PRESS EMBARGO: PUBLICATION 12AM (EST), TUESDAY, NOVEMBER 18TH

Edge

"To arrive at the edge of the world's knowledge, seek out the most complex and sophisticated minds, put them in a room together, and have them ask each other the questions they are asking themselves."

HeadCon '14

In September a group of social scientists gathered for HeadCon '14, an *Edge* Conference at Eastover Farm. Speakers addressed a range of topics concerning the social (or moral, or emotional) brain: Sarah-Jayne Blakemore: "The Teenager's Sense Of Social Self"; Lawrence Ian Reed: "The Face Of Emotion"; Molly Crockett: "The Neuroscience of Moral Decision Making"; Hugo Mercier: "Toward The Seamless Integration Of The Sciences"; Jennifer Jacquet: "Shaming At Scale"; Simone Schnall: "Moral Intuitions, Replication, and the Scientific Study of Human Nature"; David Rand: "How Do You Change People's Minds About What Is Right And Wrong?"; L.A. Paul: "The Transformative Experience"; Michael McCullough: "Two Cheers For Falsification". Also participating as "kibitzers" were four speakers from HeadCon '13, the previous year's event: Fiery Cushman, Joshua Knobe, David Pizarro, and Laurie Santos.

We are pleased to present the program in its entirety**,** nearly six hours of *Edge* Video and a downloadable PDF of the 58,000-word transcript.

EdgeVideo

[6 hours]

Edge URL: http://edge.org/event/headcon-14.

EDGE.ORG
John Brockman, Editor and Publisher
Russell Weinberger, Associate Publisher
Nina Stegeman, Editorial Assistant

Published by Edge Foundation, Inc.
260 Fifth Avenue
New York, NY 10001

Edge Foundation, Inc. is a nonprofit private operating foundation under Section 501(c)(3) of the Internal Revenue Code.

Related reading on *Edge:*

HeadCon '13
Edge Meetings & Seminars
Edge Master Classes

ABOUT EDGE.ORG

"Take a look. No matter who you are, you are bound to find something that will drive you crazy." *—New York Times*

"The world's smartest website; a salon for the world's finest minds." *—The Guardian*

"...A collection that reads like the best TED talks ever. It's an absolute pleasure to read." —Fareed Zakaria, GPS, CNN

"We'd certainly be better off if everyone sampled the fabulous Edge symposium which, like the best in science, is modest and daring at once." —David Brooks, *New York Times*

"Brockman's seeds of a new intellectualism have bloomed in the culture of ideas that has become so popular in the past years in the pages of magazines such as Atlantic and New Yorker, in numerous nonfiction bestsellers, or in the various incarnations of the TED conference...Edge has remained one of the purest outlets of intellectual thought on the Web." *—Süddeutsche Zeitung*

"A wonderful opportunity to savor the thoughts of many top scientists and thinkers of the world." *—Boston Globe*

"An epicenter of bleeding-edge insight across science, technology and beyond, hosting conversations with some of our era's greatest thinkers....(A) lavish cerebral feast ... one of this year's most significant time-capsules of contemporary thought." *—Atlantic*

"The most stimulating English-language reading to be had from anywhere in the world." *—The Canberra Times*

"The inquiry becomes an a fascinating experience. The pleasure of intelligence is a renewable source of intellectual energy." *—Il Sole 24 Ore*

"Brilliant, essential and addictive. It interprets, it interrogates, it provokes. Each text can be a world in itself." *—Publico*

"Open-minded, free ranging, intellectually playful ... an unadorned pleasure in curiosity, a collective expression of wonder at the living and inanimate world ... an ongoing and thrilling colloquium." —Ian McEwan, *The Telegraph*

"[John Brockman] A kind of thinker that does not exist in Europe." —*La Stampa*

"Not just wonderful, but plausible." —*Wall Street Journal*

"Thrilling ... Everything is permitted, and nothing is excluded from this intellectual game." —*Frankfurter Allgemeine Zeitung*

