Rationally Speaking #188: Robert Kurzban on "Being strategically wrong"

Julia Galef: Welcome to Rationally Speaking, the podcast where we explore the borderlands between reason and nonsense. I'm your host, Julia Galef, and I am taping this podcast live at the Northeast Conference on Science and Skepticism. (Say hi!)

I am very excited to introduce today's guest who's here with me. I have professor Robert Kurzban who is a professor of psychology at the University of Pennsylvania, where he specializes in evolutionary psychology. One of his books, excellently titled *Why Everyone (Else) Is a Hypocrite* -- such a great title, Rob -- it's been very formative in my thinking about rationality, and has shaped a lot of my work and my talks that I give to people over the years. So this conversation has been a long time in the works.

Rob is also the coauthor of *The Hidden Agenda of the Political Mind*, which is a book that I talked about in, I guess, my most recent episode of the podcast with Jason Weeden. And Rob, no pressure, but Jason was an excellent guest -- so the bar is just set high, that's all I'm saying.

So Rob, welcome to the show and to NECSS.

Robert Kurzban: Thanks for having me. It's a pleasure to be here.

Julia Galef: As I hinted at, the thing that I'm most excited to talk to you about is *Why Everyone (Else) Is a Hypocrite*. I thought we could just jump right into your thesis, maybe by talking about what motivated it. What is the mystery or puzzle in the world that demanded a theory to explain it?

Robert Kurzban: As a psychologist I consider myself to be a student of human nature, and the main lesson that I take from my time as a psychologist is that people are super weird and super puzzling. And one of the things I think that's most surprising or at least puzzling to me is just how inconsistent we are. Part of my background is economics, and economists view people as these coldly rational, consistent beings, and that's just not my experience of humanity.

Julia Galef: Did anyone ever consider the theory that economists just don't go out very much and meet people?

Robert Kurzban: Yeah, that's actually ... That's the main theory, as it turns out. They need to get out more.

I got interested in human inconsistency, and for me as a person who is a student of evolutionary approaches to psychology, what was exciting was this idea that the fact that mind consists of lots of different pieces.

That connects very directly to this notion that there might be different pieces that are going on and off at different times. If that's true then you start to get a window into why we're not these uniformly consistent creatures. It's

because the mind just doesn't work that way. It's a collection of parts, and trying to use that idea to explain this fundamental mystery, to me, was extremely exciting.

Julia Galef: Let's talk about what it means to say that the mind is made of parts. Are you talking about physical parts in the brain that each do something different and operate independently, or something else?

Robert Kurzban: What I'm talking about really is a functional sense of different parts. If people open up their smartphones, as I see many in the audience right now are doing, you'll see that there are ... I'm not insulted at all.

Julia Galef: They're just taking notes on your brilliant insights.

Robert Kurzban: Yeah, taking notes. That's right. You see that there's little applications that are specialized. One of them is a communications app. One of them is maybe checking your stocks. One of them is for throwing pigs at birds, or birds at pigs, or whatever that game is.

Julia Galef: I can tell you're actually productive. That must be nice.

Robert Kurzban: The idea is that ... And we know this. The mind has a visual system which lets us see the world. It has a memory system which allows us to remember scintillating talks like this one, and a language system which allows us to process language and produce it. And so that's the sense that I mean.

Now those things might be distributed all over the brain. I'm not too worried about the spatial elements. In the same way that in your smartphone you don't care where Angry Birds lives -- you just care that you can get to it when you click the thing.

So what I'm interested in is this idea that the mind consists of parts in the sense that parts that are specialized to do different jobs. That's the fundamental message. And because of the so called functional specificity, once again you get this idea that there's not a unitary entity in there. It's a lot of different pieces that are working somehow together.

Julia Galef: How controversial is that? Because to some extent, as you say: we have vision. We can process language. There's all these different things that the brain is doing that you can think of as modules, or as analogous to apps. So in that sense it can't possibly be controversial.

But maybe there's a spectrum of *how* modular people think the brain is?

Robert Kurzban: Yeah, that's exactly right. I think that almost every modern psychologist is willing to concede that there are specialized systems for vision. You poke somebody in the eye, you look in there, and you're like, "Oh, there's a retina." That does one thing all day. It just sits around and it waits for light to hit it and then it says, "Oh, there's something out there."

In that sense everyone agrees there's specialization. But as you go up, metaphorically, the cognitive system gets more controversial. For memory, people feel like, yeah, there's probably a specialized memory system -- but now let's talk about the systems that underlie social behavior. Is there a specialized system in your head that's designed to identify when someone has cheated you? Or is that a more general process?

