Rationally Speaking #155: Uri Simonsohn on, “Detecting fraud in social science”

Julia Galef: Welcome to Rationally Speaking, the podcast where we explore the borderlands between reason and nonsense. I'm your host Julia Galef, and with me is today's guest, Professor Uri Simonsohn. Uri is an associate professor at the Wharton School of Business at the University of Pennsylvania, and he blogs at Data Colada, which is a favorite blog among my data nerd friends and my social science friends.

Uri's research -- I like to think of it as breaking down into two categories. On one level, you research factors that influence human decision making, like for example the weather. I think of that as the object level. Then on the meta level, Uri researches the scientific method, and problems in the scientific method that lead studies to be less reliable than we would ideally like, especially in social science and, for example, psychology.

That's the part we're going to focus on in today's episode. For example, Uri has been called the Fraud Vigilante for his work uncovering and explaining instances of fraud in social science. We'll discuss that, but we'll also put it in the broader context of issues with the scientific method, and maybe touch on the current replication crisis, the crisis of faith that's affecting psychology and social science in general, that's been discussed in the news in recent weeks.

Uri, welcome to the show.

Prof. Simonsohn: Hi, thanks for inviting me.

Julia Galef: One thing that you said last time we spoke was that there's a connection between those two levels that I termed the object and meta level of your research, in which, on the object level, you're investigating biases in human judgment and decision making, or factors that can unconsciously influence our judgment. Then, on the meta level, you're applying some of those principals, some of those findings, to science itself. To the way that scientists think. Can you expand on that?

Prof. Simonsohn: Yes. I came to realize that after I had been doing work on methodology, that I saw the connection between those two areas that you described.

The connection is that statistics textbooks tend to think of researchers the way economists used to think of people fifty years ago: as purely rational maximizers, and objective processors of information. Economics went through a major revolution, first just incorporating limited processing of information and then becoming more and more psychologically realistic.

Statistics hasn't really gone through that. So once you start thinking about researchers as being motivated thinkers who want to be successful, want to keep their careers, but also just believe that they're right and that's why they're doing a

study… They have a very clear expectation in how that will tint how they interpret ambiguity that's present in any data analysis. You can really get an insight into how things go wrong and how to prevent it from going wrong.

Julia Galef: I think of the problems with the scientific method as falling into a few categories.

One of the biggest categories is something sometimes called p-hacking, which we've talked about before on this show, and it basically entails ways that researchers can end up, not necessarily intentionally, but getting results that fit their hypothesis more than they would have if that p-hacking weren't happening.

Then on a broader level, on an institutional level, there are ways that the process of by which studies get published can bias the field of published studies towards positive or exciting results -- and we've talked about this, it's the file drawer factor, the publication bias effect, on the show before. Both of those categories seem like they fall under the umbrella of unconscious biases affecting the way that we do statistics, and thereby what we conclude.

Then the third category is outright fraud, where researchers are literally just fabricating data. Is there a connection between fraud and the kinds of judgment and bias research that you were referring to earlier, or would you put that in its own category?

Prof. Simonsohn: Yeah, I would put it outside of it, at least ... I would put it outside of it. Most of the ways in which we relax how humans behave don't go all the way to criminal behavior or just unambiguously wrong behavior. I don't think, at least the training that I have or the research that I've read or created, you see much insight into the mind of somebody who would go their entire career creating false claims that she or him knows are false and going to conferences, presenting them, writing papers, editing, responding to reviewers, and all that for a big fiction. I think that's more psychopathic, so maybe a clinical psychologist would be better equipped to think about that.

Julia Galef: Right. Can you put these categories of problem, can you compare them to each other in terms of the magnitude of the different problems? Like how big of a problem is fraud relative to p-hacking or other unconscious influences on individual researcher's work, as compared to say publication bias and the way it affects how studies get published?

Prof. Simonsohn: I think there's two main ways you could measure them, how prevalent they are and how much impact they have.

I think it's hard to tell. Most people say fraud is not very common, but what they really are saying is they hope it's not very common, because they don't have any meaningful way of assessing it. It's very hard to do that. I suspect it's more common than most people describe it to be.

I would ballpark it… Basically I think most people would ballpark it at trivially small, and I would say a substantial small share, let's say ... And it's of course a number out of thin air, but just based on my experience I would ballpark it closer to 5% of studies.

Julia Galef: 5% of published studies are fraudulent?

Prof. Simonsohn: Yeah.

Julia Galef: Wow.

Prof. Simonsohn: Or contain fraudulent data. That'll be my rough estimate. Of course there's a lot of guesswork in there. If you ask most people, most people would say, "Oh, maybe a handful of researchers are doing it." That's what I think would be the majority answer.

Where I do agree with the majority, and especially influential people who talk about this, is I suspect very few ideas that are well-known are based on fraudulent data. I think most fraud tends to happen in the margins, in less important journals, by less important people, because that's where the incentives are. If you can't be successful any other way and you can fake data in a journal a few people, or any will ever read, then you can get away with it for a long period of time. In that sense it would be much more common, but less influential.

