Rationally Speaking #196: Eric Schwitzgebel on “Weird ideas and opaque minds”

Julia Galef: Welcome to Rationally Speaking, the podcast where we explore the borderlands between reason and nonsense. I'm your host Julia Galef, and my guest today is Eric Schwitzgebel. Eric is a philosopher at the University of California, Riverside. He's the author of several books and also writes the excellent blog The Splintered Mind. Eric, welcome back to the show.

Eric S.: Thanks for having me again, Julia.

Julia Galef: Yeah. Eric was on the show about a year ago now, talking about crazyism. You're, I hope, still as crazy as ever?

Eric S.: Yeah, maybe even getting crazier.

Julia Galef: I mean, I've been reading your blog so the answer, I know, is yes.

I wanted to have you back on the show because I have this cluster of related burning questions for you that have accumulated in the last year of reading your stuff, that are loosely related on themes of: the right way to think about weird or counterintuitive ideas, and whether we can know our own minds, and some things like that. So we have a lot to talk about.

Why don't we work backwards by starting with a blog post you wrote just recently that really caught my attention, it was called “Truth, Dare, and Wonder.” In it, you were describing these three different styles of thinking or philosophizing. I guess your post was about the context of academic philosophy but it really could apply to anyone who's thinking and discussing big picture ideas. Before I ask my questions, why don't you just explain what truth, dare and wonder represent?

Eric S.: Right. That truth or dare idea I got from another blog, which I hope you noticed, a fairly new blog called *View from the Owl's Roost.* The bloggers there put out the idea that some philosophers are “truth” philosophers and some are “dare” philosophers. The idea is that dare philosophers like to go after positions that might be extreme or exciting or interesting, but they don't really believe them. But they argue for them and they dare you to show how they're wrong.

Julia Galef: What's an example of a view a dare philosopher would espouse?

Eric S.: Right. For example, you might think -- panpsychism is this view that all of the matter in the universe is conscious. Now, there are probably some people who genuinely believe that, but you could imagine someone taking that for dare-like reasons. Like, “Here's an argument for that, I know this conclusion seems absurd but here's the argument and I dare you to prove me wrong.”

You might put that forward without really believing it in your heart -- but that still could be an interesting way of doing philosophy. You take this position that people find highly unintuitive or something like that, and then you argue for it and you defend it. So that will be dare style philosophy.

Then truth-style philosophy, which is the one they're, maybe, more in favor of ...

Julia Galef: Yeah, I can tell from the naming scheme the leanings of the person who named them.

Eric S.: Right. You're aiming at the truth and you don't want to espouse positions that you don't sincerely believe.

Julia Galef: Are you optimizing for daringness under the constraint of things you think are true? Or only optimizing for truth and not caring at all about how interesting the claim is?

Eric S.: When I pushed one of the blog post authors on this point a little bit, she seemed to be suggesting that she was especially attracted to true positions that were surprising. But I think you could also just be a truth-like philosopher who sees your job as showing the boring thing to be true. Like, in fact, not all of the matter in the universe is conscious.

If you've got dare-like philosophers out there, then you probably want some people fighting back even though the response position isn't so daring.

Julia Galef: Got it. You're motivated to argue the true things especially if you think people are missing those points. Even if the true position on that issue is not interesting.

Eric S.: Right. A lot of what philosophers do is argue for boring true things.

Julia Galef: Got it. Right. I assume, ultimately, the goal is truth and then the question is just: in the present moment, what do people think on the margin we need more or less of in our collective search for truth?

Eric S.: Right. Well, you might think that…

Julia Galef: Actually, that's a longer thread that we should get to in a minute. Before we go there, why don't you talk about "wonder"?

Eric S.: Yeah. Part of my reaction to this was feeling like people might think of me as a dare philosopher, because I espouse, sometimes, pretty wild seeming positions, or argue for them.

But it didn't seem quite the right characterization of how I see what I do. I'm really drawn to the capacity of philosophy to call into doubt things that we normally take for granted.

Julia Galef: Things that might seem crazy to our common-sense.

Eric S.: Right. I mean, the way the dare style philosophy works is almost like a game, rather than something sincere.

But I feel this kind of sincere sense of wonder when philosophy does the job -- or when psychology or any other academic discipline does the job -- of pulling the rug out from under me, and causing me to question my background assumptions about myself and about the world.

Julia Galef: Is it wonder in the sense of the emotion of awe? Or is it wonder in the sense of, "I wonder whether this could be true?"

Eric S.: I think they're related in my mind. I mean, there are certainly things you can wonder whether they're true without feeling awe about it. But the kinds of things that philosophy gets at, where you wonder -- for me at least, there's this kind of awesomeness of being able to wonder about that.

One of the things that I've been arguing recently is that if we take standard materialist views of consciousness seriously, most of them would seem to have the implication that the United States, considered as a group entity, is literally phenomenally conscious. It's not that I *believe* it's true, but I think -- maybe, actually, it *might* be true. And wondering whether it might actually be true, then also produces in me this awe in a way, at how interesting and impenetrable the universe is.

That's part of what really excites me about philosophy. And that's very different than the dare game, that you might think something as like, "I'll play this game, I'm taking this extreme position and knock me down if you can." You know what I mean?

Julia Galef: Yeah, I do. I very much know what you mean. I can think of several people who I think fit that description.

Okay, my central question, reading this taxonomy -- and it's a great taxonomy -- my central question was why isn't wonder just strictly better than both truth and dare?

Here's the case. The case for why wonder is better than truth is that even if your only goal is truth, and you don't really care at all about playing a game or being provocative or getting attention or something like that, your only goal is truth… Still, to that end, you should be reaching for interesting and provocative new hypotheses, and wondering about them. Because some of those are going to turn out to be true in some form, assuming that we don't already know all of the true things already, which I don't think philosophers believe we do.