In there lies the controversy, and I feel comfortable saying that there's animated controversy. You know the old expression: the fights in academia are so venomous because the stakes are so small. Well this is one of those cases where-

Julia Galef: No, this feels important though!

Robert Kurzban: I certainly think it's important. I think in many ways what's at stake here really is something super important which is, what is the fundamental nature of the mind? Does it consist of lots of different specialized parts all the way down? Everything from vision to the sophisticated social systems that we have? Or are there more general processes?

And of course all of this intersects with other kinds of theories including my area, evolutionary approaches. The evolutionary point of view points to specialization, because we see that throughout the animal kingdom, and there's good reasons to think that. So yeah, I feel comfortable saying that there's sufficient controversy to keep the field going. I think that's right.

Julia Galef: At the other end of the spectrum from you perhaps would be the “one algorithm hypothesis” that people like Jeff Hawkins sometimes talk about. That yeah, it might seem like we have all these different functions -- we can optimize for social goals, and then there's also vision and language and strategizing and so on -- but really it seems most likely that there's just one algorithm that can just do all these different things.

The same way that with deep learning, reinforcement learning algorithms -- machine learning researchers now can take roughly the same basic algorithm and throw a bunch of compute and data at it, and train it to do language. Or train it to classify images. Or train it to play video games.

The idea being, according to this hypothesis, that our brains work the same way. Why is that implausible to you?

Robert Kurzban: Well, I would say a couple things. The first step if we're going to have a debate like this is just to talk about what the different evidence would look like in favor of the different views. I think my reading of the evidence in terms of psychology doesn't point me in that direction. Everywhere you look it looks like there's specificity.

But there's other issues too. There's philosophical problems in the sense that it is very difficult to lay out exactly what that algorithm looks like, in a

way that can actually solve the problems on the ground that the mind has to solve, particularly in real time.

Certainly people have tried this. The history of psychology, dating all the way back to behaviorism -- the behaviorists said the same thing. They said, "Look, give me a child and I'll teach it by shaping it in different directions, with reinforcement learning and so on," and it just turns out that's not true. There's a reason that the cognitive revolution replaced behaviorism. Because pigeons didn't behave the way they were supposed to according to the behaviorist principles, and people *really* didn't.

Then it got a second incarnation in the context of the connectionist ideas that we saw in the '80s and into the early '90s. And once again my read of the evidence there is that the systems just don't have enough content in them to actually solve the problems.

Don't get me wrong. I think that learning is super important. But one of the things I think we're learning about learning, as it were, is that you need different kinds of learning systems to learn different kinds of information. So the thing that learns language is probably not going to be the same thing that learns who should be your friend, and who should be your mate, and who should be your ally, and who should be your enemy. You just need different information data structures in order to solve that problem.

Julia Galef: What would be an example or two of a more social, or psychological, function that a module might serve in your theory? So, not like vision.

Robert Kurzban: The one that I mentioned earlier is usually held up as the prototype, this idea of cheater detection. This comes out of some work that my former advisors did at Santa Barbara where what they were able to do is show, with different psychological tasks, that there looks like there's a system in your head. And all it's doing all day is looking around and asking the question, "Is there somebody who has taken some benefit without having met the requirement or paid the cost?"

And you can give people logic problems that they're very bad at, and then if you just recast them into a social language, where there's costs and benefits involved with people potentially cheating -- all of a sudden people become extremely good at reasoning about it.

That wouldn't be true on the kind of very general system that you have in mind. For example, if we were just good logicians, if I give you a problem of the structure “if P then Q,” you should be pretty good at handling that -- but that turns out not to be true.

But if you add social content to these problems, you get improved performance. And what that points to is the idea that there's a system in your head that is specifically good at doing that, and when you engage it you

start to see that our performance is much, much better. That's pretty good evidence of specificity.

Of course one of the places that you referred to as well is language, and people are still talking about whether there's a specialized language acquisition device. Chomsky of course made his career on this, and I think many people find those ideas as extended by Steve Pinker and others fairly persuasive, that it looks like there's at least some architecture in your mind that's pretty good at language.

Don't get me wrong. I think this is an ongoing research enterprise. That's why I have unbelievable job security is because there's ... Also tenure, but ...

Julia Galef: Because we'll never run out of human weirdness to explain?

Robert Kurzban: There's always more weirdness. There's always more controversy. There's always more data to be gathered… and I think we've established that I'm more charming that Jason at this point.

Julia Galef: I… like you both equally. You're both special and talented in your own ways.

Robert Kurzban: Thanks, Mom.

