably also don't want to use the simplest thing. You actually want to use something which is somewhere in between, and that might end up being more complex than perhaps the simpler sorts of strategies that Gigerenzer has identified, things that, say, rely on just using a single predictor when you're trying to make a decision.

Kahneman and Tversky, on the other hand, emphasized heuristics as basically a means of dealing with cognitive effort, or the way that I think about it is computational effort. Doing probabilistic reasoning is something which, as a computational problem, is really hard. It's Bayesian inference… It falls into the categories of problems which are things that we don't have efficient algorithms to get computers to do, so it's no surprise that they'd be things that would be challenging for people as well. The idea is, maybe people can follow some simpler strategies that are reducing the cognitive effort they need to use to solve problems.

Gigerenzer argued against that. He argued against people being, I think the way he characterized it was being “lazy,” and said instead, "No, we're doing a good job with solving these problems."

I think the position that I have is that I think both of those perspectives are important and they're going to be important for explaining different aspects of the heuristics that we end up using. If you add in this third factor of cognitive effort, that's something which does maybe push you a little bit further in terms of going in the direction of simplicity, but it's also something that we can use to explain other kinds of heuristics.

For example, one of my students here at Berkeley, Falk Lieder, has been doing a lot of work investigating whether some of the classic heuristics identified by Kahneman and Tversky can be understood as being really good points on this trade off between the effort that we have to make and the error that we make as a consequence of it.

Julia: Like what?

Tom: Looking at things like the availability heuristic or anchoring and adjustment, which are canonical things that have been used to argue against aspects of human rationality. And saying, "Well, maybe we can understand the decisions that are made in the context of those particular algorithms by viewing them as algorithms, and then asking how we should best use those algorithms in solving problems that we have to solve as humans."

Julia: Can you briefly explain the availability and anchoring heuristics?

Tom: Yeah. Maybe I'll just do one of them.

Julia: Great. How about availability?

Tom: Yeah, I'll do availability, and then I'll tell you a story about it. The basic idea behind availability is that if I ask you to judge the probability of something, to make a decision which depends on probabilities of outcomes, and then you do that by basically using those outcomes which come to mind most easily.

An example of this is, say, if you're going to make a decision as to whether you should go snorkeling on holiday. You might end up thinking not just about the colorful fish you're going to see, but also about the possibility of shark attacks. Or, if you're going to go on a plane flight, you'll probably end up thinking about terrorists more than you should. These are things which are very salient to us and jump out at us, and so as a consequence we end up overestimating their probabilities when we're trying to make decisions.

What Falk did was look at this question from the perspective of trying to think about a computational solution to the problem of calculating an expected utility. If you're acting rationally, what you should be doing when you're trying to make a decision as to whether you want to do something or not, is to work out what's the probabilities of all of the different outcomes that could happen? What's the utility

that you assign to those outcomes? And then average together those utilities weighted by their probabilities. Then that gives you the value of that particular option.

That's obviously a really computationally demanding thing, particularly for the kinds of problems that we face as human beings where there could be many possible outcomes, and so on and so on.

A reasonable way that you could try and solve that problem instead is by sampling, by generating some sample of outcomes and then evaluating utilities of those outcomes and then adding those up.

Then you have this question, which is, well, what distribution should you be sampling those outcomes from? I think the immediate intuitive response is to say, "Well, you should just generate those outcomes with the probability that they occur in the world. You should just generate an unbiased sample." Indeed, if you do that, you'll get an unbiased estimate of the expected utility.

The problem with that is that if you are in a situation where there are some outcomes that are extreme outcomes -- that, say, occur with relatively lower probability, which is I think the sort of context that we often face in the sorts of decisions that we make as humans -- then that strategy is going to not work very well. Because there's a chance that you don't generate those extreme outcomes, because you're sampling from this distribution, and those things might have relatively low chance of happening.