The only way in which fraud does have more of a burden on the rest of the scientific community is through two main ways, I think. One is through meta analysis. Meta analysis is this idea that when a literature is mature and you want to summarize it quantitatively, you go out and you try to find every study published and not published, and try to aggregate it. Then, because fake data is not constrained by reality, they will often have just gigantic, enormous effects. It can dramatically bias the overall literature. Even if you fake data and nobody ever reads your paper, you could still have an impact in the overall understanding of the literature because somebody is going to include your extreme data point into an overall average, and you're going to have disproportionate weight.

For a recent paper we had, we were trying to improve on a technique that we have to correct for p-hacking, for selective reporting of analysis, and we simulated what would happen if you throw in the studies by one of the people that I caught committing fraud. Just one study by that person would make, if I remember correctly, if you take 20 studies that are all false-positives, so none of them are real effects, it's just by chance they came about, and you throw in just one fake study, it makes it almost certain that you would conclude the effect's real, because it starts an outlandishly large effect. In that sense I think fraud is potentially more consequential.

I don't know if that answers the angle that you were interested in.

Julia Galef: It does. Although I'm a little alarmed to learn that meta analyses don't generally adjust for outliers. I would hope that they would.

Prof. Simonsohn: They do. Yeah, there's a lot of variation in how to deal with them. Outliers are a really tricky problem from the statistics standpoint because you never know when you have an outlier if that's an anomaly, or if that's really an extreme value.

There are ways to deal with it. One way, for example that I see in analysis is they sort by the effect size across studies. So you will very easily see if there's an outlier value, but then it's not all clear what to do with it. Is that the person who really understands the phenomenon, that's why they get a large effect?

You are right that it is possible to manage the problem in principle; I think in practice people will be reluctant to say, "We decided to exclude this result merely because it was large." It could easily be construed as you being biased against finding evidence for some phenomenon for example.

Julia Galef: Did you say there was another way to think about the size of the fraud problem?

Prof. Simonsohn: Yeah, the other way I think it harms is that... Even if you're somebody doing work that is not very [influential] at all, you are taking a position, a faculty position or a research position, that somebody else could take, and that they would do good work. If you think of people in the margins, and then you think of the big, the impact any one of them could have on their students and so on, that's where it becomes more influential.

For example I happened to, at a conference, run across a person who applied for the same job at the University of Michigan that the fraudster took. For that person the cost of fraud in science was very real, in that they didn't have a job at University of Michigan because somebody who was faking data had it. That's like the human angle.

Julia Galef: Right, so there are human victims and then there's the general integrity of science victim, which is maybe a more indirect but pretty substantial consequence.

Prof. Simonsohn: Right.

Julia Galef: Is this, just generally speaking, what motivated you to go after the fraud problem, or was there some more specific impetus?

Prof. Simonsohn: No, so for the program we were working on something else and we just stumbled on the paper that was so incredibly, the effects were so incredibly large that we were trying to figure out what was happening. Then, I think like most people who do any sort of science, interest and curiosity is primarily what drove it at the time and it just ... If it were faked how could you prove it? How could you provide evidence that it's fake?

Then at that point I shared the prior that this was very, very rare, so I thought just because if you come across something that's so rare and so terrible you should do something about it. If it happened today I'm not sure I would react the same way because now I no longer think it's such a rare thing.

Julia Galef: Do you think this is a common reaction on the part of other social scientists, that they see surprising or surprisingly large effects and they think to themselves on some level, "Ugh, that can't really be real," but you're the only one who really, you're one of the few people who really tried to find out if it was real?

Prof. Simonsohn: I don't know, there's been a few cases. I think some people perhaps are too quick to ... So after the paper in which we described the fraud and how it was discovered and so on, were published, I would receive a ton of emails from people. From researchers and even data from political campaigns, election data, financial statements and so on.

You can tell there's some people who are very quick to judge that something's fake because it strikes them as surprising. I don't know, I think I should say that I'm an outlier in terms of how common fraud is. I know two other people who have such negative views on this and almost everybody, like 99.9% of people I know, think it's much less of a severe problem than I do.

Julia Galef: One thing that I know you've talked about in the past is that it's troubling for social science that we can't just assume that a surprising result that we read in the literature is true, and therefore interesting. We sort of have to have this prior that a surprising result, one that contradicts our expectations, there's a large probability that it's, if not fraudulent than flawed research for other reasons, and it sort of limits our ability to make updates from the research.

Prof. Simonsohn: Right, my view on that has evolved a bit. Some people would, especially researchers who advocate for basic methods, they will say you really have to bring in your priors about a phenomenon before accepting it. I think the risk with that is that you end up being too skeptical of the most interesting work, and so you end up, in a way, creating an incentive to doing obvious and boring research.