You need to wonder about things to get as complete and as accurate as possible a world view in the end.

Then the case for why wonder is better than dare is: Well, yes, it is good to advance interesting and provocative new hypotheses. But dare and wonder both do that. And the only difference that I can see is that dare is overstating their belief in those hypotheses, just for the sake of play. Or maybe for the sake of provoking more discussion than they otherwise would provoke if they just stated their true epistemic confidence level, which is like "This probably isn't true but I think it's worth considering."

It just seems to me wonder and dare are both doing this one good thing, but then dare has this additional bad thing, which is muddying the epistemic waters by being deceptive about what they actually believe.

Eric S.: Yeah, maybe. I guess partly I think about this at a group level and partly I think about our incapacity to be neutral in the hypothesis that we're attracted to.

Julia Galef: How so?

Eric S.: Well, think about it this way, if everyone was a wonder style philosopher, then they wouldn't be so interested in defending the boring, commonsense position that probably does need to be defended. They wouldn't invest as much of their heart in it.

We'll be attracted to certain philosophies more than certain others and certain types of positions more than certain others. I don't think if everyone were similar in their biases and attractions as philosophers or as members of the intellectual community, then you'd end up with a narrower range of things that people do and you wouldn't have that lovely competitive chaos of so many different styles of voice, that I think, ultimately, is what you want.

Julia Galef: The first part of what you said sounded like maybe a compelling case for why we need truth in addition to wonder. That there might be a bunch of points that need to be made, but they're not interesting or exciting enough to appeal to a wonder philosopher to speculate about, and so that's why we need the truth philosophers. That is somewhat compelling. But I didn't hear anything in there that justifies the existence of the dare philosophers.

Eric S.: Yeah. The dare philosopher might still go after some things that even the wonder philosopher would stay away from.

Julia Galef: Things that *should* be gone after, or things that *shouldn't* be gone after?

Eric S.: Maybe they should be gone after. Maybe you need someone out there who is really saying things that are hard to even take seriously as initial things you might wonder about.

Julia Galef: Is the idea that some fraction are going to be true? Or that there's some benefit we get to our thinking, even if none of those completely out-there

ideas are true, there's some benefit we get to our thinking from considering them or arguing with them anyway?

Eric S.: Actually, I think both of those. Some fraction may turn out to be true. But then, also, I think that it's a part of the value of philosophy and the intellectual enterprise of academia on Earth, that there are people out there who are embracing things and defending things for all kinds of reasons that might seem absurd and not even worth defending. As long as they're not the majority of people.

I don't know, I guess I'm inclined to want to celebrate the diversity of motives and the diversity of positions that philosophers embrace.

Julia Galef: Okay. Here's my hesitation. I like to distinguish between my inside view and my outside view – so, let's say it seems to me that panpsychism is true. When I look around the world and I think about things, I'm like, "Based on my understanding of consciousness and how that works, it seems like everything, actually, is conscious."

But also, I'm a reasonable person and I know that most smart, thoughtful people who have thought about this think panpsychism is wrong. So if you ask me to bet on ... "Okay, Julia, 50 years from now, when you've really thought about this hard, and you've talked to as many people arguing against panpsychism as you can and you've thought about their arguments and so on, what do you think you will end up believing?" I might say, "Yeah, probably my confidence in panpsychism will end up being reduced." And that's my *outside view*, that it's less likely to be true than other theories. But based on what I've currently considered, it sure *seems* true.

It's definitely valuable for people to share their inside views, because if everyone just relied on their outside view, we wouldn't have any diversity of opinion. Like, maybe there's one thing that's 60% likely to be true, but everyone just says that that's what they believe, because it's the dominant thing, and everything else is lower credence than that. And then we never find out if it was wrong.

It's good to share the inside views. But when we have dare philosophers claiming to have strong confidence in things that they don't really have strong confidence in, that messes up our attempt to see what the average consensus actually is, and what the outside view is.

The outside view matters too. It determines what I would bet on. it determines how much time and effort I'm going to spend trying to understand something. If a lot of smart, sincere people seem to *sincerely* believe it, I'm willing to invest more time thinking about it. It just feels bad for the epistemic health of a community.

Eric S.: Yeah, I don't know. I mean, I have some sympathy with that but, also, I remember ... Just a little anecdote. I started my PhD program in Berkeley in

1991 and I met someone who was interviewing me, who's, maybe, going to be my landlord. He said, "Oh, so you're a philosophy PhD student. Are you one of those people who thinks that you've discovered the truth about things, and you're in philosophy to tell everyone what's right? Or do you see it ..."

Julia Galef: …I feel like there's a correct answer to that.

Eric S.: "Or do you see it more as this chess game with moves and [strategy] that's fascinating and interesting, and who knows where the truth is but you're excited about the moves?" I thought, "That's a really interesting question," and at the time, I answered in the second way. Maybe partly because I knew that's the answer that he wanted.

Julia Galef: Hey, you need an apartment! The Bay Area is tough. I will take any philosophy position you want, if you have a studio for me for less than $2,000.

Eric S.: Right. I guess my thought is that there's a mix of vices and virtues in any of these different kinds of ways that you could enter philosophy.

Julia Galef: Do you think there's *any* bad way to be a philosopher?

Eric S.: Yeah, probably. But not truth, dare or wonder, not at that broad description.