For example, if somebody asked you whether you wanted to play Russian roulette or not, I think you'd have to generate about 50 samples of the possible outcomes of that game to have a 99.9% chance of deciding that you definitely don't want to play that game. That's just because there's an outcome which is really bad, which is shooting yourself in the head, which occurs with relatively low probability, in this case 1/6. In general, the events that we have to deal with might be equivalently bad, but occur with even lower probabilities.

The answer is, in order to deal with that problem, you probably want to generate from a different distribution. And we can ask, what's the best distribution to generate from, from the perspective of minimizing the variance in the estimates? Because in this case it's the variance which really kills you, it's the variability across those different samples. The answer is: Add a little bit of bias. It's the bias-variance trade off again. You generate from a biased distribution, that results in a biased estimate.

The optimal distribution to generate from, from the perspective of minimizing variance, is the distribution where the probability of generating an outcome is proportional to the probability of that outcome occurring in the world, multiplied by the absolute value of its utility.

Basically, the idea is that you want to generate from a distribution where those extreme events that are either extremely good or extremely bad are given greater weight -- and that's exactly what we end up doing when we're answering questions

using those available examples. Because the things that we tend to focus on, and the things that we tend to store in our memory, are those things which really have extreme utilities.

Julia: I see. Interesting.

The thing I'm most interested in now is what this means for prescriptive rationality. Even granting that argument, that the availability heuristic was a roughly optimal solution to these kinds of problems under the constraints of time and computational effort, it's still not clear to me that we couldn't improve on those strategies that are built into our brains.

If the availability heuristic gives us biased, inside-view, intuitive impressions of risk, say from shark attacks, then couldn't it still be prescriptively a better solution to take that intuition and compare it to, say, the statistics that exist on chance of death from shark attacks if you go snorkeling 10 times a year? And have a new heuristic that you add to your brain, of trusting the explicit statistics in cases where you know that your brain is using the availability heuristic, or is going to tend to use the availability heuristic, that kind of thing?

Tom: Yeah. I think there are two things that come out of this. One is that having this perspective for thinking about the algorithms inside our heads means that we should be doing perhaps good algorithm design for humans.

We normally think about algorithm design as something that we do for computers, where we're trying to come up with better strategies for our computers to solve our problems. But if we think about the algorithms that we use being those things that are the sweet spot between error and computational effort, it's not going to be successful to teach us really complicated strategies that require a lot of computational effort. Because the reason why we end up using the heuristics that we have is because they end up being a pretty good trade off in that way.

Actually, we're also doing some work at the moment about looking at how people engage in what's called rational measure reasoning, which is basically deciding what strategies they're going to use in solving a problem and how they could actually acquire pretty good strategies for these different kinds of solving problems.

I think one approach is to say, "Let's try and come up with good algorithms that people can use," but in thinking about those good algorithms, have this dimension of cognitive effort and computational ease. And trying to understand what the nature of the human computational architecture is like, so that we can end up making recommendations of algorithms that end up being good algorithms with respect to both of those criteria.

I have a book coming out in April with Brian Christian, where we look at those kinds of notions. It's called *Algorithms to Live By*. It's explicitly exploring that kind of question of what good algorithms for human lives might look like.

I think the other idea is that, to the extent that we've already adopted these algorithms and these end up being strategies that we end up using, you can also ask

the question of how we might structure our environments in ways that we end up doing a better job of solving the problems we want to solve, because we've changed the nature of the inputs to those algorithms. If intervening on the algorithms themselves is difficult, intervening on our environments might be easier, and might be the kind of thing that makes us able to do a better job of making these sorts of inferences.

To return to your example of shark attacks and so on, I think you could expect that there's even more bias than the optimal amount of bias in availability-based decisions because what's available to us has changed. One of the things that's happened is you can hear about shark attacks on the news, and you can see plane crashes and you can see all of these different kinds of things. The statistics of the environment that we operate in are also just completely messed up with respect to what's relevant for making our own decisions.

So a basic recommendation that would come out of that is, if this is the way that your mind tends to work, try and put yourself in an environment where you get exposed to the right kind of statistics. I think the way you were characterizing that was in terms of you find out what the facts are on shark attacks and so on.