Julia Galef: You're welcome.

Just to get a little more clarity on the idea of a module, if you make the claim "The mind has all these different modules which serve different functions," is that saying anything above and beyond the claim "The mind serves different functions"? Is the word module doing any work there?

Robert Kurzban: The word module is only a shorthand, so that I don't have to keep saying "functionally specialized integrated computational system" over and over again. So it's a placeholder for that idea -- but like anything else the crucial thing about the commitment to modularity is that it makes predictions in terms of the empirical work.

If you're an evolutionary biologist and you make a claim that some structure is specialized for some task, you're making an empirical prediction that the structure is going to show design features that it is really good at doing that task. And so what this does is it gives us a lot of leverage in terms of predictions that we can then test in the laboratory.

That I think is the crucial part of this. That yes, it is a framework for understanding the mind, and in that sense it's useful in and of itself -- but as a scientific matter, where it becomes incredibly useful is the idea that, look, we now commit out loud to these properties, and that allows us to go in and test them. And other people can do their tests, and then we can arbitrate these issues. I think in that sense it's not just a word. It has content in terms of the way we do our business of science.

Julia Galef: Last thing on the how modular is the mind question. I imagine if you just had one really good reinforcement learning algorithm and threw a ton of data and computational power at it, it could look like it was really good at a bunch of different specialized tasks. But that doesn't necessarily mean there are specialized modules designed for those tasks.

Robert Kurzban: Look, again I think these are empirical questions. There's other things there. So, learning only gets you so far, because the system has to do something with the learned information and it also has to structure its learning.

Philosophers have wrestled a very long time about the problem of induction, and the thing that a child faces as they're learning about the world is that they really don't know what it is they ought to induce from any given set of stimuli. There are limits, I think philosophical limits, to what reinforcement learning can do.

And I should also say developmental psychologists are showing us all the time that systems are coming in, in young babies and infants, way sooner than they should be given the amount of information they have to work on. So it looks to many of us that there's a lot of computational flexibility that is coming online very quickly, and it's hard to tell a good story about exactly how it is the organism learned from a blank slate.

I think if people want to look into what I consider to be the best compilation of data against this argument, it's Steve Pinker's *The Blank Slate* which I suspect your listeners and people in this room are probably familiar with. One of the agendas of that book, or the agenda of that book, was to say, "Look, if you think that it's just a blank slate and then you're going to learn everything, you've got to explain all these data which push the other way," and I think that should be the sort of thing that persuades you. Of course the philosophy should be persuasive to some extent about the problems of induction, but the evidence that psychologists and economists and sociologists and others have gathered, that should be the thing that makes up your mind.

Julia Galef: Let's move on to hypocrisy and self deception. How does the modularity hypothesis help explain that?

Robert Kurzban: First I think it's important to define terms. People use these terms in different ways. I think of hypocrisy as the case in which you say doing X is wrong and then you yourself do X, so it's an inconsistency between what you morally condemn and then what you actually do.

The story that I tell about that -- and I use the word "story" just to mean an explanation, I don't mean that in a pejorative way -- is this: Look, one of the specialized systems in your head is this moral system. We all go around the world and we try to identify when people do wrong things. We're very sensitive to other people's moral failings, but by the same token when we

ourselves are choosing what to do, we don't always, shockingly, use our own moral compass in deciding what we're going to do.

The argument is you have one system in your head which is designed specifically for moral condemnation. And that leads to saying things like, "You shouldn't tweet insults at people, particularly if you're a high status individual and they're a lower status individual." Some people might have that view. I'm just speaking randomly here. Totally randomly.

Then you have a guide of your behavior, which let's say, might be this very hot emotional system, which causes you to engage in very aggressive behavior on social media for example. So on the one hand you might say, "Look, here's my moral principle and that's guiding me in what I say, and then here is my behavioral system and that's guiding me in terms of what I do."

Those things are not necessarily going to [agree with] one another. There's nothing that prevents people who condemn doing X from themselves doing X when they decide to do it, when it's in their best interest. And that I think is the crux.

… And I just gave away the whole book so now you know why everyone else is a hypocrite. You don't have to buy it. That's the whole crux of it which is that we're moral creatures *in part*. That is to say we're good at condemning other people. But we're not always moral creatures in behavior, because the part of your mind that guides condemnation is a different part from the one that guides behavior, and that's the fundamental conflict.

Julia Galef: It makes sense to me why that would be a useful inconsistency to have. From my own self interest, I want to get away with as much as possible, and I also want other people to not get away with stuff.

