Rationally Speaking #209: Christopher Chabris on “Collective intelligence & the ethics of A/B tests”

Julia Galef: Welcome to Rationally Speaking, the podcast where we explore the borderlands between reason and nonsense. I'm your host, Julia Galef, and I'm here today with Christopher Chabris.

Chris is a cognitive psychologist and professor at Geisinger Health System in Pennsylvania. He writes about social science for publications like the New York Times and the Wall Street Journal, and he's the author of the book The Invisible Gorilla: How Our Intuitions Deceive Us. Chris, welcome to Rationally Speaking.

Chris Chabris: Thanks for having me. It's great to talk to you.

Julia Galef: I have a bunch of things lined up that I want to ask you about. Maybe let's start with some of your recent research on collective intelligence. Can you tell us how you define collective intelligence, and how do we know it's a thing and it matters?

Chris Chabris: Sure. So this work I'm going to talk about is all done in collaboration with Tom Malone from MIT, and Anita Woolley from Carnegie Mellon. I should say that right off the bat.

And the second thing I should say is that as a researcher I try not to get hung up on defining concepts, because I find that defining them precisely, even though that's generally a good idea, can often sort of get us hung up on whether we agree on the definition and distract us from the empirical phenomenon. So I'm going to define it a little bit by describing an empirical phenomenon, and that phenomenon comes from studies of individuals.

So we have the concept of intelligence, the psychological concept of intelligence, as a measurable thing about people, because when you give a bunch of people a bunch of different cognitive tasks, it just turns out empirically that for whatever reason, people who do well on one of the tasks also tend to do well on the other tasks.

They're not perfectly correlated, so it's not as though the person who gets the highest score on task one necessarily gets the highest score on all the other tasks and so on, but there's a general tendency for the performance on different kinds of cognitive tests to be positively correlated. We call the capacity, the inferred capacity that can lead people to do well on a variety of tasks, we call that intelligence.

So in our research, we've basically just tried to apply that simple concept of intelligence, the way it works with individuals -- some people being, colloquially speaking, smarter than others -- and just apply it to small groups or teams. Which is groups of two, three, five, six people working together to achieve common goals.

It turns out, as we hypothesized, some teams just seem to be generally smarter than others. They generally do better on different kinds of tasks. Teams that do well on one kind of task tend to do well on other kinds of tasks also. Spoiler, that's what we found in our research, but we can go into more detail now.

That's really the phenomenon of collective intelligence. It's some capacity that is reflected in the fact that some teams tend to do better than other teams on a wide variety of tasks that they might have to perform.

Julia Galef: What is the scope of tasks that we think collective intelligence predicts performance on? Because with IQ, it predicts performance on a lot of things, but certainly not all things, or it predicts performance differentially on different tasks. So what have you looked at, and what's your sense of what things it might not predict?

Chris Chabris: Yeah, so you're absolutely right about individual intelligence or IQ. It's more important for some things and less important for others, and it is kind of hard to find something that individuals do that has a measurably good and bad side to the continuum that is not related somehow to IQ.

It could be related very tenuously to IQ. So for example, being good at recognizing faces is pretty unrelated to IQ, at least as far as we can tell, and there have been several studies on this.

When it comes to collective intelligence, the picture is kind of similar actually. We found a couple of the tasks that really are the best measures of collective intelligence are solving abstract puzzles. That is, let's say three people sit around a table and they literally get one piece of paper with a matrix reasoning puzzle on it, say, which is a common kind of IQ test item. It's abstract, it's non-verbal. It's not even pictorial, it's just a bunch of lines and shapes. It turns out that groups that do well on that kind of task tend to do well on the others.

Another one is a task that measures speed and coordination of the group members. This is my favorite one, because it's kind of the funniest in some ways. We gave them printouts of a Wikipedia article, and they had to type as much of that into a shared Google doc as they could in a limited amount of time, without duplicating or leaving gaps or making typos and mistakes in it, and so on. So it's kind of a speed, accuracy and, especially, coordination task for team members, which doesn't really seem very intellectual on its surface. But just as the speed with which you can, let's say, respond to blinking lights is correlated with individual intelligence, this kind of speed and coordination task seems to be an indicator of collective intelligence.

The further you get away from those kinds of things, it seems like the lower the relationship is with a general collective intelligence factor. So moral reasoning might have less relationship than abstract, logical reasoning, let's say.

Julia Galef: How do you measure performance on moral reasoning?

Chris Chabris: So that's a difficult one also, because there's not necessarily a correct judgment. That's sort of more of a process measurement, like how many things groups take into account when they discuss their decisions, and so on. We don't know what's in the individual members’ minds, but we can keep track of what they mention and what they discuss, and so on.

Julia Galef: Got it.