"Fantastically stimulating...It's like the crack cocaine of the thinking world.... Once you start, you can't stop thinking about that question." —BBC Radio 4

"The brightest minds in the known universe." —*Vanity Fair*

[MORE: http://edge.org/edge-in-the-news]

______

Edge Foundation, Inc. is a nonprofit private operating foundation under Section 501(c)(3) of the Internal Revenue Code.

http://edge.org | Follow us on Twitter @edge https://twitter.com/edge

CONTENTS

Sarah-Jayne Blakemore: "The Teenager's Sense Of Social Self"

The reason why that letter is nice is because it illustrates what's important to that girl at that particular moment in her life. Less important that man landed on moon than things like what she was wearing, what clothes she was into, who she liked, who she didn't like. This is the period of life where that sense of self, and particularly sense of social self, undergoes profound transition. Just think back to when you were a teenager. It's not that before then you don't have a sense of self, of course you do. A sense of self develops very early. What happens during the teenage years is that your sense of who you are—your moral beliefs, your political beliefs, what music you're into, fashion, what social group you're into—that's what undergoes profound change.

EdgeVideo

[36.22]

SARAH-JAYNE BLAKEMORE is a Royal Society University Research Fellow and Professor of Cognitive Neuroscience, Institute of Cognitive Neuroscience, University College London. **Sarah-Jayne Blakemore's *Edge* Bio Page**

Lawrence Ian Reed: "The Face Of Emotion"

What can we tell from the face? There's some mixed data, but data out that there's a pretty strong coherence between what is felt and what's expressed on the face. Happiness, sadness, disgust, contempt, fear, anger, all have prototypic or characteristic facial expressions. In addition to that, you can tell whether two emotions are blended together. You can tell the difference between surprise and happiness, and surprise and anger, or surprise and sadness. You can also tell the strength of an emotion. There seems to be a relationship between the strength of the emotion and the strength of the contraction of the associated facial muscles.

EdgeVideo

[26:27]

LAWRENCE IAN REED is a Visiting Assistant Professor of Psychology, Skidmore College. **Lawrence Ian Reed's *Edge* Bio Page**

Molly Crockett: "The Neuroscience of Moral Decision Making"

Imagine we could develop a precise drug that amplifies people's aversion to harming others; you won't hurt a fly, everyone becomes Buddhist monks or something. Who should take this drug? Only convicted criminals—people who have committed violent crimes? Should we put it in the water supply? These are normative questions. These are questions about what should be done. I feel grossly *unprepared to answer these questions with the training that I have, but these are important conversations to have between disciplines. Psychologists and neuroscientists need to be talking to philosophers about this and these are conversations that we need to have because we don’t want to get to the point where we have the technology and then we haven’t had this conversation because then terrible things could happen.*

*Edge*Video
[44:00]

MOLLY CROCKETT is Associate Professor, Department of Experimental Psychology, University of Oxford; Wellcome Trust Postdoctoral Fellow, Wellcome Trust Centre for Neuroimaging. **Molly Crockett's *Edge* Bio Page**

Hugo Mercier: "Toward The Seamless Integration Of The Sciences"

One of the great things about cognitive science is that it allowed us to continue that seamless integration of the sciences, from physics, to chemistry, to biology, and then to the mind sciences, and it's been quite successful at doing this in a relatively short time. But on the whole, I feel there's still a failure to continue this thing towards some of the social sciences such as, anthropology, to some extent, and sociology or history that still remain very much shut off from what some would see as progress, and as further integration.