But what about something where there seems to be more conflict? Like, I ostensibly have this goal that I want to achieve, but at the same time I do things that are inconsistent with that ostensible goal. Or I have a self image of myself that isn't necessarily true or justified based on the information?

There are a lot of cases like that, where ... Maybe that's not really what we call hypocrisy. We might call that self deception. But those cases seem a little harder to explain from a self interest point of view.

Robert Kurzban: I agree and I would say that those things have different explanations.

Let's take the first one. We all want to be in good shape as an abstract long term goal, but none of us really wants to go to the gym. Some of us really like to go to the gym, and many of us say that we love to go to the gym, at least on our Tinder profiles. "I love to go to the gym." Right?

The explanation for that is still a modular story, which is that you have some modules in your head which are really forward looking. Those guys really look ahead. Your frontal systems are like, "I want to be healthy in the future." Then you have these other systems which are designed to cause you to conserve energy and not want to exert. You're like, "I just want to stream Netflix right now." And those two things are in conflict, and it's because you've got these two different modular systems.

So I think the modular approach explains why we have, on the one hand, long term goals that push us in a particular direction, and then we have these short term reward systems which push us in the opposite direction. That's I think how I would explain that kind of inconsistency.

The other thing you talked about-

Julia Galef: Just to bookmark that, what's the alternate explanation? Why don't all behavioral economists and psychologists and evolutionary psychologists ... How could they not acknowledge that we have inconsistent goals, such that it seems like the best way to model that is that different parts of our mind want different things and they're in conflict?

Robert Kurzban: I would say that the way that I put it is not super controversial. That is to say, I think many people would think of it more or less that way.

I use a slightly different language from the way that other people do. So people like Danny Kahneman and Amos Tversky and so on, talked about dual systems theory...

One of the big debates, and you alluded to this earlier, is how many different systems are there? I want to say there's a ton of them, and people like Kahneman and Tversky, well just Danny, want to say two. When they're talking about multiple systems, their granularity is just a lot coarser than mine is. I want there to be a lot of systems in there, all of which have different discount factors or what have you. They sort of think well, it's actually just a couple of them, and they're in conflict from time to time.

But more or less I think there's a certain amount of agreement there. What we agree on is that people are not consistent temporally, that they're doing very strange things over time. And I think a lot of economists, one of the things they would say is look, this is just irrationality. And they might have other ways to model it, but I think this is a particularly valuable way.

But I do want to distinguish that from the other thing you said, which is also interesting, which is self deception -- which I think is a slightly different story. The way I think about self deception is that in the modern world and also in the past, one of the things that we do as human creatures is we meticulously cultivate a certain reputation. And one of the main sources of information that other people have about us is our own stated beliefs about our traits, about our future, and so on.

And if you think that what you say about those things is going to influence other people's judgment about your reputation, then sometimes it can be very valuable to broadcast things about yourself that are not necessarily true. No one knows how good a driver I am, but if I go around acting as if or saying that I'm in the top 10% of all drivers in the country, then you might have reason to believe me, and I mention that-

Julia Galef: Is that what you say on your Tinder profile, Rob?

Robert Kurzban: I have a 4.95 Uber rating. I looked in there right now. Excellent Uber passenger.

I mentioned that example in part because there's research on this. Seventy percent of people who were surveyed put themselves in the top 15% of drivers. We know for sure that can't be true. As a statistical matter that can't be true, so then the question is, why do people believe things that must be false? This community should be extremely interested in that question. The skeptics community is obsessed with the question of why do otherwise sensible people believe things that are false?

One of the things I love about that research is that it really strongly suggests that yes, you also believe things that are false, in particular about yourselves, and why? Well, if it's true that my having a false belief about myself improves my ability to persuade you that I'm a successful, articulate, good driver kind of guy, well then that's good for me.

So on this view, the self deception part is really a reputation maintenance part. And a lot of these things are hard to penetrate. It's hard for anyone to see how good a driver you are. You could check how good an Uber passenger I am by going into my app, but I have a little fingerprint sensor on my phone so I don't think you're going to be able to get into there. But in general, reputations are hard to verify.

Julia Galef: On the driver example in particular, it occurs to me that it could technically be possible for a majority of the population to not be wrong that they're above average drivers, because they could be defining quality of driver differently. Maybe some people are like, "I get to where I'm going faster than other people," and maybe they do. And other people are like, "I get into fewer accidents than most people," and maybe they're right too.

Robert Kurzban: That's right. Some people-

Julia Galef: But we can be very selective in how we define “good driver” such that we fit the criteria.