Chris Chabris: Brainstorming is another one we use, which is also a common thing that groups do together, even though it's not necessarily the best way for groups to generate good ideas. But it's often done.

And again, there you have a little bit of a problem of measuring the outcome. So it could be that a group that generates only one idea still comes up with a great idea, but the usual ways of measuring the output are number of different ideas and things like that.

Julia Galef: Right. Now, for IQ ... I'm far from an expert on IQ, but my impression is that we think we have some understanding of some of the underlying mechanisms that would make someone good at this whole wide variety of things. For example, I think working memory is probably part of IQ.

Now, that statement may or may not be correct, but I'm wondering what the equivalent would be for collective intelligence. What is the theory for what is causing groups to perform well on this wide variety of tasks, or poorly?

Chris Chabris: So we have done some work on that from the very beginning, and I guess I should say that it's mainly correlational work, as is most of the work on what causes IQ, because it's hard to randomly assign people to conditions that actually make them smarter or less smart, to do randomized experiments. Although, the evidence so far suggests you're right, that things like working memory and processing speed, and even brain volume are all related to IQ.

Julia Galef: Oh, good.

Chris Chabris: Collective intelligence, I think the story is in many ways much more interesting, because since we're talking about a capacity of a group of people, it's always possible to try to decide who should be in the group as a way of increasing or decreasing the collective intelligence of the group. It's also possible to arrange different kinds of environments or systems for the group to use in interaction, which you can't really do with the different parts of your own brain, right? It's hard to tinker with those and engineer them. But since we have the parts of the group in essence right in front us, we can start to play with that kind of stuff.

Correlationally, we have found in our initial studies that important things for collective intelligence were how well the group took turns. So we recorded every interaction that all the groups in our studies had, and we quantified things like how evenly distributed the amount of speaking was. So if one person did most of the talking, that seemed to be bad for the group. If each person spoke about an equal number of times, that seemed to be better. Those groups tended to score higher on the test.

Julia Galef: Interesting.

Chris Chabris: Another one was having group members who score higher on tests of social intelligence, and we used a very common measure of the capacity called theory of mind, and the test we used is called the “reading the mind in the eyes” test. So it's kind of an advanced test of social perception, detecting complex mental states in people just by looking at their eyes. And having members of your team who score higher on that test seems to be associated with having a more intelligent team.

Again, it's hard to say if any of these things are causal, because while we randomly assign people to be on teams, we just measure these capacities and then do correlations after the fact.

The third thing in our initial studies that popped out was having more women on a team. So teams with more women tended to score higher, and we replicated these findings several times now. So even though they were first published in our initial studies, our group, including some studies that I was not part of and some that I was, have replicated those basic effects more than once.

Julia Galef: So, first off, are you talking about literally, there's a monotonic relationship between number of women, or percentage of women in the group, and performance? Or is it like you want at least half women? Like, if you had all women, would that be better than half women?

Chris Chabris: So, yeah. So it's interesting. Sometimes people interpret the results that we published as evidence in favor of diversity.

Julia Galef: That's probably because they're thinking of a baseline of mostly men, so they think more women equals more diverse — which it does, from that baseline.

Chris Chabris: Yeah, so if your baseline is zero women, then adding women would be great as far as our results say about collective intelligence. But we actually ... The statistical analysis found a significant linear relationship, meaning the more women, the higher the collective intelligence, but no quadratic relationship. So although it ... If you look at our graphs, as I've looked at them many times myself, it does look like there's a slight drop when you get to 100% women. It's not clear that that's statistically significant. Maybe with a lot more data, it would be, and it's a benefit to have a mix, but certainly it's not ... We don't have evidence that a 50-50 balance is best, let's say.

One thing we do have some evidence for, though, is that in our studies, although this is not necessarily universally correct, and we could get into the weeds about the characteristics of the tests we use and so on. But in our studies, we found that some of the effect of having more women on your team seemed to be due to women also being higher in social intelligence, at least according to our measures of social intelligence. So it's not necessarily purely women, per se. It could be that having more people who are more socially intelligent would also be a benefit, and that's just ... It goes along with a slight effect of the sex difference between men and women, and that social intelligence measure.

Julia Galef: How does the effectiveness, or the importance of social sensitivity trade off against things like individual members' skill at the tasks that the group is working on? Like, if you were putting a team together and you had to choose between people who are really high in sensitivity but just average at their skill at the task, versus people who are average at social sensitivity, versus 90th percentile skill at the task. Which is better?

Chris Chabris: That's a really good question, and I wish I could say that we had better data on that. One thing we tried to do in this research was to study from the beginning a fairly wide array of tasks, in a deliberate attempt to capture what's common to all of those tasks.

We are far from the first people to study group performance or what are the characteristics of effective groups, or what things might make groups more effective. The novelty was having each group do a bunch of different tasks, and sort of look for the commonalities among them.