We might want to tweak around the ratio of them. If philosophy was almost all dare philosophers, that would be a problem. The fact that there are philosophers out there who see it as a game of chess moves, and are fascinated by the chess movingness of it and that ... I don't know, I think that they have a different complement of virtues and vices epistemically than I do, as a more wonder-oriented philosopher, and than a more truth-oriented philosopher.

Julia Galef: Could you at least agree that dare philosophers should have to declare somewhere on their website or CV that they consider themselves a dare philosopher? So that we know to take their claims with a grain of salt, in forming our picture of what the field *actually* believes?

…I'm not going to get you on that one, am I?

Eric S.: I wouldn't want to enforce that.

Julia Galef: Fine. Do you have an opinion about, on the margin, what we need more or less of? Like, we probably could change the ratio by changing what we reward or punish – socially, in the sense of what kinds of claims or papers do we give attention and admiration to? And then, also, just what papers tend to get published in the top journals, that kind of thing. If we could use those levers to change the ratios, in what direction would you change them?

Eric S.: Well, since I'm sympathetic with wonder, I guess I'm inclined to think we need more wonder philosophers. And that might be a bias on my part. I think that there are challenges that wonder-oriented philosophers have in publishing, because it's a little hard to publish something that says, "Well, we should have a 5% credence that this bizarre-sounding thing is true." It's a little easier to say, "Here's the compelling argument this is true," right?

You can do that as a truth philosopher. And as a dare philosopher, you can do that in this kind of insincere way. But if you're being sincere and you're invested in, "Well, why shouldn't we explore this bizarre seeming possibility where it maybe only deserve a minority credence?" There isn't a lot of room for that in the discipline, as it stands.

And I can understand why journals aren't super excited about those kinds of papers, but I feel like it'd be nice to have more room for that. One of the fun things about being a blogger is that you can just go out on a limb a little bit like that more.

Julia Galef: Yeah. I was thinking that, actually, as I read your post, that blogs seem the most natural home for the wonder-style philosophy.

Eric S.: Yeah. I think that might be partly why I'm attracted to blogging.

Julia Galef: That all makes sense now.

One sort of related question I wanted to ask you -- related in the sense of how should we interpret or react to bold and weird claims -- you had another post a little earlier that was really interesting, in which you argued against reading weird philosophy charitably. And there's a lot hanging on the meaning of what does it mean to read something "charitably," but -- can you summarize that case?

Eric S.: Yeah. When we read, say, older philosophy or philosophy from a different cultural tradition, I think there's this tendency to want to read, especially if we like the philosopher, as saying things that are true. Which, of course, means true by our lights, or…

Julia Galef: Or at least reasonable, right?

Eric S.: Or at least reasonable, or plausible, by our lights.

You might think of one thing that would come from that is that if there are, say, four different ballpark plausible interpretations of an author from a different tradition, or much earlier in our Western tradition, then you might have a tendency to want to choose the one that's closest to our current contemporary view.

Julia Galef: Right. Just feeling that what you're doing is trying to interpret them reasonably, to assume they're being reasonable. Of course, reasonable has an implicit definition based on your framework.

Eric S.: Right. If you're reading Descartes or Kant or Zhuangzi or whoever, you like them and you want to say, "Oh, they're saying true, reasonable things," and so you interpret them with the principle of charity, the idea that you attribute to people that you're interpreting mostly true or at least plausible attitudes… and you say, "Well, he didn't really mean this strange seeming thing."

Julia Galef: Right. It would be like reading your claim about -- conditional on materialism, the United States being conscious. And reading that and going, "Well, I know Eric is a very reasonable guy and I like him, so he probably meant it *behaves* as if it is conscious. He didn't mean it's literally conscious," like that.

Eric S.: Right, exactly.

Julia Galef: Because that would be crazy.

Eric S.: Right, exactly. We take these things that we visibly think of as being crazy views and we say, "Well, the philosopher couldn't really have meant that so let's interpret them more charitably, more reasonably," and preventing us from seeing how different the philosopher's view might really be.

We kind of tame the philosopher, translate into modern terms and then we actually lose an important part of the value, I think, of reading cross-culturally and reading the history of philosophy, as you get exposed to views that are radically different from your own.

You get a sense of how, things that you might perceive as crazy, people actually thought were maybe, literally true, in other cultures or other times. If you're overly charitable, then you lose one of the important values you can get from reading broadly in philosophy.

Julia Galef: It sounds, now at least, that you're arguing against reading charitably just because it messes up your picture of the history and sociology of views. But I thought the case was like: you will actually end up with a less truthful, less accurate model of the world, if you read people charitably.

Because sometimes people will say things that seem so crazy to you that you assume that *can't* possibly be literally what they mean -- but, in fact, that is what they mean, and it is actually true in some important way. And you would miss that if you read it "charitably".

Eric S.: Yes, I do think that. I mean, there are several things that you lose when you read too charitably. One of those is the opportunity to interpret them correctly by over-assimilating them to our current views.

Another is the opportunity to, perhaps, discover a truth that was available to someone in a different time and culture that isn't available to you or seems crazy to you now in your current time and culture.

Still a third thing is the opportunity to, even if it's not true and even if it's not the author's actual view, to stretch your mind by contemplating something that's really bizarre-seeming and out there.

Julia Galef: Yeah. One of the reasons this post struck me is that I'm frequently an advocate of this practice called steel-manning, which is a play on the idea of straw manning. Where in straw manning, you're caricaturing someone's view in a way that's dumb and easy to knock down, and then you knock it down. And steel manning is when you try to construct a stronger and more reasonable version of what they're ... sorry, “more reasonable,” I'm going too far there, but just a stronger version of what they actually literally said. And then you consider that.

Eric S.: Very charitable.