Robert Kurzban: That's right, and the way that we finesse this problem is we also have research where the criteria are much more objective. And what you find is that for those traits that are a little bit more objective, it's true that those

effects diminish, but more or less they never go away. We're always just a little bit more optimistic about ourselves, and we get it wrong in little ways.

So I agree with you that things like driving have multiple different dimensions that you could evaluate yourself on, but that's another case where you've got to look at the entire body of evidence in this work. And I think that you would say that more or less, even in objective cases -- so for example, optimism. We know how many people are going to break their legs in the next year as a fraction, and people wildly underestimate their probabilities of breaking a leg. So step carefully out there.

Julia Galef: Wait, sorry. Do they correctly estimate the general rate in the population of leg breaking and then think their rate will be lower? Or are they confused about how likely people are to break legs?

Robert Kurzban: They underestimate the base rates, and then they further underestimate their rates.

Julia Galef: How do we distinguish between these two hypotheses: One, people self deceive in this way so that they can more effectively convince other people that they're better than they really are. Versus, on the other hand: people self deceive because they just want to feel good about themselves, and it feels good to believe that you're in the top whatever percent of drivers.

What would distinguish?

Robert Kurzban: Good question. In psychology, the idea that we're motivated by the drive for self esteem has been one of the largest bodies of literature there is out there. This has been the source of a tremendous amount of work.

But by and large, when people go in and try to test that hypothesis by looking at people's beliefs and looking at how that affects our self esteem, by and large that work has shown that these effects are extremely small or even zero. There have been some meta-analyses on these bodies of research that look at the entire set of published findings in an area, and the conclusions in those meta-analyses suggest that you just can't explain people's beliefs or behaviors given the motive to maintain self esteem.

I think there are some other kinds of evidence that point away from it as well. For example, the kinds of effects you see in this literature depend on whether or not the claim that you're making is a public claim, and so people tend to inflate what they're saying about themselves to the extent that they think those kinds of claims are going to be known by experimenters or by other people and so on. If it were just a motive to make yourself happy, that should be irrelevant. You shouldn't care who's going to know. You should just care about the belief itself.

I think those are two important pieces. I will say there's a theoretical piece. As an evolutionary psychologist, my view is that natural selection has never

cared how happy or sad an organism is. It just cares about its survival and reproduction. It would be very strange if natural selection designed an organism to seek happiness, something that's internal to the organism. As opposed to outcomes in the world, things that are going to lead to its survival and reproduction.

So I think on two empirical grounds and one theoretical ground, the “try to feel good so I'm going to be wrong,” I don't think it does very well.

Julia Galef: Would your model predict then that people would not self deceive about whether their life was going well? People would not be likely to avoid thinking about stressful topics, regarding how much of a pickle they're in, if they're going to be able to solve their problem?

Or -- would your model also predict that that's something that would make people look bad to others, if others knew that they were in a tough situation?

Robert Kurzban: Well there's good reasons for people to think about their problems, because it helps them drive towards a solution, so I think that's a tough example.

Julia Galef: Oh, I agree with that, but people often don't. People will often not think about health problems, for example, because it's just too upsetting to think about the fact that things might actually be worse than you'd like to think they were.

Is there anything that would be upsetting to someone to think about, but wouldn't make them look bad to their fellow primates, that we could use to distinguish?

Robert Kurzban: I admit I haven't thought through that in detail. I will say one thing that people tend not to think about is cases in which, yeah, it's bad to think about -- and also there's nothing I could do about it if it were the case that I'm actually in serious trouble. People don't seek out information that they can't do anything with, or at least they rarely do.

Many people, for example, who forgo getting medical tests done, many of them have beliefs about how because of their insurance or for whatever other reason, there's just nothing they would change anyway. And so they forego these tests. And I'm not denying that part of that could also be that there is reputational damage, once you find out for example, if you have something which is ultimately going to be fatal.

It's just a sad fact about human nature is we tend to prefer to spend our time with people who we think are going to be around longer, and that's because we're strategic creatures. How many times are you trying to build a friendship with someone who just said that they were leaving the state in three weeks? Maybe that's a cynical view of human nature, but I think it's important to be realistic about it.

Julia Galef: What about the “just world” hypothesis, the folk hypothesis that if bad things happen, people deserve them -- and people will rationalize some explanation for why the victim was actually at fault?

My understanding is the theory of why people tend to do this is that it's upsetting to think that the world could be random and good people could have bad things happen to them. This seems like a kind of self deception. And it also seems like something that fits better with the model of people wanting to believe things that are nice, as opposed to people wanting to believe things that are strategic in the way they appear to others.