So in order to do that, we had to pick tasks that did not really require specialized expertise. Because otherwise, we might not even be getting group performance. We might just have one person doing everything and everybody else sitting back and watching, and that's not the kind of group interaction we wanted it to simulate.

Of course, that would be a good strategy in some cases, right? Like, if one of you is a surgeon and the rest of you aren't, well that guy should do the surgery and the other people should watch, rather than everybody sticking their hands in the patient's body and messing around. And what goes on in the real world sort of exists on that continuum.

Now, previous research — not by us, but really pretty good quality research I think — has found that often groups do not access the expertise and the ability and the knowledge of the most expert people in the group. That factors like personality factors, and social factors, and people expressing confidence, and people being the first to speak and things like that, can often override substantial differences in expertise or knowledge. And that people also often conceal ... Maybe not deliberately, but they sort of fail to surface their own special knowledge and expertise about things groups are working on together.

So we think that collective intelligence is sort of a phenomenon at the group level which is quite influenced by these kinds of social interactions. Even when you have experts, you have people who are clearly superior performers, they might not get to express all of that. And the group's efforts might be even worse than the best individual.

Julia Galef: So the groups that you were studying were, I think, basically strangers working together for the first time. Do you think that the effect of social sensitivity might wash out once group members get to know each other, and are better able to read cues? About, like, who has a thing to say but isn't talking, because the conversation is too chaotic for him. Or who do I expect would have good input here, because we've worked together a bunch of times, et cetera.

How much do you expect this to hold up over repeated work?

Chris Chabris: Yeah, I think what you think describe is kind of the ideal of how we would like groups to evolve over time. That, ideally, people should start to pick up on those things, and they should arrive at patterns of interaction that, maybe not optimize, but improve their ability to use individuals' expertise and knowledge, and so on.

We haven't really done a lot of long-term studies of groups, but we do have some ... There's a study that I wasn't involved in, but I really like that a bunch of my colleagues did, which was a study of ... What's the Riot Games game that everybody's playing?

Julia Galef: Is it League of Legends?

Chris Chabris: League of Legends, right. Yeah, so League of Legends teams, with the collaboration of the company, which I heard was wonderful to work with, League of Legends teams took our collective intelligence tasks. Which had nothing to do with expertise in League of Legends or anything like that.

And teams with higher ratings in the game and higher levels of achievement did better on our test, and also did better in the future. So it wasn't just retrospective, but also continued to perform well in the future.

So one might imagine that being really good at League of Legends is kind of like a very specialized thing and you learn a lot about your teammates and so on, but it still seems to make a difference, to have a team that does well in sort of like this generic collective intelligence test.

Julia Galef: So that sounds very plausible to me, but the thing that's surprising about it is, if I'm understanding how gaming works — it's remote, so people don't have an ability to read social cues off of each other's faces. Which is what I thought social sensitivity was capturing. Why would that still hold?

Chris Chabris: Yeah, that's another great question. I'm sure you're tired of hearing, "Great question".

Julia Galef: No, no. I never get tired of that.

Chris Chabris: So in our original studies, we used this “mind in the eyes” test, which we sort of initially interpreted as, in a sort of a narrow way, as perhaps just a test of what it seems like on its face, so to speak, which is the ability to read subtle cues from facial expressions. Therefore, you would wonder, if we can't see the other people in our group, what differences does being good at that make?

We did a study to test exactly this. In this study, we randomly assigned people to, once they were in a team ... They came to the lab, and they were put on a team. And then a team was randomly assigned to either sit so they were all facing each other and they could see each other's faces, or to sit in cubicles facing the wall and not even really knowing which other people in the room were on the same team as them.

Then they did this collective intelligence battery in an online forum, so that groups in either condition, face-to-face or cubicles, did the exact same battery, and there was a chat room that recorded everything they typed. In either case, online or face-to-face, purely online or online plus being able to see each other's faces, the mind in the eyes test still was correlated with the collective intelligence of the group.

Julia Galef: Interesting.

Chris Chabris: As a reminder, this is the “mind in the eyes” performance of the individual, so every individual does the test by themselves-

Julia Galef: Right.

Chris Chabris: They average the score of the team members, and that average score of the team members is still positively correlated with the team’s collective intelligence. Even when they never look at each other during the collective intelligence tests, and don't even know who in the room is on their team.

It seems to be measuring something deeper than just perceiving facial expressions, even though that's the medium that it uses for the test items. It seems to be measuring some deeper capacity and social intelligence theory of mind ability — the ability to understand and represent what other people are thinking, what they know, what their emotions are. Which may come through in text or in other subtle behaviors that get expressed online.