Julia Galef: Yeah. A classic example of steel manning might be if someone says “Men are like X and women are like Y,” you could just respond to what they literally said and be like, "Oh, well, it's not true that all men are like X, because here's one counterexample of a man who's not like X." And then just be like, "Well, I refuted them."

Or, you could think, “Well, they're a reasonable person, maybe they meant that men in general are more, on average, like X and women are more Y.” It's not so easy to refute that with a single counterexample. And you have to actually think about, well, how would I tell whether the averages are different? How would the world look differently? And that's a more interesting thing to do than just find one easy counterexample and then decide you're done with it.

The thing that steel manning, I think, is contributing is -- one of the things, is: Well, A, it prevents you from accidentally straw manning them. So it's like a corrective for our tendency to sometimes straw man people unconsciously.

And then, B, often people are not perfect arguers, and they might say things that aren't quite what they mean, or neglect to mention a premise that's important to their argument, or something like that. And if your goal is to get at the truth instead of just to win the argument and refute them, then you want to try to do some of that work for them, if they haven't done it fully themselves.

I guess I'm wondering: If you agree that's a good goal, do you think there's a way to get the goods of what steel manning is supposed to do, is trying to contribute? Without the potential harms that you talked about in your warning about charitable reading?

Eric S.: Yeah. I mean, I like the idea of steel manning and that is probably an excellent thing to do sometimes. Just running with the metaphor, what if you befriend instead of attack the straw man? Right?

Julia Galef: I really like your tendency of taking a dichotomy and making it a trichotomy! It's like a theme for you in your work.

Eric S.: Yeah. Well, usually things aren't as simple as one and two.

Julia Galef: But three is exactly the right number, things are always as simple as three.

Okay. What does it look like to befriend a straw man?

Eric S.: Right. I don't know if this is such a great idea with the example you started with, but just to run with that… The person said something that you might, on the face of it, interpret as saying “All men are like X.” The steel man view is to say, "Well, they just mean a lot of men are like X or men on average are more like X."

But another thing you could do, and this is what I will be thinking of as maybe befriending the straw man, is think: “Well, is there any way that I could think more plausibly about the possibility that all men are X?”

I guess just one example, I don't know why this comes to mind, when I was an undergraduate, one of my friends was taking a women's studies class and he was very upset because the women's studies professor said, "All men are attracted to rape." He's like, "I really don't think that."

Julia Galef: Ugh… Sorry, go on.

Eric S.: Right. Now, he could have steel manned what she said. But one possibility is that she really meant that literally, and that what she wanted him to do was, really, much more seriously consider the possibility that, literally, all men are attracted to rape. That would be more like dancing with the straw man, right?

Julia Galef: I mean, this takes us back to the dare philosophy thing. Where it just feels, in some way, like defecting on a sort of implicit epistemic contract that we have with each other. Like, yes, maybe he's going to give more consideration to her claim because she framed it in this bold, provocative way -- but why couldn't she just have said, "On the margin, more men should consider whether they are attracted to rape. And not all of them are but more of them are than they think," or something. If that's actually what she meant.

Eric S.: Yeah. It's not firsthand so I don't know what she really meant. But I could imagine someone -- and, this is not an area I'm expert in -- who want to play or entertain bold, crazy seeming, extreme seeming views. And maybe you're attracted to certain essentialist views and certain psychodynamic views. Maybe from the Freudian tradition, you could see how an essentialist,

Freudian feminist might literally think, “There are some unconscious things that all men share that we will not see unless we take that claim at face value.”

Now, I'm not inclined to think that's true, but I think it could be interesting to really consider whether it *might* be true. Instead of ...

Julia Galef: That’s very “wonder.”

Eric S.: Right, but allow your little bit of wonder credence on that, instead of instantly steel manning it or straw manningly attacking it.

Julia Galef: Okay. That was reasonable.

One more tangential but interesting point about the crazy claims topic: So you're writing or preparing to write a book called *How to Be a Crazy Philosopher.*

Eric S.: Yeah. I've been rethinking the title on that.

Julia Galef: Okay, so that's what I wanted to ask about, in fact. You just reposted on your blog that you had gotten some pushback on using the word “crazy” in the title on the grounds that it was ableist, i.e. it was stigmatizing or it was offensive to people with mental illness.

I've seen and been part of a bunch of discussions of this general shape. That a certain word or practice that isn't generally considered to be offensive in normal American society is, in fact, offensive and that we should stop using it. Sometimes it's the word crazy, sometimes it's the word stupid, sometimes it's a practice like wearing a sombrero for a party or Halloween. Or an American making tacos if they don't have Mexican heritage, or something like that.

I often find these discussions pretty frustrating. But I also have a policy of at least trying to consider these arguments, because I suspect that the current set of things that I think are offensive is probably not the complete set that I would find offensive, if I genuinely thought about all the arguments and made my best judgment. I don't think my views are currently complete and correct.

So I do try to consider them. But I often still, upon reflection, think, "No, that's just not a reasonable case for why this word is offensive, I just don't agree after having thought about it." And that seems to be your take -- at least when I last read your blog, maybe your thinking has evolved -- that you just weren't quite persuaded that this usage of the word crazy was offensive.

I'm wondering what you think our policy should be. Should our policy be: genuinely think about it and if you don't agree with them, just say, "I'm sorry, I respectfully disagree, I'm going to keep using this word?" Or should

our policy be, “Well, I still don't see why it's offensive but if you say it is, I'll stop using it?” Or should it be, like, “I'll only stop using it if you seem to be objecting in good faith, as opposed to just, I don't know, a troll or something?” What do you think?

Eric S.: I think you don't want to be wholly deferential, because some people probably go too far in saying that things are offensive. It's probably okay to make tacos, for example, even if you don't have Mexican heritage.