Robert Kurzban: Yeah, I think there might be some of that. And I think that there might be cases in which people believe things because they think it would be nice. But I think there's other stories there. Another sad fact about people is that we're always looking for excuses not to help people that would cost us to help, and you see this both in the interpersonal level.

Here we are in Manhattan, and as I was walking here of course you have a couple panhandlers. And most people walk by and they have a ... And maybe it's a true story. They have a true story about why they're not going to help. Having to do with, look, this is a person who's not genuinely in need. Or maybe it was their fault. Those stories allow them to walk by and to choose not to do something beneficial, and that is useful as a social matter.

We all want to be able to tell a plausible story about why we don't want to do the thing that's costly for us to do, and I think the just world hypothesis is one of those stories. If I have the belief that those people over there, they certainly had it coming, then I can say, "Look, I don't think that the federal government should be giving foreign aid to those individuals. If they wanted to live better they'd grow corn." Or I don't know exactly how that would go, but I think the just world hypothesis, like many of these kinds of stories -- sure, one possibility could be that we want to feel better. But I think that we should entertain the notion that it has to do with having a good way to explain to other people why we've chosen not to act when we could do something that would be valuable.

And I think these stories are super important, and I think these stories we tell also in the political sphere. If we have a particular political position, we never say, "Well I think we should cut taxes on the wealthy because I'm wealthy and I'd rather have more money." No. We say, "Look, that's going to stimulate economic growth and we're all going to be better off." The first explanation is just not a persuasive explanation, but the second one is.

Julia Galef: I'm curious how much we disagree here. Here are three possible things that you could be saying: You could be saying A, that self deception is just purely helpful – it just only helps us, it doesn't hurt us. Or B, you could be saying that it both helps and hurts us, but overall it helps us more. So, on net it helps us. Or lastly (C) you could be saying self deception both helps and hurts us, and it's not clear what the net effect is good or bad.

Which of those three?

Robert Kurzban: I'm going to say number four, none of the above. Which is, I actually think that the way to have this conversation is to get away from the notion of self deception.

Let me try this. I think that a better way to think about self deception is what I've called being "strategically wrong." You're being wrong in a way that is helpful for you. And that by the way does put me closest to your number one, so I'll go that far.

The way I think about this is when we talk about self deception, in almost every single case, what we're really talking about is something like look, this person has this belief which somehow they really shouldn't have. They should have a belief that they're more likely to get a broken leg. They should have the belief that they're a worse driver. They should have whatever belief that's closer to what the reality is in the world.

But they don't have that belief. And then the question is why? And my argument is well, it's because having the false belief is useful for persuading others about how wonderful you are.

In that context, yeah, I actually do -- and I'm reluctant to say this since we're taping, but I do actually think that these strategically wrong beliefs are the product of evolved systems that were specifically designed to be wrong in this way that, yeah, is helpful in the long run.

I'm not saying that self deception can't be in some cases very damaging. Robert Trivers has a book that came out, *Folly of Fools*, two, three years ago. And he takes a position which is not completely unlike mine, but is unlike mine in the sense that, first of all, he casts the net of self deception much more broadly. He takes cases where people are just simply wrong about facts of the matter to be self deception.

I think that doesn't quite get the point. But then he talks about cases in which he says these beliefs have horrible outcomes. He has this story in there about some pilots who had the belief that they were cleared for takeoff, which turned out not to be true because of a radio malfunction, with catastrophic results.

I don't consider that to be self deception. I consider that to be a case where they were definitely wrong. There's no doubt that they had a false belief, but-

Julia Galef: I guess maybe we should distinguish ... I would call something self deception if there's a rationalization process going on. Where they have all the information that they need, to know what the right answer is -- and in fact if they had less of an emotional or personal stake in the question, if it were just a purely logical, like an abstract word problem, maybe they would get it

right. But because they have some motivation to get a particular answer, they end up deceiving themselves.

Robert Kurzban: Yeah.

Julia Galef: But then there's also self deception *about yourself* which is the main thing we've been talking about. So that's I guess not what he was talking about.

Robert Kurzban: Yeah, and again he casts the net more broadly.

Let me put it this way. I think there's another important lesson that modularity has to teach us about this. So, to stick with the driving example. I actually think that it's completely plausible that in people's heads you have two beliefs that are mutually contradictory but they exist at the same time.

I think that stored in one part of your head could be the idea that I'm a very safe driver. A super safe driver. Then another part of your head could be a belief which is more accurate. Let's use skill instead -- "I'm a skillful driver."