I have a bit of a twist on that. In practice it may not be not so different, but I do think it is different, which is I think we should bring in the priors and our general understanding and skepticism towards developing the methodology, almost blind to the question or the hypothesis that's being used.

Let's say you tell me you ran an experiment about how preferences for political candidates shift. Then I should bring to the table how easy it is to shift political preference in general, how noisy those measures are and so on, and not put too much weight on how crazy I think it is that you tell me you're changing everything by showing an apple below awareness. My intuition on how big the impact of apples below awareness are in people is not a very scientific prior. It's a gut feeling. So then if I start judging your scientific evidence with my gut feelings, that doesn't

seem right.

I probably do have a lot of experience, especially if I'm a political scientist or I'm an experimentalist, on how easy it is to move the dependent variable under different circumstances. So if you tell me, for example, even though Likert scale is one to seven, how strongly do you agree with this or that, they sound kind of flaky. When you work with them, you know that moving a Likert scale like that more than, say one and a half or two points, is really, really hard. If I read a paper, and the paper says ...

Julia Galef: Just to clarify, by moving a Likert scale two points you mean trying an intervention that will cause people's average response on the one to seven scale to move an average of two points.

Prof. Simonsohn: Absolutely, right. It would be some people we showed this and they said four on average, and people we showed that and they say six on average. I said two points.

If I wrote a paper that shows an effect of three, say, my prior is you don't get that unless it's an incredibly obvious thing, like what's taller, a building or a house. Unless you're asking an incredibly obvious question, I do allow myself to be skeptical, but not because the manipulation necessarily doesn't resonate with my intuition.

I don't know that the distinction is clear, but when it's my prior about the specific intervention you're claiming, there I try not to trust my intuition. And the other one is, what do I know about the reliability of the measures, how easy it is to move the independent variable? There I do, because in the latter case it's based on data and the other one is just my gut feeling.

If you want to be surprised by science and change your mind, then it's interesting if somebody comes up and shows something that you weren't expecting and it's reliable and replicable.

Julia Galef: To push this logic to the extreme, if we take a study that purports to show evidence for psi phenomena, for people being able to predict the future, for example, surely my intuition about the prior implausibility of that being real should factor in somehow, right? How could it not?

Prof. Simonsohn: Yes. I think that one, the example's more of a grey area, in terms of you could argue the hypothesis is so out there that your intuition about what moves things, it's not where the psychology starts or where some of the methodology starts -- in terms of is precognition really a psychological phenomenon, or does it challenge our understanding of how the world operates?

I agree. I agree that your example challenges my description, which is… a heuristic, right? The idea is when you tell me about, let's say preferences for how much you are willing to pay for shoes, I may have an understanding of that and I know how

much it moves one way or the other, and that's what's constraining. Then if it's about an ad or if it's about experience, that's a thing I don't have a lot of, beyond my loose intuition for.
A guess of precognition is about what changes dependent variables. From everything we understand, we know only things that happened in the past can change dependent variables, no matter what the dependent variable is. Imagine you thought precognition was possible, then whether it has to be arousing stimuli or only for men or for women, that part I wouldn't trust my intuition on very much.

Julia Galef: I see. A lot of the frustration with psychology and social science in general has been directed at these kind of frivolous or sexy studies that will get reported in science news about -- you know, I'm going to pull an example out of my lived experience in this very moment.

So, there's a window behind me and I can feel the sun on my back, so I could imagine a study of this kind showing that, "Oh, if there's a heater or the sun on someone's back, that will cause them to respond to surveys saying that their better days are behind them. Because they have this feeling of warmth and positivity being behind them."

I mean the framework that I'm thinking of here is basically that our behavior, and our choices, and our view of the world, can be strongly influenced by all sorts of random factors. Like, women are more likely to wear red on days 6 through 14 of their menstrual cycle, that kind of thing. I do feel like my prior on that not being the case is certainly a weaker prior than my prior on precognition not being real, but I do have a noticeable prior that that's not really how the world works.

Prof. Simonsohn: Let's go with the window behind you example. What part of our experience would give us feedback as to whether windows behind us do or do not influence our perception? It's hard. It sounds implausible, but that's just our gut reaction to it.

A lot of really interesting findings in science, social and not social, seem crazy when we are presented with them. I'm not challenging your instinct, I have the same instinct. They do sound far-fetched to me. What I try to do, it's almost like an exercise in self-control, and say, "Okay, I'm going to suspend that prior, and instead bring in this other prior, which is to ask how big is your sample?"

For example, those… cycle studies, if they had a really large sample, and they have a very carefully designed control group, and then you saw that they got the effect, and then the authors shared the very natural skepticism that any small cause would have a detectable effect, so they went out and they carried out a very similar replication that addressed a natural concern you may have… If they did that, and so they addressed my priors and methodology concerns, I would say there's good reasons to update my beliefs.