At the margins, to use the phrase you're using earlier, one might want to shift a little bit toward deference. When I look at my own use of the word crazy, I don't feel like I'm using it in a way that should be offensive. But there are several people who seemed to think that it's offensive in that way and I guess I feel concerned enough that I might be wrong and that they might be right, that I've decided that maybe I shouldn't highlight my usage of it. I still will use it and can use it -- I'm on the hook for it with some of my earlier stuff. And I'm not quite ready to abandon it entirely but maybe I shouldn't put it in my book title. That's where I am right now.

Julia Galef: I see. It sounds like it still, ultimately, comes down to wanting to be guided by whether the thing *is actually* offensive. I know we're not quite defining offensive; that'll take too long in conversation. But, still, it's supposed to be guided by whether it's, in fact, offensive -- and not guided by doing the thing that a minority tells you to do. A minority, literally, in the sense of a minority opinion.

So you suggest deference because you think that there are going to be cases where your inside view says they're wrong about the offensiveness, but your outside view says there's a decent chance they're right.

Eric S.: Yeah. I think that's about right, yeah.

Julia Galef: I kind of like that, actually. Because I also had the sense that some kind of benefit of the doubt was correct, but I'm uncomfortable with the moral hazard caused by saying, "Well, even if I don't agree, I'll give the benefit of the doubt.” Because my intuition is that that encourages people to object about things, or to be offended by things, that they otherwise wouldn't have been offended by, if you know that that's a way to get people to change.

Eric S.: Right. Yeah, and I agree with that. You don't want to be too deferential partly because of that hazard you're talking about.

Julia Galef: You don't want to be blindly, or commit to being, deferential.

Eric S.: At the same time, a certain amount of outside deference maybe -- especially you're hearing it from several people, and people whose opinion you respect for other reasons.

Julia Galef: Yeah, I like that. Cool, okay. I don't have a good segue into this next topic, but it's interesting and important and we should talk about it anyway, that's my segue.

Some of your especially interesting blog posts, articles and books as well, have been about the challenge of self-knowledge. Of having introspective access to your own properties as a person, and even to what you're thinking or feeling or experiencing in your mind at that very moment. That accessing those things is much harder or less reliable than we tend to think it is.

You actually did this over 10 years ago, but I only just recently discovered it –something you've done that's really cool was this collaborative project with ... I think he was a psychologist, not a philosopher, named Russell Hurlburt. Who disagreed with you, and thinks that, no, in fact, we *can* have reliable access to what we're thinking and feeling.

You guys wrote this book together in which it was almost like an adversarial collaboration, where you're trying to figure out why ... Well, I won't explain your methodology, I'll ask you to do that. To start off, I'm just curious to hear the basic case for why you think introspection is not reliable and why other people disagree with you.

Eric S.: Right. Yeah, that book was a lot of fun, be happy to talk about it. It's a really interesting experience and exercise.

Julia Galef: So cool, I wish there were more books like that.

Eric S.: Yeah. I want to answer the main question that you asked at the end but, yeah, the book was just… the idea methodologically of getting together with someone who has a very different view from your own. And not just doing pro, con, rebuttal response, like a couple of conversational turns. But actually writing something collaboratively together, with hundreds of conversational turns, where you're editing the other person's words, they're editing your words and you're really trying to get at the truth of their opinion. That was just really interesting and extremely rare.

Julia Galef: I'm swooning here, so great.

Eric S.: Yeah. Russ was a wonderful collaborator for this, he's very non-defensive in certain ways… so that was awesome. We can talk more about that if you want. But, let's see, I also want to answer the question about why I think that people have poor self-knowledge.

I guess I partly got into this because there's this long philosophical tradition of thinking that people have perfect self-knowledge of their own stream of conscious experience as it's occurring within them.

The classic example of this, and this is often associated with Dick Hart and the tradition that comes after him, is -- if you're feeling intense pain and you

think “Am I in pain?” it seems impossible that you could be wrong about whether you're in pain or not, right?

Julia Galef: Right. You could be wrong about whether someone is stabbing you -- like maybe you're delusional, and you just think you're being stabbed, but if the pain is there in your mind anyway, then you're feeling pain. Whether or not you're right about the source of it.

Eric S.: Right. Or if you're a brain in a vat, you might be wrong and there's no external world out there at all, but at least you're right that you're having these visual experiences *as though* there's an external world. You can't be wrong about that, like “I'm having this visual experience of red right in the middle of my visual field, how could I possibly be wrong about that?” That kind of intuitive appeal.

I think it's often exactly the cases of canonical intense pain, and canonical red as experienced in the full view center of the field -- those are the two philosophers’ favorite examples, not accidentally I think -- that invite this idea that philosophers have found attractive, that people can't be wrong. They're infallible about their own stream of conscious experience as it's going on through them.

In the 20th century, psychologists had done a good job of bringing up doubts about our knowledge of our attitudes, especially our unconscious attitudes, like from Freud. And also the causes of our behavioral choices like you see with people like Nisbett and Wilson.

Psychologists had not really, I thought, nailed down the case that we could be radically wrong about our own stream of currently ongoing conscious experiences. I thought maybe we could.

And I was partly led to thinking that maybe we could because at the time I was starting to think about this, I was a graduate student in philosophy, but I was working in Alison Gopnik's laboratory in developmental psychology -- this was at Berkeley -- and John Flavell was down at Stanford, and I'd been an undergrad at Stanford. Gopnik and Flavell both thought that children about three or four years old could just make these whopping mistakes about their own experiences, their own attitudes, their own stream of conscious experience.