And then there's another part of my head, there's somewhere in there that knows that I'm actually not that good. And so when I actually get behind the wheel of a car, which I hardly do at all since I use Uber so often, but when I get behind the wheel of a car I am excruciatingly ... I'm very vigilant. I don't try to show off, and so on.

So on the one hand, I might broadcast to the world, "I'm a super skilled driver," but then when I actually get behind the wheel, my behavior is driven by the true representation, the true belief, that I'm actually kind of impaired in this way.

I think that one of the big problems in this area is this idea that there's just one belief in there somewhere. That it's one ring to rule them all. When in fact you can have beliefs that are in different parts of your head that are doing different jobs. One does a public relations job. One does a behavioral guidance job. And that seems like it's a weird thing but I don't think that it's implausible because of the architecture of the mind.

Julia Galef: I can think of examples of, say, failed entrepreneurs, who -- they weren't doing that strategic switching back and forth, between self deception and truth seeking, depending on which was useful to them. They weren't being confident in meetings with potential investors, and truly believing in those moments that their startup was definitely going to succeed and had no problems, and then once they got home trying to be as accurate as possible with themselves about all the potential risks and pitfalls and flaws.

They were just confident the whole way. And in retrospect they think that they were overconfident, and were sort of in denial about the risks or problems.

That feels counter to your model.

Robert Kurzban: Right. That seems like a case where ... Yeah, it's hard though. You don't want to have hindsight bias. So for every entrepreneur who was super confident and then had a bad outcome, you want to ask the question, for how many of those entrepreneurs did that confidence actually lead to good outcomes? How many of those ...

Julia Galef: I think it's a minority though.

Robert Kurzban: Yeah, but if the stake is a unicorn -- if it's if "I win I get a billion dollars, if I lose, I lose a thousand dollars of someone else's money," then ... I think people's cognition is really sophisticated. Are people perfect at mathematics? No. But are they pretty good at computing the probabilities? Yeah.

And I think in these contexts, one might make the argument that recent history has shown that a ridiculous amount of confidence and optimism has led someone to an incredibly good outcome that they somehow shouldn't have gotten, given the actual skills that they have in the job for which they were applying… to the American people, let's just say. Not to get, uh…

Julia Galef: I just don't know what you're getting at, you're being so cryptic!

Robert Kurzban: I think that cases like that illustrate that, sure, confidence in some settings which is not justified can lead to disastrous outcomes, as in the airplane case.

But by the same token, I think William James was one of the first people that talked about this, is that humans are deeply social creatures and we are subject to self fulfilling prophecies. The example that he gives is in the context of mating, where he talks about a suitor who is just so sure that the woman he is pursuing is going to fall in love with him that he brings about that outcome. And part of it is because there is something compelling about that kind of certainty. There's something kind of weirdly compelling about the person who leads you to believe, "I know for sure we are meant to be together forever. You're my lobster."

Julia Galef: Don't use that quote on your Tinder profile, please. That's my free advice to you.

Robert Kurzban: Appreciate it.

Julia Galef: We only have a few minutes left, so I want to make sure: just to be clear, I think we do disagree. In that I'm less bullish on the benefits of self deception on net than you are.

Although I agree that in some contexts it can be helpful. It's just not clear to me that... A, it's not clear to me that that isn't outweighed by the costs of self

deception, and B, separately, it's not clear to me that we couldn't intervene on the system and make some improvements to get a better outcome. That we couldn't do better than evolution did.

One reason that I think we should expect that the solutions evolution found would not be optimal now is because of the many differences in our decision making environment now, compared to the evolutionary environment. Just to throw some examples off the top of my head, it seems quite plausible to me that there are many more skills that we can intentionally improve on now than there used to be. Skills that, if you really put in the time and effort, you can get better at it.

And if that didn't used to be the case in the evolutionary, the adaptive environment, then maybe it really was just better to try to convince everyone that you were much stronger than you actually were, or much more fierce in battle, or much more reliable or something like that. Because you can't improve on it, so you might as well just believe you're already great and get the benefits of social persuasion.

Whereas now, I guess, the situation has changed. The number of skills available to us to train is much bigger, and they're more complex. But also we maybe have the mental capacity now to decide to do deliberate practice on public speaking or on math or something. Again and again, until we improve. And we won't do that if we already believe we're perfect.

That might be one difference. Another difference is just in our social environment -- that looking good to the people around us was really important back when we were a small tribe, and social disapproval or scorn could mean getting cast out of the tribe. But now if you go to a bar and get rejected by 10 people, it doesn't really matter that much. But it *feels* like it does because our brains seem to have been optimized for really, really being risk averse when it comes to rejection.