Julia Galef: Right. Going back to the percentage of women factor, are you aware of any correlational studies of real-world teams of women and whether they tend to perform better than men?

For example, has anyone looked at start-ups that were founded by more than one person, two or more people? You could measure the percentage of women in the founding team, and then you could look at probability of being profitable five years later or something like that.

Is there any kind of real-world correlation that would back up the experimental finding?

Chris Chabris: There are some studies that I'm aware of, and I should say that the start-up study that you mentioned was an idea that we also had a few years ago that I wanted to pursue. I thought the ideal environment for doing that kind of study was to look at a bunch of founding teams that were all at the same stage of starting up, maybe everybody who applied to Y-combinator… Where teams were accepted into a batch, and then you could follow how they go along and so on.

I contacted Paul Graham and other people and so on, and tried to drum up some interest in having everybody do a 45-minute collective intelligence test at the beginning of their incubator time. Then we'd just passively gather information on what happens to their companies over the next year or months and so on. Never got that off the ground. I think-

Julia Galef: Oh, too bad. It's a good idea.

Chris Chabris: I'm not sure start-ups really want to be studied, or that the people who are funding the start-ups want them to be participating in studies. As opposed to inventing stuff and marketing it, and so on.

But I like your idea of looking at observable characteristics like number of women and other factors. There is some data from studies of boards of directors. There's some argument that companies whose boards of directors that have more women do better, and I would like to think that that's because the collective intelligence of the board is increased — and therefore, whatever influence the board has on the company generates positive results and so on.

I think all those connections are probably a bit tenuous, and the causality could actually be reversed. It could be that companies that are doing really well, in a sense, can afford to now attend to questions of diversity and representation and so on, that struggling companies may just not have the attention or other capacity to pay attention to.

Julia Galef: Right.

Chris Chabris: There was one other study, by the way, that I find a little bit more convincing. There's a guy at Harvard Business School whose name escapes me right now, because I don't have his book in front of me. He did a study of equity research analysts on Wall Street in the '90s, but these are the people who analyze companies. They say buy, hold, or sell, and set price targets and things like that.

Their performance is measured in — it's a little bit of a fuzzy metric — by how highly their customers rate them. Still, it turned out that when these analysts got very highly rated, they tended to be poached by competing banks who would hire them. When that happened, their performance tended to go down — but if they were women, their performance recovered faster than if they were men.

One interpretation of that is that what you're really measuring here is not the performance of this one person, but the entire team that they're a part of. Therefore, there could be some effect of women adding more to team collective intelligence or something, that leads to better output.

Again, these are correlational studies. The data is not as good as we would want, so maybe someday those start-ups will want to study them.

Julia Galef: I want to raise a general concern that I have about studies that find that women are better at something than men. My general concern is I worry that the opposite result would be unlikely to be published.

If a researcher did a study that seemed to show, "Oh, hey, when you add women to a team, the team does worse," is that paper going to get published? It just seems so inflammatory that my suspicion is that either a journal would be reluctant to publish it, or would subject it to much more stringent standards to make sure it's a real result, to avoid publishing a false inflammatory study. Which — stringent standards are good, but if you're applying them unequally, then that affects the ratio of findings that you end up seeing. Or maybe the researcher himself or herself wouldn't try to pursue that finding because of the potential fallout.

I just don't know how to interpret the findings that I see published and shared that show that women are better at a thing, because I don't know what the denominator is. I don't know what the other potential studies found or would have found if they had been allowed to be conducted.

How do you think about this? Do you think my concern is a real one?

Chris Chabris: It's a very sensible concern. We should always be concerned about publication bias, and there are so many filters and publication biases of all kinds, right? There's so many filters that occur between the conceptualization of a study, and not only what gets published in journals, but maybe especially what gets publicized after it is in journals or in conferences or something like that. You're even more likely to hear about some kinds of studies, and they are likely to get published in journals and so on. I think that's always sensible.

Going back in time, you could say that probably there was a time in the past when the opposite publication bias existed, where if someone found in their data that women were better than men at something, then that might have been less likely to get published.

Julia Galef: Yeah, that’s plausible. That makes a lot of sense to me. But it doesn't get rid of the concern that any research about which gender performs better than the other is hard to interpret. The denominator is especially unknown.

Chris Chabris: Yeah, it's harder to interpret. I agree with that. And we always don't know what was not published. It's very hard to know what was not published, and even more so, what's not being done. What hypotheses are not getting tested.

I can say in our case that we didn't have a hypothesis about that from the outset.

Julia Galef: You did or didn't?

Chris Chabris: No, we did not have a hypothesis.

Julia Galef: You did not. Okay.

Chris Chabris: It was surprising, and it replicates. At least in our data, that replicates. It's not really a huge effect. It's not the biggest effect you would ever see, but it does tend to replicate.