Especially Flavell. I thought Flavell made a good case, and we could talk about the case if you want. The summary version is I thought Flavell made a good sense of four-year-olds could be radically mistaken about their stream of experience, and it's a little hard to reconcile with the idea that maybe adults will be completely infallible.

Julia Galef: So he wasn't just making the case that they can be wrong because they're bad at communicating? Like they might say, "I'm angry," but they don't

really even understand what the word angry means and they're just feeling excited or something?

Eric S.: Right. It's not that in the Flavell. Just for one example from the Flavell -- and he does this in so many different ways, you can just tell that he's trying to help them find the right answers and they're just failing over and over again. But one example is in one experiment, he's got ... I think, they're four-year-olds and he says, "I'm going to ring a bell," and he's got a library bell under the desk and then he waits five seconds and he rings it. Then he says, "I'm going to ring the bell again," and he waits five seconds and then he rings it. He says, "I'm going to ring the bell again," and then he waits 10 seconds without ringing it. He says, "Are you thinking about anything?"

A lot of the children will say, "No."

"Are you thinking about a bell?"

They actually will say, "No, I'm not thinking about a bell." But it was like, of course, they got to be thinking about the bell!

The way he describes it -- if I'm recalling correctly, I hope this isn't just my imagination playing tricks on me, but the way he describes it or at least how I picture it, is the children ... they're shaking, waiting for this bell to be rung, because this is such a weird thing that this adult is doing. How could they not be thinking about this?

Julia Galef: Can we be confident that they're not just saying what they think he wants to hear? I mean, this is probably a challenge in trying to perceive people's internal experiences anyway, that we can only get at them through self-report, but it seems like it might be a bigger problem with children than adults.

Eric S.: Yeah, that's true, and I don't think that experiment is decisive by itself. And maybe the entire body of work of Flavell is not decisive. But to me, it's suggestive at least, and that's part of what inspired me to think about the adult case too.

Once I started thinking about the adult case, I felt like I did see evidence that we are often quite badly mistaken about our own stream of conscious experience, even as it's ongoing. The philosophers' tendency to focus on a full view presentation of color in the middle of a visual field and an intense canonical pain ...

Julia Galef: They're making it easy for themselves.

Eric S.: They're choosing the two easiest cases.

One of the things that I invite readers to do when I'm making the case for this is to form a visual image of their house, or their apartment, as viewed

from the street. And most people say that they can form these images. Not everybody does, but most people say they can.

And then think about how stable is it: Is it fully colored? Is it fully colored all the way into the periphery, before you think to assign color to it? Is it flat and two-dimensional like a picture would be, or does it have more depth to it? Is it like an image, an afterimage? Where is it located in space?

Actually, when I've interviewed people about their imagery, some people say, seemingly to their own surprise, "I have this visual image and it seems like the image was located in front of my forehead, I know this doesn't make any sense," because you think that image will be in your head if it's anywhere. Some people will say, "Yeah, I have a visual image, it's not like that image is anywhere." Some people will say, "Well, it seem like it's in my head." Then a substantial number who might worry will say, "I know this sounds weird but it seem to me the image was in front of my forehead."

Julia Galef: I get that, that makes sense to me intuitively.

Eric S.: All right. There are all these interesting questions about what it's like to experience imagery that, I think, are not obvious. And you could imagine people going pretty wrong about some pretty basic structural features of the imagery.

In the '70s, it was part of our culture in the US, sometimes they'd think that memories tended to be black and white like TV at the time. If you think about memory images, I think most people now will not be inclined to think that memory images tended to be black and white. But back when the most salient media were often black and white, people did maybe seem to think that their images and their dreams ... Go ahead.

Julia Galef: Did people think back then that the memories of people back before any media existed at all were black and white? Or is it just: memories were in color and then, suddenly, we got black and white TV and then memories went black and white?

Eric S.: This is actually a personal memory of my own and I've seen some evidence from it in popular culture, especially in Paul Simon's song "Kodachrome," is a nice example of this. I haven't found systematic discussions of memory being black and white in, say, the psychological literature or in journalism, so it's a little hard to evaluate carefully what people were thinking.

For the dream case, there was actually a literature that's very interesting where people in the '50 in the United States and the '40s thought that dreams just generally were black and white. I don't think that they thought it was just dreams in the United States, as influenced by media. I think they just thought dreams are a black and white kind of thing. Most people thought that in the 1950s.

It's related to the presence of media in the culture, so if you look pre-20th century, very few people will say that dreams are black and white. If you look 21st century, very few people will say that dreams are black and white. You look at the arc of it and it relates to the dominance of black and white film media in the culture.

And we got some cross-cultural evidence for this. This guy emailed me and said, "We should try this in China," because this was about the year 2000. He said, "Well, in rural China, most people are exposed to black and white media, their TVs are black and white, whereas in urban China, most people -- especially the wealthier people -- are exposed to mostly color media." So we asked about their dreams and we found rural people in China in the early 2000s tended to say that their dreams were black and white, and urban people tended to say their dreams were colored.

Julia Galef: That is weird. Do you think that any of this problem -- just generally of people misreporting their internal states -- could be chalked up to misremembering? Like, if you could somehow ask people right at the moment they're having the dream, they would report, "Yes, I'm dreaming in color"? But then if you ask them 15 minutes later or something, there's this revising process that happens and so they remember it having been black and white?

I mean, it's still people being bad at reporting what their recent internal states were. So in that sense, you would be right. But it wouldn't quite violate this almost self-evident notion that people have, that you can't be mistaken about *feeling* a thing or *experiencing* a thing.

Eric S.: I think that is possible with a dream case. Actually, once REM sleep was discovered, people decided to ask about coloration of dreams by using REM awakening, instead of using retrospective report or more distantly retrospective report. And the rates of color dream recording went way up.