Because I don't have a strong, I don't have well-founded beliefs about how different phases of a cycle influence female preferences for clothing. Where would

that prior come from? It's just my gut reaction to it doesn't resonate with me, but it's not very well informed.

Julia Galef: I suppose that's true.

Prof. Simonsohn: You see what I'm saying?

Julia Galef: Yeah. In fact, now that I'm introspecting, I think that some of my skeptical prior about studies like that comes from the fact that I don't trust a lot of methods in social science.

Basically I have this model where, there are different reasons a study can become known to the public. Either it can be really well conducted, and therefore published in a good journal and discussed by scientists, and therefore more likely to filter into the public, or it can be a sort of fun, sexy result and that alone can be enough to propel it to public awareness.

So finding out that a study is sexy will make me somewhat less confident that it's true, I think. Sort of like if an actor can either achieve fame by being incredibly talented or by being incredibly attractive, and if all you know that an actor is famous and attractive, that might cause you to downgrade your expectation of them being incredibly talented. That kind of thing.

Prof. Simonsohn: I think that's reasonable. I think it's reasonable to say, if the only thing I'm going to tell you is I have a very sexy finding, I'm not going to tell you any other detail, it is perfectly rational to assume the methods are weak, based on that information. But then if you say, "Okay, I'm going to evaluate," now I'm going to open the paper and read it. At what point should I challenge the findings?

Julia Galef: That sort of screens off ...

Prof. Simonsohn: Right.

I was thinking this may be a good example of the contrast between the two. A few years ago I did research evaluating claims that our names really impact what we end up doing. For example, if your name is Smith, you're more likely to marry somebody else who's name is Smith, and if your name is Dennis you become a dentist, and so on.

My main reaction to it was, so they show that people tend to marry other similar last names. When I read that claim, I don't have a strong intuition about how we choose our partners and to what extent last names influence our perception of others. I don't have well-informed priors.

But I do have priors on how difficult it is to show an effect is causal, and so I thought, "I can't imagine a way to make this case compellingly." Even if it were true, how would you ever prove it... I was skeptical of the study for that reason, it

just seemed like a very strong, [hard] to document claim.

I ended up showing that all the evidence for it was spurious. At the time I wasn't necessarily a skeptic of the main hypothesis, I was just skeptical that you can easily document it. I think that makes sense. I think that captures my reaction to it.

Julia Galef: Right. Good. So let's dive into the question of how to react when data are too good to be true. This was one of my intuitions of how to detect either fraudulent data or data that are just a serious victim of p-hacking or other kinds of cherry picking. That if the results are much stronger than we would expect, even if that effect were real, then that's a strong red flag. Does that seem fair to you?

Prof. Simonsohn: Yes. Fraud definitely, at least the fraud that has been detected… has the flavor that is usually too good to be true, but the caveat is we only know the fraud we detect. And "we only catch very slow thieves" kind of thing. It could be that there's very smart fraud out there that's hard to detect and that's not too big to be true, and we don't detect it.

With p-hacking, with the selective reporting of an analysis, it's often the opposite where effects are just credible enough to, at least in terms of statistical significance right, they're just below the .05 threshold that we arbitrarily impose. Those are easier to spot where if you're trying multiple things just to get, if you exclude some variables or some observations or other to get your effect, or if you try one versus the other measure to get your effect, you will tend to land just south of .05, and so that will make it easier to spot.

Julia Galef: Can you tell the difference between data that's fraudulent and data that's maybe just the result of unconscious massaging on the part of the researchers? Is that detectable objectively?

Prof. Simonsohn: I mean the problem with fraud is that made up data doesn't follow any mathematical law. It's just whatever you come up with. That's what makes detecting fraud tricky in that there's infinite ways to fake data. There are, let's say, five standard approaches to identifying data as problematic, and those are all very different from selective reporting of analysis. You could very easily fake data so that it looks like it's been p-hacked instead of faked.

In fact, Dietrich Stapel, this famous psychologist from the Netherlands who got caught a few years ago, and I was not involved in that case, he would describe that when he would fake data he would try to not make it look too good. He would consciously do that. He wasn't very good at it because his data did definitely look too good, but at least he was trying to. He was trying to not be too good to be true.

Julia Galef: I know that I've heard about methods exploiting the fact that humans don't really know what randomness looks like, for example, so you can sometimes look at the second digit after the decimal point to see if the frequency of different digits in that position is what you would expect generated by random chance alone, that

kind of thing. Do you do anything like that? Or if not how do you actually prove that something was fraudulent, other than ... I mean it could fail to replicate but that's not proof of fraud.

Prof. Simonsohn: Right. In fact it could successfully replicate even though it was fraudulent.

Yeah, so in the work that I did, I did something that's similar to what Fisher did when analyzing Mendel's [gene studies]. I'm guessing a lot of your listeners will be familiar with that debate.

Julia Galef: Why don't you recap it anyway.