Julia: I agree that's not what our brains are optimized to respond to, though, for sure.

Tom: Yeah. I can give you an example of doing this, though.

Julia: Please.

Tom: One thing that I did for a while is: Wikipedia has a random page. You can go to this link and it loads a random Wikipedia page. I set my homepage on my browser to just generate a random Wikipedia page. What I get then every time I open the browser is just a random sample from Wikipedia. This is interesting from the perspective of, first of all, I learned a lot about the kinds of things that's on Wikipedia. There's a lot of small towns that have web pages ...

Julia: Interesting.

Tom: ... and things like sports are very strongly represented.

The other thing that I got out of it was: We were talking earlier about randomness and causal relationships. I got a better sense of what randomness really looks like for semantically interesting phenomena. One thing that came out of that was that I was surprised at how often the randomly generated pages felt like they were something that had a personal connection to me.

For example, having observed this phenomenon for a while, I then actually did a trial where I just generated pages and saw how many pages I have to generate until I got something that was related. It only took generating a few random Wikipedia pages to hit the page on the Western Australian legislature. I grew up in Western Australia. I was like, "Okay ... "

I think it was really useful in terms of calibrating a sense of coincidence. I think

there's an interesting idea there, which is, for thinking about things like plane crashes and car crashes, if you can manage to filter out all of the stuff which you haven't seen in person, you can end up having a much better sense of those statistics.

More generally, it's interesting to think about what kinds of experiential methods we can come up with for actually getting ourselves exposed to, say, good statistics, so that the algorithms that are inside our heads end up getting better inputs for solving the sorts of problems we want to solve.

Julia: Excellent. Yeah, a very good point, which unfortunately we don't have time to expand upon because we are all out of time for this section of the podcast. We should consider revisiting this conversation after your book comes out later this year.

Tom: That sounds great.

Julia: All right. Well, we're going to wrap up this part of the conversation and move on now to the Rationally Speaking pick.

(interlude)

Julia: Welcome back. Every episode we invite our guest on Rationally Speaking to introduce the Rationally Speaking pick of the episode. That's a book or website or movie or something that's influenced their thinking in some interesting way. Tom, what's your pick for today's episode?

Tom: I can tell you about the book which got me into computational cognitive science, and that's a book called Matter and Consciousness by Paul Churchland. It's a book which is basically an introduction to philosophy of mind. It talks about things like questions about how you can actually study things like minds, and how you would characterize thoughts and so on. I read it when I was an undergraduate in Australia doing a class in philosophy of mind.

Over the summer I just went back and looked at it, because there was a chapter that we never read in the book. This chapter was about basically the implications of neural network models for thinking about these sorts of fundamental questions about how the mind works. To me, that was incredibly exciting. It has a whole appendix where it goes through how these artificial neural network models work.

As a high school student I'd done a lot of math and so on, and then when I went to university I was really just interested in learning about things like philosophy and psychology that I'd never had the chance to study before. This was the moment where I saw that it was possible for those things to come together, that it was possible to think about studying minds from a mathematical perspective, and so I just spent that summer doing calculus and linear algebra, and reading books about neural networks and deriving models, and coming up with a grand unified theory of how the mind works. That was just I think a very exciting experience for me.

Then on the first day of the next semester I found the person on campus who did

research that was most closely related to that, and at 9:00am on the first day of class cornered him in a corridor and talked to him about letting me be a research assistant. That's what set me off on this path of thinking about the mathematical structure of the mind.

Julia: Excellent. It's too bad that my former co-host, Massimo, isn't here, because he's always frustrated that scientists don't have enough respect for the relevance of philosophy to their work -- and here you are, a scientist who embarked on his path because of reading a book of philosophy of mind. I'll have to make sure Massimo hears about this episode.

Well, thank you so much for your time, Tom. It's been a fascinating conversation, and I hope to have you back on the show soon.

Tom: Thank you. It's been great.

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