But that was also, unfortunately, for testing my hypothesis, that was also during the '60s during which the film media were undergoing quick change in their coloration. So it's a little hard to know whether the change is due to the change in the REM awakening method or the cultural change. For imagery, you're recording it as it's ongoing.

Julia Galef: Or emotions, actually. I'm thinking of someone yelling, "I'm not angry!"

Eric S.: Yes, that's a classic example, and I've certainly experienced that. For example, I think my wife reads my face better than I introspect my own emotional state. If my wife thinks I'm angry, I usually am.

Julia Galef: It's better evidence. Outside view, I am angry. Inside view, I'm not.

Eric S.: Exactly. Even just the basic label of "Are you angry or not?" can be hard. But then when you start to think about… not just the label, but it's like some

people think an emotional experience is a state of bodily arousal, and exactly what type of bodily arousal will be associated with anger, and is it the same in every case of anger, and to what extent does it involve cognitive stuff versus more, literally, visceral stuff… I mean, what is it like to be angry? First layer, we don't even know very well whether we *are,* but second layer, even harder, is what is the phenomenology of anger?

Julia Galef: I'm also curious about the conscious versus unconscious question, with anger or with anything you're experiencing. Is the claim -- that people can be wrong about what they're experiencing at that moment -- does that boil down into, "People experience things unconsciously in addition to consciously, and so you're not aware of all the things you're experiencing, because some things are unconscious?" I mean, I'm probably walking into a trap here but that seems obviously true to me.

Eric S.: That's charitable.

Julia Galef: Because couldn't you just explain the "I'm not angry" person yelling that, by saying he's experiencing anger unconsciously? Or are you actually claiming he's *consciously* experiencing anger and he is wrong that he's not consciously experiencing anger?

Eric S.: I'm inclined to say the latter. I don't know what unconscious experience is. I would use the word experience and consciousness and stream of experience and phenomenology all synonymously to refer to "what it's like."

I do think one can probably, maybe, have unconscious emotional states; however, that's not the kind of case I'm thinking of. I'm thinking of: there is a phenomenological aspect of your experience that's emotional, that's going on with you right now, and it's not unconscious the way early visual processing might be unconscious. It's in there in your phenomenology, in your experience -- but you're wrong about it. Why couldn't that be the case?

It seems to be plausible that it is the case, when I think about harder cases like my emotional experience right now. I'm not in an intense emotional experience right now so it's not totally vivid to me what my experience is, it's not totally obvious to me what aspects of my phenomenology, what's going on with me right now -- viscerally, emotionally.

I mean, I can make some guesses but it seems like it could be wrong. The same way it seems plausible to me that I could be wrong about the features of my visual imagery as I'm thinking about my street as viewed from the house.

Suppose that our stream of experience changes quickly and is complex. And our linguistic and conceptual categories for thinking about it, the tools that we have to think about it, they're really designed for thinking about the outside world and not for introspection. So we struggle to get our minds

around this swift, shy, changeable, disjointed mass of experience that we have.

When I look at, say, this coffee cup in my hand, I could tell you all kinds of things about its features, its structure. And when I introspect my own emotional state around an imagery state, I can't tell you with nearly the same amount of certainty about its features.

This kind of Cartesian picture -- that we know first and most certainly our own experience, and what we know secondarily and much less certainly is the outside world -- in my view, that's exactly backwards. What I know best is the ordinary things about middle-sized objects around me. And this experience that I have, I don't know that nearly as well. And to the extent I do know that it's often because I reach inferences about it based on my knowledge of the social world and the physical world around me.

Julia Galef: Is a way to resolve this seeming paradox -- that people can be wrong about what they're consciously experiencing -- is a way to resolve that just the modular mind? That there's not one single unified consciousness?

Sort of like, as it's highlighted very starkly in cases of patients with damage to their corpus callosum. Where one hemisphere of their brain is aware of some information that was shown to one eye and the other hemisphere is aware of other information, and they aren't communicating. Could something like that be happening, where part of your brain is experiencing anger and this other part of your brain that's consciously reporting, is not even aware of it?

Eric S.: Maybe, but I'm almost inclined to go the opposite direction from that -- but in a way that comes around, maybe, a full circle to a similar conclusion. I call this the “crazy spaghetti” view of introspection.

Julia Galef: Of course you do.

Eric S.: Instead of thinking of the mind as modular, I'm going to think there's something to the idea of modularity, especially for early sensory input processes -- but instead we got this massive chaotic tangle of processes going on in our mind that interact and interfere with each other, and cooperate in this incredibly complex way. Instead of these tight-knit modules that then feed into a center or something like that, instead think of it as this chaotic tangle of crazy spaghetti.

Then somehow out of this chaotic tangle comes some kind of self-report. But what's driving that self-report is a whole mix of processes -- including your presuppositions about what must be the case, including just your ordinary linguistic habits, including all kinds of stuff, only some part of which is some kind of sensitivity to what the experience itself is that you're reporting.

Julia Galef: Yeah, that's actually pretty plausible. Or, that feels like a type of brain that could exist, and that would produce the kinds of phenomena that you've been pointing out.

Eric S.: Yeah, that's my view. I got a little picture of the crazy spaghetti model in one of my papers which is, basically, just a giant tangle. Here's my picture of the mind I scribble on a page…

Julia Galef: A brilliant artist, brilliant!

Last question about this before we wrap up, did your collaboration with Russ change either of your minds?