Prof. Simonsohn: Okay, hopefully I remember sufficient detail. Fisher, who's one of the founding fathers of statistics, frequestist statistics, p values and so on, noticed a pattern in Mendel's genetic studies that was troubling. And that was that Mendel's predictions were coming through *too* well in the sample.

You can say, even if Mendel were exactly right about the proportions of each trait that should be observed, the sample should have random error and they should deviate from those predictions, and they were systematically too similar to the predictions. To the point where Fisher computed, "Okay let's imagine if it was right, how likely you'd be to get this good evidence or better." It's like the opposite of the p value which asks, "How likely if your theory's wrong, sorry, that your theory's right, that you'd be this far from the theory." Fisher was asking how likely you'd be to be this close to the theory.

He concluded that Mendel's theory was impossibly similar to the predictions. There's been debate that, at least four, five years ago over it still, there was a paper in statistics where they were debating if it was right out fraud or if it was selective reporting in this case. Some of the analysis I did were of that flavor, where I thought that the samples in the studies just didn't have enough variation. Another way to detect fraudulence is data not fulfilling expectations of mathematical properties.

They can also deviate in terms of more conceptual or psychological properties. If you know a domain, you know that they behave in certain ways. For example, if you ask people how much they will need to pay for things, which people in my field do a lot as a way to capture how much people value things, how much they like them, or how the liking or interest changes within different abstracts. If you ask them how much they would pay for things they tend to answer in multiples of five or ten. If you were to fake data and you aren't aware of that, you may say, "Oh, people valued t-shirt, and they said they would pay $17, $18, $12," and not have that very marked tendency.

One of these cases I collected evaluation data from 20 different studies and all of them have very, very, very pronounced spikes at the multiples of five, and his data there were zero. Then I'd run a replication of his study, and not only did I not get

the effect he got, but I did get spikes at multiples of five. That was additional evidence that the data were not ... I think this is a euphemism that people use quite a bit, but the data were “not collected as described in the article.”

Julia Galef: That is, wow, that is very ... I was going to say that was the smoking gun, but that's also a way to describe it.

I have to say, listening to this is making me feel the way I feel when I watch a crime procedural and the detectives come up with all sorts of clever ways to finger the culprit even though he thinks he's covered his tracks. My reaction's always, "God, I'm never going to commit a crime because there's no way that I would know everything I had to do in order to get away with it. There's always something I'm going to forget or slip up." As you talk I'm thinking, "Man, I'm never going to make up data, because there's no way I'd get away with it."

Prof. Simonsohn: It's funny, so when I was working on this, well I haven't for a couple of years, but when I was, statisticians would say, "Oh, I know, they're so dumb. It would be easy to fake undetectably." They would go on to describe something ...

Julia Galef: Famous last words.

Prof. Simonsohn: I know. They would go on to describe something and I’d be like "Oh, that would be totally detectable." For example, they would say, "Oh just generate random data with, even in Excel, just use a normal distribution and generate random data." I would say, "Yeah, but if you used it for example, for violations, you would immediately get caught, because they don't follow the normal distribution rate..."

I do think it's possible to fake undetectably. I think it's very hard to do it at your first attempt. If you don't have feedback, if you don't have somebody saying, "Oh, this is how I would get you. This is how I would get you," I think it's actually way harder than it seems.

Julia Galef: Let's talk briefly about solutions. One solution category that has come up before is the idea of preregistering, so getting researchers to state ahead of time what effect they're looking for and what methods they're going to use to test for that effect.

The pessimism that I've heard from people, and that I kind of feel as well about that solution, is just that the incentives aren't really there. That as long as journals are going to keep publishing non-preregistered studies, what benefit is there to researchers tying their hands ahead of time, when they could otherwise leave themselves free to kind of data mine and cherry pick until they get something publishable. What do you think?

Prof. Simonsohn: Right. One big distinction is that preregistration is all about selective reporting of analysis, p-hacking, but it won't do anything with fakery, right? If you're faking you can preregister, you'll get it no matter what.

My view has evolved on preregistration. I used to be quite skeptical and now I've co-created a website for preregistration that's called AsPredicted.org. When we created that, part of the reason was we did see a selfish incentive for preregistering your studies, and it is that once the readers of your work have become more skeptical, and justifiably so, more educated about how selective reporting matters, they will start looking for science that you selectively reported. If you don't selectively report, if you are transparent, then you need a way to *signal* that you are transparent.

The preregistration becomes a little bit like the organic label in the organic farmer apple. If I went to create an apple, it's harder to produce, more expensive, and it's smaller, how do I get credit for it? Well, I label it.

Actually I was just completing a paper where we have a preregistered study, and we had collected three variables, three alternative measures. One of them we thought, I actually disagreed with my co-authors, I thought it wouldn't work and they thought it would.

If you're worried about people worrying that you p-hack, you may choose not to collect that variable to avoid any suspicion. Instead, we preregistered that we care really about two variables, and this third variable was exploratory and we were not going to include it in our analysis. Preregistration bought us the freedom to include things in our study that we were not planning on reporting in this study, but we wanted to use to inform future research.