EdgeVideo

[39:34]

HUGO MERCIER, a Cognitive Scientist, is an Ambizione Fellow at the Cognitive Science Center at the University of Neuchâtel. **Hugo Mercier's *Edge* Bio Page**

Jennifer Jacquet: "Shaming At Scale"

Shaming, in this case, was a fairly low-cost form of punishment that had high reputational impact on the U.S. government, and led to a change in behavior. It worked at scale—one group of people using it against another group of people at the group level. This is the kind of scale that interests me. And the other thing that it points to, which is interesting, is the question of when shaming works. In part, it's when there's an absence of any other option. Shaming is a little bit like antibiotics. We can overuse it and actually dilute its effectiveness, because it's linked to attention, and attention is finite. With punishment, in general, using it sparingly is best. But in the international arena, and in cases in which there is no other option, there is no formalized institution, or no formal legislation, shaming might be the only tool that we have, and that's why it interests me.

EdgeVideo

[31:58]

JENNIFER JACQUET is Assistant Professor of Environmental Studies, NYU; Researching cooperation and the tragedy of the commons; Author, *Is Shame Necessary?* **Jennifer Jacquet's *Edge* Bio Page**

Simone Schnall: "Moral Intuitions, Replication, and the Scientific Study of Human Nature"

In the end, it's about admissible evidence and ultimately, we need to hold all scientific evidence to the same high standard. Right now we're using a lower standard for the replications involving negative findings when in fact this standard needs to be higher. To establish the absence of an effect is much more difficult than the presence of an effect.

EdgeVideo

[42:15]

SIMONE SCHNALL is a University Senior Lecturer and Director of the Cambridge Embodied Cognition and Emotion Laboratory at Cambridge University. **Simone Schnall's *Edge* Bio Page**

David Rand: "How Do You Change People's Minds About What Is Right And Wrong?"

What all these different things boil down to is the idea that there are future consequences for your current behavior. You can't just do whatever you want because if you are selfish now, it'll come back to bite you. I should say that there are lots of theoretical models, math models, computational models, lab experiments, and also real world field data from field experiments showing the power of these reputation observability effects for getting people to cooperate.

Edge Video

[34:37]

DAVID RAND is Assistant Professor of Psychology, Economics, and Management at Yale University, and the Director of Yale University's Human Cooperation Laboratory. **David Rand's *Edge* Bio page**

L.A. Paul: "The Transformative Experience"

We're going to pretend that modern-day vampires don't drink the blood of humans; they're vegetarian vampires, which means they only drink the blood of humanely farmed animals. You have a one-time-only chance to become a modern-day vampire. You think, "This is a pretty amazing opportunity, but do I want to gain immortality, amazing speed, strength, and power? Do I want to become undead, become an immortal monster and have to drink blood? It's a tough call." Then you go around asking people for their advice and you discover that all of your friends and family members have already become vampires. They tell you, "It is amazing. It is the best thing ever. It's absolutely fabulous. It's incredible. You get these new sensory capacities. You should definitely become a vampire." Then you say, " Can you tell me a little more about it?" And they say, "You have to become a vampire to know what it's like. You can't, as a mere human, understand what it's like to become a vampire just by hearing me talk about it. Until you're a vampire, you're just not going to know what it's going to be like."

[48:42]

L.A. PAUL is Professor of Philosophy at the University of North Carolina at Chapel Hill, and Professorial Fellow in the Arché Research Centre at the University of St. Andrews. **L.A. Paul's *Edge* Bio page**

Michael McCullough: "Two Cheers For Falsification"

What I want to do today is raise one cheer for falsification, maybe two cheers for falsification. Maybe it's not philosophical falsificationism I'm calling for, but maybe something more like methodological falsificationism. It has an important role to play in theory development that maybe we have turned our backs on in some areas of this racket we're in, particularly the part of it that I do—Ev Psych—more than we should have.

EdgeVideo

[43:37]

MICHAEL MCCULLOUGH is Director, Evolution and Human Behavior Laboratory, Professor of Psychology, Cooper Fellow, University of Miami; Author, *Beyond Revenge.* **Michael McCullough's *Edge* Bio page**

Also Participating

FIERY CUSHMAN is Assistant Professor, Department of Psychology, Harvard University. JOSHUA KNOBE is an Experimental Philosopher; Associate Professor of Philosophy and Cognitive Science, Yale University. DAVID PIZARRO is Associate Professor of Psychology, Cornell University, specializing in moral judgment. LAURIE SANTOS is Associate Professor, Department of Psychology; Director, Comparative Cognition Laboratory, Yale University.