Eric S.: Yeah, that's something that's a little hard to know, consistently with my skepticism about self-knowledge. Here's one thing I think is true: I think that before I collaborate with Russ, I was on the cusp of going to a very extreme version of skepticism about self-knowledge. I still have what some people might see as an extreme level of doubt but I think Russ helped me pull back from going too far to the extreme on that.

I guess I do think that we can have self-knowledge through introspection and that there are ways to do it, including Hurlburt's preferred way, that probably do get us a certain amount of knowledge, I think. Maybe more than I would have thought before.

Julia Galef: Your position now would be: introspection is much less reliable than many people think, but it's also more reliable than I previously thought? And there are ways to do it…

Eric S.: Yeah, I think he did help me moderate. I did probably moderate my view somewhat as a result of the collaboration with him.

Julia Galef: Was that because he pointed out approaches to introspection that you hadn't been considering, that you agreed were more reliable? Or was it that he convinced you that the approaches you had thought were very unreliable were actually somewhat more reliable?

Eric S.: The way we did this book was we actually interviewed ... we gave this person a beeper. Melanie. We interviewed her about her experiences when this beep went off at random moments in her life, and then she wrote down little notes, and then we asked her what her experience was.

And Russ has this wonderful way of asking people about their experience, that's very open and non-judgmental, and full of openness to possibilities that you might think were crazy or strange or impossible at first. He kind of hears participants starting with their presuppositions about what experience “must” be, and answering in terms of those -- and then Russ pushes back against that and says, "Well, to me, it sounds like you're just assuming this about your experience. And maybe that's right and maybe it's

not, but let's beep you again. And here's some possibilities you might consider about what might be in your experience that are different from what you're pointing at now. And they might be right and they might be wrong, but just think about it."

He has this way of doing that, and then people come back the next day and the next day and the next day, and sometimes they shift. As you see them shifting a little bit away from those first reports, there's something convincing about that sometimes, I think.

I think that's part of what my exercise with Hurlburt did for me, and you can see that just playing out in the book, because we present most of the dialog word-for-word as we're interviewing this participant. Then we've got these little side boxes where we argue about what's going on, and connect it to existing literature.

Julia Galef: That's so perfect. Yeah. As you say, it's not just that you guys make arguments with each other, but -- you're not both, figuratively, facing each other, you're facing this real case and collaboratively trying to figure out how to interpret that, and reacting, both to each other at the real case. It's wonderful.

Okay, we'll link to that book as well as your blog… and, also, this is a good segue, to your pick for this Rationally Speaking episode. What is your pick -- the book, article, blog, something that's influenced your thinking?

Eric S.: Right. We didn't talk about moral self-knowledge, which is something I've been thinking about a lot recently, but this pick is related to moral self-knowledge: what do you know about your own moral character? Unsurprisingly, I'm kind of skeptical about people's knowledge of their moral character.

This is a book, it came out in the year 2000 and it had a huge impact on me and now I use it as the very first reading in this giant introductory class I teach called "Evil." Right now I'm teaching it. It's 375 students, which is about a typical size for this class.

This is a book, it's called *Without Sanctuary,* and it's a book of lynching photography from about 100 years ago. A lot of the lynches are racially motivated, though not all of them. People will take pictures of the victims of the lynching and they make postcards out of them and then circulate them among their friends. Often in these pictures, the crowd will be around standing proudly by the victim.

What James Allen and his co-authors did in this book was just find as many of these postcards as they could, and try to research the backstory of all the victims of the lynching.

It's just really striking because here you have someone who's been murdered, sometimes without even having been accused of anything serious. Like the person might be the mother of someone who took a potshot at a police officer, or someone who expressed approval of a black man having killed a white man who was abusing him. So the person who expressed approval of that could be murdered in a lynching. Then you get the pictures of the bystanders proudly in front of this murdered corpse.

Julia Galef: Like, they think they've done something righteous.

Eric S.: They think they've done something righteous. They're posing for this photograph and they're circulating it. They're bringing their kids and they're collecting souvenirs like the person's knuckles and shreds of cloth from their clothes and stuff. Then you read these horrible stories of how much torture was involved in these things sometimes.

For me, it's this amazing puzzle – like, what is going on? Because it seems so horrible, and so obviously horrible. But these people seem to have no moral self-knowledge of the gravity of what they've done. It's an emotionally moving, hard book.

Julia Galef: Yeah, that is not a beach read.

Eric S.: It presents this real challenge. Then I structure this class called Evil around it -- it's like, okay, there is our question, why are these people smiling.

Julia Galef: Well, you know the book and I don't, but it seems like it could be, potentially, not about self-knowledge, but about having a different framework of morality. And they are correct that they're being moral by this standard, that their society or their subculture adheres to.

Sorry, that's probably a long thread, but if you want to give a brief response…

Eric S.: That's more relativist than I would be inclined to say.

Julia Galef: I see, maybe it's a different kind of self-knowledge than knowledge of your internal states, so it's sort of different?

Eric S.: I guess I'm inclined to think that there are moral truths, and that one of them is that you should not torture and murder someone because they expressed approval of a black man having shot his employer and the employer abusing him. I just think that's a moral truth, right?

Julia Galef: Yeah, well -- even assuming moral realism, that there are moral truths, discovering those truths can be a ... that seems more of an issue of lack of knowledge about the *world* than it is knowledge about your *self* and your psychology and internal state, to me.

Eric S.: Yeah. It's also a lack of knowledge about your own moral position in the world, your own moral character.

… Okay, so we're not going to solve that right now! But that's my pick, anyway.

Julia Galef: Okay, excellent. Excellent and grim. We'll link to that and to your blog and the book that you wrote with Russ. Eric, thank you so much for coming back on the show, this is a pleasure.

Eric S.: Yeah, 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.