I think it signals where you're not selective reporting. It allows you to collect additional information or just science that you wouldn't normally find.

We launched AsPredicted.org on December 1st, and it has over 100 authors who contribute their preregistration already which is, we're very excited about that.

Julia Galef: That's wonderful. Are you actively trying to increase its adoption in the field?

Prof. Simonsohn: It's growing fast enough that we haven't done any ... We basically had a blog post about it and tweeted about it. There's an aspect of the design that's kind of viral, that we didn't build it in but we got lucky about it, which is all authors have to approve any given preregistration. If I'm co-authoring with somebody and I preregister it, they all get emails and they find out about it. I think it's spreading through that, I don't know, the way that Hotmail, back in the day, spread out. You would get those emails people sent saying it came from Hotmail. I think it's having that, not by design but by chance.

Julia Galef: Or like playing Candy Crush on Facebook. Would you like to show your friends your wonderful success at Candy Crush, yes or no?

Except for good instead of evil.

Prof. Simonsohn: I actually expect once a few papers start coming out with that ... We worked really hard to make it incredibly simple to use and to enforce. There are other options for people to preregister their studies, but it's very costly as a reader to check if the things were done as predicted, as preregistered, because that can be like a 40 page document. We at AsPredicted have a single page document. The idea is that every reader, within a minute, should be able to compare the study that was published with the study that was preregistered.

Julia Galef: At this point I'd like to dive into the current replication crisis in psychology. Obviously, the problems of social science have been discussed for years now, but the recent context is that there was a paper published in Science a few months ago, titled, "Estimating the Reproducibility of Psychological Science," in which the authors took 100 papers from the psychology literature and tried to replicate them. They found that only 40% of those papers actually replicated. Meaning, that in only 40 of the cases were they able to find the same effect that the original study found. This has caused much gnashing of teeth and rending of garments in the social science world over the last few months.

The most recent update to the debate was a commentary published a few weeks ago by a couple of social scientists, saying, "You know, this actually isn't that bad. There are a bunch of reasons why studies can fail to replicate, even if the effect is real. So this 40% figure shouldn't actually be that troubling. There's no real replication crisis."

Uri, I was hoping you could speak to this particular issue and talk about whether in fact you think this 40% figure is troubling, or not. Then maybe, more broadly about the process of replicating in general. How concerning is it when you try to replicate a study and fail to find the same effects? What should we conclude from that?

Prof. Simonsohn: Starting with the specifics of the original paper and the critique… I was originally having the conversation with somebody and we came up a good analogy for it which is, the original paper said 40% of studies replicated. That would be, imagine I tell you that in soccer playoffs this team won 40% of games.

You'd be forgiven for assuming that, "Oh, they must have lost the other 60% because you can't tie in the playoff." But it turns out that in soccer playoffs, because there's two games, you can tie in a given game, so then if I tell you, "Oh, they won 40%, they tied 30%, and they lost 30%," you would be surprised by that. Even if there was no ill intent on my part, I just didn't realize you didn't know about the soccer rules.

I think that the original paper said 40% of studies replicated. They didn't say that 60% failed to replicate, but a lot of people interpret it that way, and that's not ...

Julia Galef: Sure, I did.

Prof. Simonsohn: Right, and that's not a good read, that's not a justified read of the evidence.

Justified read of the evidence is 40% replicated, 30% didn't, and 30% we really can't tell one way or the other. I think that's part of the discussion.

Some people have made a bayesian, have used a bayesian approach to think of the problem, I've used a different approach. It doesn't really matter, as long as you're willing to accept that some scientific studies are inconclusive as opposed to supporting or not supporting a conclusion, you will conclude because so many of the studies had small samples that their replications were just inconclusive. If we take the evidence at that, 40% of success, 30% failure, 30% unknown, that seems to me, that's probably better than I would have expected.

Julia Galef: But you are unusually pessimistic about the social sciences.

Prof. Simonsohn: Yeah. I don't know, I think people who are in this business of trying to improve how science gets reported, I think it's ... And part of the reason is it's really hard, and that's another part of the critique, is that it's hard to really replicate a study in one of these… social sciences where facts are so affected by context and also by measurement. You can have a study, and then you run it very similarly, but you run across a problem that was just not present in the original. For example, you can get a floor effect, which is used in psychology a lot and a little bit in economics where all the responses are so low that you just can't get any lower than that, and so you fail to replicate ...

Julia Galef: So you can't really detect any variance.

Prof. Simonsohn: Right. And because in social science, as well as hard science, harder sciences, the floor is going to be dependent on the sample. Maybe if you ask Americans today the floor is in one place, and if you ask Swedish people two years ago the floor is a completely different place. That doesn't really falsify the psychological hypothesis, it just means you have to adjust your measures, your sample size, or who your sample is. Even if all effects were true, just these factors…

This is the biggest debate, whenever somebody publishes a failure to replicate in psychology the original authors typically will say, "Well, it's because there's this big factor of change between my sample and your sample." That's easy to get into unfalsifiable explanations.