I share your general concern, that the social desirability of the research outcome, among whoever is the gatekeeper, someplace along in the process, can definitely affect that. And I would be somewhat concerned.

I think one solution is increasingly open data. So there are more and more large data sets being made more and more open and available, so people can look at sex differences and other differences in the original data. If articles about sex differences in this data set have not been published, people can download the data for themselves and look at it and start to point it out. Not sure of really a good universal mechanism for fixing that.

I should say, I've looked at sex differences in other contexts, too, such as spatial reasoning, like performing mental rotation tasks and found both the normal improved performance or better performance by men in some of those tasks, but also some indications of why that might be that don't necessarily have to do with some absolute better performance in spatial cognition.

Again, we're getting into a little bit of the weeds of some of this stuff, but I think it's a fair concern, and it's a good thing to think about when you read about these kinds of results.

Julia Galef: Cool, let's shift tracks at this point. I wanted to ask you about an op-ed that you wrote. Actually, it's a topic that you've touched on in several pieces that you've written, about companies experimenting on their customers, or on the public.

Your argument was basically that people get upset when they find out that, "Oh, Facebook was doing A/B testing where some users were subjected to more emotional content than others, and Facebook was studying how this affects people's posting habits and things like that."

People were really upset about this, and your argument was that they shouldn't be upset. You want to lay out the case?

Chris Chabris: Sure, so this work was done in collaboration with my wife, Michelle Meyer, who's a bioethicist and legal scholar, and she's actually done more of this than I have. She should get the majority of the credit for this line of thinking.

But the basic idea is that there are often ... Often, I don't know. There have been many high-profile cases, especially in the world of people who focus on research ethics and things like that, of randomized experiments that have been run either by companies or by other kinds of organizations, even by medical researchers, where people object to the idea of the experiment. They say the experiment itself was unethical, shouldn't have been run, and is really, really bad for a variety of reasons.

What we noticed and what Michelle noticed, especially, is that people were complaining about A/B tests. That's just an experiment where people are assigned to either an A or B condition.

But they rarely complain about just changes in policy or practice, which affect everybody without comparing them to anything else.

Julia Galef: Right.

Chris Chabris: If Facebook changes its algorithm one day, they're, in a way, just running a really bad experiment where we're all in the A condition, and there's no B condition to compare to. We don't object to that. We don't object to doctors deciding to practice medicine one way and not another, but sometimes when people do randomized experiments even in medicine, there's objection to that.

Michelle and I pointed out a few cases of this, and we call it the “A/B Illusion,” when people object to an A/B experiment, but they would not object to just imposing A or B on everybody as a matter of practice.

One example we gave in one of our pieces was companies being reluctant to run beneficial experiments from which they could learn a lot, because they don't want people to find out that they've been running an experiment. Instead, they don't run an experiment; they just go with lesser quality data or no data at all, or just intuition. Or as someone said, the “HIPPO,” the highest paid person's opinion, just governs the outcome.

That's to us, and probably to a lot of people, not the most enlightened way to figure out what policies and treatments and practices are likely to work best, either for the company, for the company's bottom line, or for their customers. Companies very often are concerned about the welfare of the customers. They honestly try to make products that improve people's lives, so everybody has a stake in this, in this illusion, I think.

Julia Galef: This reminds me of an old anecdote — I have no idea if this is real or not — where some prestigious, esteemed doctor ... I think it was a surgeon ... was reporting some change in surgical methodology and arguing that this would be a good thing, and some student in the back raised his hand and was like, "Why don't you try it on only half your patients to see if it works well?"

The presenting surgeon took umbrage at this, and he was like, "You're seriously telling me that we should subject half of our patients to worse treatments just for the sake of experimentation? I'm not going to subject half the people to worse treatment."

And the student just replied, "Which half?"

Chris Chabris: Yes, exactly. Yeah. Yes. Exactly, so we can talk about all the cognitive biases and thinking traps that might lead us to believe, after an experiment has been run, that we knew all along what the results would be or we should have known what the results would be or we should have known that people would be harmed. We could make up a lot of explanations for that and so on.

To me, I think it's interesting that randomized experiments, kind of like randomness in general, tend to trip up our intuitive thinking processes. I think part of the explanation for that is that the first randomized experiment was done something like 200 years ago, and it wasn't even really followed up on very much. It was less than a hundred years ago that the proper theory and statistical tools for doing randomized experiments were invented, and there's still a little bit unintuitive for people to think about.