For those and other reasons it seems to me that the intuitive calculus that our brain is doing, when deciding when to self deceive, might just not be optimal.

Robert Kurzban: Those are really interesting examples. I think in particular that last one is interesting. In the modern world it is bizarre how infrequently that happens, given the cost of rejection is so low. In places like New York City, men should be having relatively short interactions to try to judge interest, and then moving on after they get rejection.

I agree with you. The problem there is that you've got this evolved psychology where if you were rejected by someone in a world of 50 mates, that was a big deal. You just lost a big fraction. In the modern world it means nothing of course. Being swiped left doesn't really matter.

Julia Galef: Just briefly to clarify, I understand these are just evolutionary stories. But given that the model itself is based on an evolutionary story, I'm saying -- accepting that premise, I think we should then also expect that it would be an imperfect solution.

Robert Kurzban: Yeah.

Julia Galef: OK, go on.

Robert Kurzban: The example that I thought you were going to go to, which I would have agreed with, is gambling. There's a case where other people are structuring the world, I think, to take advantage of our overoptimism. Gambling in the modern world-

Julia Galef: It's an adversarial environment now.

Robert Kurzban: This is a good example it seems to me, because there -- overoptimism, that is a bad idea, and in fact it costs people who really can't afford to lose money tons of it everyday. And that's because if you do have a system in your head that is just a little bit, on the margin, too optimistic, and you're faced with an adversarial world where people want to take advantage of it, boom. All of a sudden you've got Las Vegas and Monaco and Atlantic City where they make all of their margins on the fact that everyone is wrong about how good an idea it is. Holding aside the issue of the value of the entertainment, or what have you.

But there is a case where I agree with you. The modern world has presented the ancient psychology with this tremendously difficult problem where specifically the errors that it's making is leading to catastrophic -- well, negative outcomes, let's say. So that part I would agree with you.

Julia Galef: Sometimes catastrophic.

Robert Kurzban: Yeah. I agree. Don't get me wrong. I 100% agree and have taken the position that the fact that the modern world is so different from our ancestral world, it leads to some bad decision making. That I agree with.

What I'm reacting to in some parts with this work is this idea that I think many people -- again coming back to the tradition of Kahneman, Tversky, and others, have taken this view that there's all these flaws and there's all these biases. And the work that I've been proposing says, look, some of these things might not actually be as bad as you thought they were. That there can be advantage to error.

And so in some sense it's less staking out the extreme position, as opposed to tacking back towards the middle.

Julia Galef: Yeah, and this is actually a running theme in my podcast -- looking for ways that apparent irrationality might be rational in some contexts, or under certain conditions.

I had Dan Sperber on, talking about the argumentative theory of reason, that rationalizing and making bad arguments might actually be useful because it's designed to persuade other people to agree with you, not designed to figure out the truth about the world. And I've done a couple episodes with Tom Griffiths who argues that apparent cognitive biases might actually be really good solutions under bounded conditions, where you have limited time and resources and you have to avoid serious downside risks.

I think this is an additional very important pillar in this set of reexamining apparent irrationality.

Robert Kurzban: Yeah, and in that notion we were talking a little bit before the show about sources that are relevant. So along the same lines, Gerd Gigerenzer and his group, and the Max Planck Institute, have a whole body of work that flies under the flag of "heuristics that make us smart."

And the idea there is that yeah, these little shortcuts, these little heuristics that we have are in the service of getting to good outcomes. And that's a whole nother research community that has more or less come to the same kind of conclusions -- that yeah sure, there's going to be shortcuts. There's going to be imperfections in the system. But you can tell by the way they frame it that they're talking about how this gives us an advantage.

And so for listeners who are interested in a body of work that pushes that, I think, in a really productive direction, looking at Gerd Gigerenzer's group and the books that they've produced I think would be a really rewarding experience.

Julia Galef: I was just about to wrap up and ask you if you wanted to give a pick for the episode, which is a book or blog or a journal article that has influenced your thinking. Is that your pick, or do you have another pick?

Robert Kurzban: I have another one. I think that in the modern era people are talking so much about artificial intelligence that people have forgotten the good old fashioned artificial intelligence guys, and so part of what influenced me was Marvin Minsky's book, Society of Mind, which dates all the way back to the '80s. A smattering of applause. I think for people who don't just want to look at the cutting edge but want to take a look at some of the background work from someone who was trained in computer science which I think still has relevance today, I think that would be a very rewarding read.

Julia Galef: Excellent. Well, Rob, it's been such a pleasure getting to chat with you finally and thanks for coming on the show and to NECSS.

Robert Kurzban: Thanks for having me.

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