On the other hand, we don't know when that's true but we know that's true sometimes. 30% failure is not terrible because of that reason. I wouldn't say, that's not my estimate of how many of their hypotheses were wrong, it's maybe an upper-bound of that.

Julia Galef: Although I can imagine other factors that should make us more pessimistic about that 30% figure instead of more optimistic. Tell me if I'm wrong here, but it seems to me that there might be just a regression to the mean effect where if there's random variation in how strong the effect seems when you do a study and you end up publishing when the effect is unusually or abnormally strong, then the next time

that you look for that effect, chances are it's going to seem weaker because you published when it was strong.

Prof. Simonsohn: People do raise that response, which is to say it's true that things change in social science, but why do they always change for the worse? Two years ago in Sweden it was really hard to get the effect and today in the US it's very easy to get the effect. That's fair. That's fair.

But even if that were the case, out of that 30%, right -- I'm thinking as I speak -- when the original author gets unlucky because of publication bias, we don't make a record of it. We're not compensating hypotheses that we thought were false then we realized were true. To some extent it's still only working against you. In the sense that… among those failures [to replicate], some of them must be explained by when you got it weaker.

I'm not trying to underestimate how -- I think it's a serious problem, and I think the solution is disclosing how studies are run, preregistration, more replication, and so on. I'm completely on board with all of those things, but we don't have to convince people that only 40% of studies replicate to make that case. Even if 80% of them replicate, we want to know which ones are more likely to replicate, and the best way to do it is to report studies with the least subjectivity of reporting as possible.

Julia Galef: Another bias that occurs to me that might exist here is a kind of status quo bias, in which -- I think it was actually Andy Gelman who made this point that I might be stealing here, but he says, "Well, you know, if we have a study that gets published and it shows an effect and then we try to replicate it and the result is inconclusive, we still sort of assume, "Well it's probably true because we didn't disprove it."

But if we imagine the order of those studies being reversed, where the replication, so to speak, was actually the first study and it found no effect… And then we did a second study that found an effect… Wouldn't we be differently anchored? Our default assumption wouldn't be that this effect is there, right? We would just think, "Well, we have two studies."

Prof. Simonsohn: I like the point and I like the way of building the argument, and I think it's true. In particular I think if somebody has a study and it fails to replicate, I don't think it should be enough to say, "Oh, these three other things changed." I think unless it is just blatantly obvious that those things really mattered and of course it was incompetently performed, unless you can really have that strong argument, you should go out and test. As a person who wants to continue believing in the effect, you should go out and show that that moderator, that other variable that changed, really is important. I agree with that.

There's a way to turn the argument on its side and argue it the other way, which is, if the failure to replicate had come first, right, almost assuredly the author who ran it would have run a follow up study. When I run a study and it fails, I don't abandon the project, I look at the data and I see, did I measure things correctly, maybe I

should assign a stronger manipulation, a larger sample, and so on.

The replicator doesn't do that. When the replicator fails to obtain a result, that's where the project ends for the most part. If we're going to treat replication as if they had come first, we should look closer at things like do you replicate, for example, the range of values of the dependent variable, is it where it should be? Any sort of quality check that you'd apply to data. Is the quality check similar in replication as it is in your original. Forget original, is the quality check sufficiently high for us to put trust in it?

I think the argument, it's a good one, and 80% argues against original authors being so defensive about replications and dismissing them so quickly. But part of it also I guess is original authors try hard to get effects. Not in the bad sense of the word, in the sense of really understanding what's happening, and replicators are not interested in the subject typically. When they don't get it, they think it's case closed.

Julia Galef: I think what I want to use my last question for is a general takeaway for our listeners about how much to trust different levels of evidence in social science.

For a long time I've been skeptical of single studies. For the most part I'm not going to trust a single study in isolation without knowing the context and what other studies have investigated the same phenomenon. Unless it's just a really, really well done study.

Then for a lesser amount of time, I've been skeptical of meta analyses, which as you noted earlier in the episode can be influenced by things like outliers, which may or may not be fraudulent, and they have other problems as well.

Then more recently I would have said, "Well, maybe we can't trust individual studies, maybe we can't trust a meta analysis, but surely we can trust a consensus that's been around for over two decades. In which study after study after study and multiple meta analyses shows again and again this phenomenon exists."

One of the sub-fields, one of the consensuses in social sciences that has been prominent in this crisis of faith in psychology has been the idea of ego-depletion, which is that your will power can be sort of used up in a local, over a short period of time. So you want to conserve will power by not trying to stick to your diet when you also have to stick to a really hard task at work or something like that.

That *was* a consensus -- and now it's been cast into doubt by attempts to replicate it that have failed.