Sarah-Jayne Blakemore: "The Teenager's Sense of Social Self"

The reason why that letter is nice is because it illustrates what's important to that girl at that particular moment in her life. Less important that man landed on moon than things like what she was wearing, what clothes she was into, who she liked, who she didn't like. This is the period of life where that sense of self, and particularly sense of social self, undergoes profound transition. Just think back to when you were a teenager. It's not that before then you don't have a sense of self, of course you do. A sense of self develops very early. What happens during the teenage years is that your sense of who you are—your moral beliefs, your political beliefs, what music you're into, fashion, what social group you're into—that's what undergoes profound change.

SARAH-JAYNE BLAKEMORE is a Royal Society University Research Fellow and Professor of Cognitive Neuroscience, Institute of Cognitive Neuroscience, University College London.

THE TEENAGER'S SENSE OF SOCIAL SELF

I'm Sarah-Jayne Blakemore from University College London. Today I'm going to be talking about the adolescent brain, which is the focus of my lab's research. I'm going to talk about the history of this young area of science, and I'll also tell you about some of the current questions for the future in this area.

I did my PhD on schizophrenia, and I also did a post-doc on schizophrenia. I became interested in the fact that schizophrenia is a devastating psychiatric disease that has its onset right at the end of adolescence. Normally people develop schizophrenia, on average, between about 18 and 25 years. This is interesting because it's a developmental disorder, but it develops much later than most developmental disorders. I became interested in whether that might be something to do with brain development during the teenage years going wrong in people who go on to develop schizophrenia.

This was about 12 years ago. Back then, I delved into the literature and, to my surprise, there was little known about how the human teenage brain develops. There were a handful of studies back in the year 2002, a small handful, but they were intriguing because even though there were only a few of them, they all pointed to significant and protracted development of the brain right throughout adolescence and into the 20s. This was an interesting finding because, prior to those papers, most neuroscientists would have assumed, and the dogma at the time I was an undergraduate and a graduate, was that the human brain stops developing some time in childhood and doesn't change much after mid to late-childhood.

What these papers suggested was that the dogma was completely wrong. In fact, the human brain continues to develop significantly across almost the whole cortex throughout the teenage years, and even into the 20s. This was an intriguing finding, but it also pointed to a massive gap in the field. There were very few papers, little was known, and there were so many questions that had yet to be answered.

I decided back then—in the year 2002, 2003—to change the focus of my research from adult studies on schizophrenia and other mental illnesses to developmental studies. In retrospect, that was a risky maneuver because I'd never done a developmental study before. It was the encouragement of my friend and mentor, Professor Uta Frith, that gave me the confidence to make that change, and also a fellowship from the Royal Society allowed me to take this relatively risky avenue.

In the past 12 or 15 years, a huge amount has been discovered about the development of the human brain throughout the teenage years. Many labs now work in this area, and there's been an explosion of research. We know a lot about the development of the adolescent brain, and I'm going to talk to you about that today, and about the questions that still remain, because there are many.

Most of the work has been done with structural imaging—structural MRI. That is the method that has changed the game in this area of research because before we were able to scan the living human brain with MRI, we weren't able to understand how the brain changes across development. Now we can. We can scan kids of all ages, as long as they keep still, which is not always the case. We're able to look at changes in brain structure, and also changes in brain function across the life span. That technology was the turning point in our understanding of the development of the brain. Now there's a huge amount of experimental behavioral studies on cognitive and socioaffective changes during the teenage years.