I think it partly has to do with the fact that they're a brilliant social invention, but a very recent one, that maybe we should spend more time teaching people about really in schools or something like that, or try to make people understand more. They're a really powerful ... I think you would probably agree, and people listening to this podcast probably would agree ... a really powerful evidence-generating mechanism, knowledge-generating mechanism for human society and for all of us. I think we should really try to get people to understand them much better than they do and then not react emotionally with fallacious reasoning in cases like this

By the way, I should add, this is not to say that there's nothing ever wrong with any A/B test. The principles of ethics say that there are various situations when it would be unethical to do an experiment. For example, if you know that one of the treatments is clearly superior. If the evidence, properly construed, shows that one of them is clearly superior, then it's probably not right to give people an inferior treatment, especially when health is involved.

There are lots of reasons why you can't just willy-nilly experiment and whatever experiment should be okay. There could be reasons why one might be unhappy, let's say, with what Facebook is doing with its platform, but they probably don't have much to do with the fact that they're doing A/B tests. We shouldn't let our dissatisfaction with whatever's going on with Facebook spill over into just disapproving of running A/B tests online in general. That would be a mistake.

Julia Galef: Yeah, you also made a great point recently when Starbucks announced that they were going to start giving implicit bias training to all of their employees to reduce the incidents of unfortunate things like ... I guess it was two weeks ago, that two black men in a Starbucks were ... They called the police on these guys for loitering, even though everyone loiters in Starbucks.

You pointed out that we really don't know if implicit bias training works, but if you're going to do it anyway, you might as well A/B test it so at least we'll get some more real world data on it. It's a waste to do all this training without information collection to boot.

Chris Chabris: Yeah, my colleague here at Geisinger, Matt Brown, and I wrote a piece in the Wall Street Journal about this. It seemed like Starbucks reacted with a very proactive response to this whole incident in Philadelphia, and they announced they were going to close their stores for a whole afternoon and everybody in all these 8,000 stores are going to get training. My first thought was, "What a great opportunity to run an experiment and see if any of this training actually works."

I don't think they were committed to implicit bias training per se, which has a very checkered, I think, evidence base, even according to some of the leading scholars on the topic. They are not convinced that somehow training people to reduce implicit bias actually reduces discriminatory behavior. That's a big open question.

I think it's a great example of how we sort of just rush to assume that we know what's going to work, even when it doesn't. Starbucks could have run any number of experiments. They could have delayed this by a month. They could have, even today, they could decide to hold back some of their stores and just delay the training in some of those stores to see whether the training they're gonna give everybody actually works.

It would help social scientists. It would help other companies who want to actually do effective training to have one of the biggest retail organizations in the world do a real experiment. And it's not really that much harder to do an experiment than it is to train 200,000 employees.

Julia Galef: I know, that's the maddening thing. That it would be so easy and so valuable and we're just not doing it.

Chris Chabris: Yeah, exactly. The added expense of the experiment is really not that much compared to what they are already investing.

Telling a company to go and invest precious resources in an experiment is kinda hard to do, from the outside, not knowing what all of their considerations are and so on. But having seen what they're already putting into this, now it's easy to say, "Well, if you just add the experimental component it would be great."

But it doesn't look like they're going to do it, unfortunately.

Julia Galef: I did want to push back a little bit on your defense of Facebook — and you also defended OkCupid's A/B testing on its users in the same op-ed.

It seemed to me that there are two separate things that we are talking about when we talk about the public's negative reaction to experiments.

One is A/B testing, or random experiments, when the participants know that they're in an experiment. There's, as you've correctly pointed out, people have an aversion to the idea that if we even have a suspicion that one thing might be better than the other that we should just give that thing to everyone — even though, in practice, we're often wrong about which is better, and it's much better to know than just to guess. So that's one thing.

But then a separate concern is: is there a problem with experimenting on people when they don't know they're in an experiment? And one argument you could make — that you, in fact, did make — is, “Well people are in experiments anyway. Not A/B tests, but “A” tests, when a company tries a thing on everyone.”

But it feels to me like that there's a real difference when the company is doing a thing that doesn't feel like it was part of the bargain when you signed up to use their service. So the Facebook example feels kind of borderline to me, although I would lean on the side of I guess it's okay for Facebook to do this A/B testing.

But the OkCupid example feels more over the line into “not okay” territory to me. So that was, OkCupid told some of its users that its matching algorithm had determined that they were a good match with someone else — when in fact, they weren't a good match according to the algorithm. And OkCupid discovered that actually, people hit it off just fine with those who they were secretly not a good match with according to the algorithm.

Which was interesting and useful to know. But it still felt like a violation to me, because the company was being actively deceptive instead of just giving some service to some people and not to others. What do you think?

Chris Chabris: I agree with you that active deception is ethically problematic. I won't disagree with that —

Julia Galef: Oh, I should also add that I think it's strategically unwise — just from a scientific perspective, forget the PR issues. Because, let's say you conduct an experiment without letting people know that you're experimenting on them, and you find a result. What are you gonna do with that result?