I'm wondering, Uri, if you have any general heuristics for like, should the takeaway from the problems with social science that we've been talking about be that you just have to retain a high level of skepticism about everything? Or are there some kinds of research or some levels of evidence that we can be pretty confident in?

Prof. Simonsohn: I think whenever, and it's a high bar I guess, which makes sense, whenever a skeptic replicates the effect, that's a good ... Whenever somebody has all the psychological reasons to *not* find it, when they find it. I think it's going to be a small share, but that's something that it's definitely the rational thing to do is to update in light of that evidence.

Julia Galef: I've heard of something called adversarial, is it adversarial research? I forget the name, but basically two labs ...

Prof. Simonsohn: Adversarial collaboration.

Julia Galef: Yes, thank you. Two labs, one which believes the effect is real and one which believes it isn't, collaborate and sort of agree ahead of time on a protocol, a set of research methods, and then report their results. That seems promising, is that common?

Prof. Simonsohn: It's rare. I've spoken to some people who have done it and they don't like it as much.

Julia Galef: I'm not surprised.

Prof. Simonsohn: I think part of the reason is that -- and this goes back to the point how replicators and original authors react differently to when their studies don't work -- it is hard to understand things.

When you're running a study and it doesn't come off as expected, you immediately see problems that you hadn't seen before. Part of it is self-deception and it's bad, but part of it is when you really need to figure something out, that's when you figure it out. If you talk to any creator who has a successful idea, if you tell them if they got it right right away they'll say almost invariably no, that they had a lot of failed attempts until they figured it out.

I think one problem with adversarial collaboration is that it assumes that after one study your beliefs will immediately update. And usually the moment that adversarial collaboration collects the data, whichever side didn't get exactly what they expected, they will see a problem they hadn't seen before. That doesn't need to be disingenuous or bad, it is just a natural process of updating our understanding.

I don't know. Social science doesn't usually deal with urgent matters so I think it's fine for us to say… I don't really understand why the newspapers have to cover studies the moment they come out. I think they wait four years until somebody, an opposing side, shows the effect too, I think nothing bad will happen.

Julia Galef: I think that's partly a coordination problem, right? Although I think there's also some weird psychological quirk where people are more interested in something that's new even if there's all this other news, all this other science that they haven't heard of before that isn't new, it's just been sitting around in textbooks for

decades. And it's not quite clear why they should be more excited about the thing that was just discovered than the thing that was discovered ages ago that they'd never heard of before.

Prof. Simonsohn: I mean, this is a falsifiable prediction, but I suspect that if you want to run the experiment there will be very little benefit of reporting on a recent study versus an old study.

Julia Galef: Benefit in terms of page views or something?

Prof. Simonsohn: Readership. I think if the New York Times, tomorrow, reported on a very sexy finding from five years ago I don't think it would have any fewer readers than if it's a new study. In fact, maybe more, because nobody else will be reporting on it.

Julia Galef: Interesting.

Prof. Simonsohn: It will require readers to be so sophisticated that they remember all these different findings and their subtleties how they're different, and they don't. In fact, because I was involved in the debunking of those name studies, I have a Google alert on that, and periodically, at least once a year, somebody writes a story in a major outlet about them and they're 14 years old. If you say people choose whom to marry based on their name, your first reaction is not, "Wait, is that a recent finding or is that an old finding?"

Julia Galef: That's true, that's interesting. I had just taken this as a given that people want to read new stuff, but maybe this is an assumption on the part of the media that isn't fully warranted. I'll have to think about that.

Prof. Simonsohn: I can't quite remember the details, but isn’t there a newspaper who ran the same editorial cartoon many days in a row to see if people would notice?

Julia Galef: No, I hadn't heard about that.

Prof. Simonsohn: That's a little different. That's about people actually reading it. I suspect… actually now I'm very curious and it would be worth testing it.

Julia Galef: Cool, well we are actually quite over time. I gave in to temptation to continue the conversation, but I'm going to force myself to wrap things up now. We'll move on to the Rationally Speaking pick.

[interlude]

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

Prof. Simonsohn: My pick is by Paul Meehl, who was a psychologist at the University of Minnesota,

and he gave his last PhD seminar in 1989, and somebody video taped it. They put all the video tapes, all the recordings online for downloading. It's this PhD seminar on the philosophy of science.

And what makes it really interesting is that he, Meehl, puts psychology within the bigger picture of science in a way that I don't think anybody's doing anymore. He would put all of our approach to understanding psychological phenomena from the perspective understanding scientific phenomena more generally. To find it your listeners can go to University of Minnesota and search for Paul Meehl, which is spelled M-E-E-H-L. I've also made a quick URL with the same files in MP3 format if they just want to listen to it, and the URL is tinyurl.com/simonsohnpick. My last name, pick.

Julia Galef: Excellent. Well, Uri, thank you so much for joining us. It was a fascinating discussion and I'll link to both Data Colada and also your pick on the Rationally Speaking website.

Prof. Simonsohn: Great, thanks a lot. I really appreciate it.

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