In my lab we're particularly interested in adolescent-typical behaviors. What I mean by that are behaviors that you stereotypically associate with teenagers, things like risk taking, heightened self-consciousness, and peer influence. There are a lot of nice examples of these behaviors, and I'm going to read one. This is a letter that was written to the *Guardian*, which is a British newspaper, a couple of years ago. This is a reader who says:

> There's nothing like teenage diaries for putting momentous, historical occasions into perspective. This is my entry for the 20th July, 1969.
>
> 'I went to arts center in yellow cords and blouse. Ian was there but he didn't speak to me. Got rhyme put in my handbag by someone who's apparently got a crush on me. It's Nicholas I think. Ugh.
>
> Man landed on moon.'

The reason why that letter is nice is because it illustrates what's important to that girl at that particular moment in her life. Less important that man landed on moon than things like what she was wearing, what clothes she was into, who she liked, who she didn't like. This is the period of life where that sense of self, and particularly sense of social self, undergoes profound transition. Just think back to when you were a teenager. It's not that before then you don't have a sense of self, of course you do. A sense of self develops very early. What happens during the teenage years is that your sense of who you are—your moral beliefs, your political beliefs, what music you're into, fashion, what social group you're into—that's what undergoes profound change.

We're particularly interested in the effects that peers have on adolescent decision making. It's well known that adolescents do take risks, and they probably take a disproportionate number of risks. However, if you give them an optimal situation, an optimal environment, they don't necessarily take more risks. If they're in a lab and there are no distractions, there are no emotionally motivational salient factors going on, they perform similarly to adults. They take about the same number of risks, depending on what task you use. When you give them some motivational context, for example, a couple of friends standing behind them, that's when you see heightened risk taking in adolescents. You don't just see it in the lab. We all know from epidemiological data and data from car insurance companies that that's borne out in real life, as well. Adolescents, for example, have more car accidents than older people, but the situation in which they have those car accidents is normally when they have a same-age passenger in the car with them.

We're interested in why adolescents are particularly susceptible to peer influence. One of the ways we've looked at this is to look at what happens when adolescents are ostracized by their peer groups. We've done this by using the pretty well-known game called Cyberball, which is a game of catch—a ball game—that you play over the Internet with what you think are two other people. In fact, they're not, they're programmed by the lab. You can program those two other players to either include participants in this ball game, or exclude them from the ball game.

When adults play the Cyberball game and they've been excluded from that game of catch, they feel sad, their mood lowers, they feel more anxious, and you can measure that. This has been done many times by labs around the world. We compared adolescents and adults in this Cyberball game, and we found that exactly the same response was found but even more so in adolescents. Their mood dropped even more than adults' mood did, and they became even more anxious than adults after being excluded from this game. That suggested that adolescents might be hypersensitive to being socially excluded. When you think about that in the context of adolescent decision making, it sheds adolescent decision making and risk taking in a more rational light.

Whenever you make a decision you weigh various pros and cons, various advantages and disadvantages. With something like speeding down the motorway, or texting while driving, you might think, "I'm going to get to my meeting on time if I speed," or you might get a kick out of speeding. On the other hand, you might crash, you might get caught by the police. You're weighing up these pros and cons, but there's also the social factor. We know that people behave differently in groups compared with when we're on our own, and having someone else observe your behavior changes your behavior on cognitive tasks. What we think is happening is that, in adolescence, that social influence is particularly heightened. This is a framework that my student, Kate Mills, and I have been working on recently.

If you take, for example, smoking. Say you have a 13-year-old girl, and all her friends are smoking. For her, what is the more risky decision, saying yes to a cigarette when she knows the risks associated with smoking—as all 13-year-olds do these days—or saying no and potentially ostracizing herself from her peer group? We think that because of the hypersensitivity to being ostracized by the peer group, saying no is probably more of a risk for adolescents.

One of the things coming out of adolescence studies is this idea that these stereotypical behaviors we associate with adolescents—risk taking and peer influence—are there for a reason. There's probably a good reason why adolescents care so much about being included by their social group and take more risks when they're with their friends. I'm not an evolutionary biologist, but it makes sense when you think about the need, the drive to become independent from one's parents, to go and explore the environment, and to affiliate with your social group during this period of life. One thing I'm not saying is that risk taking is bad, or that peer influence is bad. It's probably an important and adaptive process that we all need to go through in our transition between childhood and adulthood.