Presumably, in the future — let's say you're a medical scientist and you want to give someone a drug, a placebo without telling them it's a placebo. Great. So you find out that if people don't know they're getting a placebo, and don't even know there's a chance of getting a placebo, it helps them.

Well, what do you do with that? In the future, presumably at some point, you're gonna have to tell patients they might be getting a placebo, unless you want to lie to everyone for the rest of time. So, you have this result that shows that the placebo works if patients have no idea there's even a chance of getting a placebo. But now, in the real world, they know there's a chance of getting a placebo and so the result that you got in your experiment is no that longer applicable.

And I think the same thing probably applies to research like OkCupid's, where people don't know that there's even a chance that they're getting a random result.

Chris Chabris: Yeah, so first of all, let me say, informed consent is obviously a desirable thing whenever possible. However, a lot of experiments lose their validity when that's not involved. Or it's just not practical to do that. And they don't involve significant risks.

So in the case of OkCupid, it's funny, the assumption that there's deception going on is in a way, based on the assumption of the validity of the algorithm. Right? So if the algorithm really does match us up well, well then telling us a good match with someone we're really a bad match with is deceptive. But if you don't really know how good the algorithm is, then the ou-

Julia Galef: Well, the claim is that the company thinks they're matching people well.

Chris Chabris: Yeah.

Julia Galef: And they may not be, but they think they are.

Chris Chabris: Yeah. Or they don't know. They've created an algorithm, which sort of seems sensible on its surface, it seems like a sensible policy. Right? But they don't actually know. I would certainly not advise companies to go doing exactly what OkCupid did all the time.

I actually have the impression that they sort of enjoyed having the reputation of being the dating site that did those kind of things. They publicized it a lot in their blogs. Christian Rudder wrote a book and so on.

But I do think that there are many cases where informed consent isn't possible. And there are many cases where we want people to experiment. I think we sort of want the chef to try out different things in the restaurant and change the menu around and so on. We don't sort of want there to be sort of one menu that's set and fixed and it's never experimented with. We don't necessarily want to sign a consent form when we go into the restaurant that the chef might vary the ingredients a little bit from night to night or-

Julia Galef: Yeah, but, if the chef told us that we were getting the best or most expensive fish, and he actually gave us the cheaper fish, that might be a useful experiment to do, but I think people would still object.

Chris Chabris: Well, and those, yeah, you could do those things actually, under sort of pretty normal informed consent standards. Right? It turns out expensive wines don't taste that much better than the cheap wines when people don't know. Right?

Julia Galef: Yeah.

Chris Chabris: So that kind of work is done. So, yeah, I'm not saying that sort of every sort of undisclosed manipulation and so on is a good and okay one. I think companies should certainly weigh that.

But at the same time, we should realize that, as our friend Duncan Watts says, "The world is just the A condition of an unrun A/B experiment." And people are doing things to us all the time that they don't ask consent for. Like when OkCupid made their dating algorithm, they didn't disclose the algorithm in all its details and say do you consent to be matched up under the rules of this algorithm. Instead, they said, “We have an algorithm,” I don't know exactly what they said, but it must have been something like, “We have an algorithm that will match you up,” and you sort of assume that it's a good one.

But, that's a somewhat unwarranted assumption, just like we assume that a lot of medical practices, for example, are evidence-based and based on good information when often they're not. I mean, think how long it took them to start washing their hands in hospitals and so on. There's a lot of stuff that still goes on that is not as evidenced based as we might think.

And I think one thing Michelle would say, if she were here on this, is that probably we should do more, in general — companies, hospitals, institutions should do more to explain to their users, customers, employees, whatever, that they are organizations that try to learn and improve over time. And that one of the ways they do that is by doing low-risk experiments, that are not meant to put you in danger but are meant to compare different policies, ideas, so on, and figure out what works best so that everyone can benefit. I think if that were communicated a lot more clearly and continuously, that might help.

Julia Galef: Yeah, well, I can't argue with that.

Chris Chabris: I hope to say something that can't be argued with. That's my favorite.

Julia Galef: Chris, before I let you go, I wanted to ask you for a recommendation of a book or article or blog or something like that that you don't agree with or you have substantial disagreements with, but you still think is worthwhile and worth engaging with. Because it makes interesting arguments, or it advances an interesting hypothesis, something like that. What would you recommend in that vein?

Chris Chabris: Well I've heard you ask this question before so I did a little preparation. And I came with not one, but four answers.

Julia Galef: Oh. Fantastic.

Chris Chabris: So I will, actually I will rattle them off a little bit quickly. I'll start with Nassim Taleb, who probably has been discussed on your podcasts before and is familiar to listeners.

Julia Galef: The Black Swan.

Chris Chabris: Yes, he wrote the Black Swan, Fooled by Randomness, Antifragile, Skin in the Game, a variety of books with wonderful titles. He's a very interesting guy. He says a lot of things that one can disagree with, but I think that it's still very rewarding to consider his ideas, and try to get beyond some the rhetoric and occasional bombast and drama and so on, and think about what he's actually saying. And it probably will change people's worldviews a little bit if they haven't been exposed to it already.

Julia Galef: Yeah, he's a good test case for me. Because I find his style, especially on Twitter, so abrasive and obnoxious that I have to really work to consider his claims on the merit and not let my judgment be colored by my impressions of him as a person.

Chris Chabris: Yeah, you can use it as training. You could use it as training for active open-mindedness. And separating out the person from the arguments and so on.

Julia Galef: Exactly.

Chris Chabris: It has so many different uses, his work. And I'm sure it's extremely unpleasant to come under attack from him. And I know people who have. But I also know him, and I just think there's something to be gained from thinking about his thought and that he's somewhat of a significant thinker. And it's worth reading his stuff and thinking about it.

The second one I'll mention is Malcolm Gladwell. And I've-

Julia Galef: Interesting, because you wrote a serious critique of Malcolm Gladwell a few years ago. I think that's how I first encountered you, actually.

Chris Chabris: Yeah. I did write a couple of pieces about Malcolm Gladwell and one of his books which was called David and Goliath, which came out a few years ago. And I have had a lot of objections to claims that Gladwell makes, and the way he thinks about evidence, and the way he sort of implicitly leads his own readers to think about evidence and to think about how the mind works and so on.

But, I do read all of his stuff. I can't say I've read every single article he's written, but I do try to make it a point of reading all of his stuff because he always manages to find interesting things that nobody knows about. Or to show well-known things in a different light. So it's worth thinking about what he has to say, and looking at the items he's picked to talk about and write about.

And it's also, in a sense, good training, to look and see whether are there reasons to disagree with him. Don't accept him at face value. He's also a great writer so it's good to have examples, just of engaging prose, and the way to write so that people will actually want to read you and so on.

So I use Gladwell's stuff in teaching. For example, when I used to teach seminar on writing and so on, we would use that stuff and try to dissect sort of what makes this effective writing, on the one hand, but also what makes it potentially not a correct, or even a misleading account of human behavior and how the world works.

Julia Galef: Maybe one of the keys to finding useful people you disagree with is just finding different ways to read their stuff, like with different purposes in mind. So if your purpose in reading Gladwell's stuff is to find interesting stories, or data, or anecdata that you can then add to your own world models and interpret yourself — rather than taking his interpretations of those data — that might be a more useful way to use a Gladwell.

Chris Chabris: Exactly. And he performs a great service of going out and gathering some of the best of that stuff. And sometimes he's right and sometimes he's wrong. So if you just go into it, not assuming that he's right, but trying to think about how he might be wrong, then you're in good shape, I think.

And I guess, very quickly, I'll mention Ray Dalio's book. He's the founder and CEO of Bridgewater the hedge fund. And he wrote a book, I think it's called Principles. It's the first volume of this Principles book, I don't know if you know it or not, but-

Julia Galef: Yeah.

Chris Chabris: ... he describes a very, sort of the very unique way that his business has operated, and the way that he thinks businesses should operate. And I found a lot to disagree with in it, because I think it's, a lot of it is sort of based on fairly simplistic models, again, models of the mind and the brain. And of what personality differences between people really mean, and what are important and so on.

But to hear the way a very successful person sort of has thought about this, and he's clearly thought about it a lot, it was a very interesting counterpoint, to how academics think about the same stuff. Here's a guy who's tried to put into practice — what does he come out with on the other end? How does he try to use this stuff? And it was really thought-provoking.

Julia Galef: Cool. Were those four, or three? Did you have a fourth one?

Chris Chabris: That was three, yeah. I guess the last one I'll mention, 'cause I happened to write down four, was Edward Tufte, who's written these beautiful books on information design and graphics. I think the first one was called The Visual Display of Quantitative Information and then he's got one called Envisioning Information. He gives seminars and so on.

So I think he's kinda wrong about a lot of the visual perception psychology about sort of what kind of charts are ... convey their message most effectively and how should graphics be designed to be easy for people to understand. But he produces incredibly beautiful graphics and he tries to visualize information in unusual ways.

So again, it's sort of like the same kind of thing, look at it and marvel at the beauty and the eloquence and the interesting stuff that's in there, and then think about, “Would this really work, as a way of communicating a message to an audience, and how could you do it differently?”

Julia Galef: Fantastic. Well, I always love getting four for the price of one. That's a good day, in Rationally Speaking land. Chris, thank you so much for coming on the show. It's been a pleasure having you.

Chris Chabris: Oh well, thanks for having me, it was a really fun conversation.

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