One of the questions that we've been interested in looking at is why and how is it that social influence has its effect on decision making and behavior in adolescents? There are lots of theories about why this is, and I'm not going to go into any detail. One of the things that we're interested in is the development of the social brain. What I mean by that is the network of brain regions that are involved when we do theory of mind—when we think about other people's minds, their intentions, their beliefs, their desires, their emotions.

There is a circumscribed network of brain regions that are activated when we do a mentalizing task. What various labs around the world have found is that that network of brain regions undergoes significant development, both in terms of structure—in gray matter and white matter development—and also in terms of function during adolescence. Specifically, a number of labs now have replicated the effect that a certain region of the social brain called the medial prefrontal cortex, is more active in adolescents than in adults when thinking about other people's minds. Even though adolescents and adults in these studies are just as good at the mentalizing task, they use a different level of activity in medial prefrontal cortex in order to do the task. Again, we don't know why this is, but we think it might have something to do with the cognitive strategy, or the mental approach to the problem. The way they solve these problems might require different levels of activity in the regions of the social brain network.

The development of the social brain during adolescence suggests that during this period of life the brain is particularly susceptible to social pressure, but also to the social experiences that adolescents have around them, and the social opportunities that are given to them. This brings me on to one of the major questions. There are many, many questions that still remain to be looked at in this area, but one of them is whether adolescence represents a sensitive period for brain development. You'll hear a lot of people talking about adolescence as representing a second sensitive period of brain development, but we don't have very much data on that.

We know from studies on early development of the brain, both from humans and non-human animals, that the brain undergoes different "sensitive periods" of development, meaning there are periods of development where the brain is particularly susceptible to certain types of environmental stimuli. We know lots about this in domains such as sensory input, and also language input in the first few years of life, just to name a couple of examples.

Given that the brain is undergoing a lot of development during adolescence, particularly in areas like prefrontal cortex and other cortical areas, many people have suggested that this might represent a window of opportunity—a second sensitive period for learning in cognitive and social domains. There is very little evidence on this yet. It makes a lot of sense, but it's still an open question. It's something that we and other labs are currently looking at, but again, there's not much to say about it yet because there are a couple of studies suggesting it might be in some domains.

If it's true that adolescence represents a sensitive period of brain development in some areas of cognition and social behavior, then that has implications for things like education; when to teach what, what's the best moment to teach calculus, or algebra? It also has implications for the social environment; should adolescents be experiencing certain types of social interaction experiences during that period of life? It has implications for things like the legal treatment of teenagers. At the moment, if teenagers do something naughty they are incarcerated with other teenagers who have done something naughty. Yet we know that they are particularly susceptible to peer influences. Is this the most rational thing to do? It's probably not a particularly productive solution.

If adolescence is a sensitive period for brain development, that is a double-sided coin because although it represents a period of opportunity in which the brain is particularly susceptible to acquiring new information in certain domains, it also might represent a period of vulnerability, in which the brain is particularly vulnerable to certain environmental inputs.

I've been briefly talking, in no detail at all, about average teenage brains. Most of the data we have comes from averaging over teenage brains, but there is no average teenage brain. There is no average teenager. The individual differences are much greater than the averages. We're only just starting to look at individual differences and how individual differences in both genetics and environment influence brain development. We know that they both do. We know that genetics—your genotype—and also your environment influence your brain development. Environmental influences are almost infinite and difficult to study, but there are things like stress, alcohol, drugs, your social group, your family environment, your culture, your peers, who you hang out with, all these things invariably will be shaping the way the adolescent brain is developing. They also might play an influence in triggering the onset of mental illnesses in people who are genetically predisposed to them.

Now I'm going back to where I began: mental illness. It's not just schizophrenia that has its onset during this period of life. It's also many other mental illnesses. Most mental illnesses—depression, anxiety, eating disorders, addictions—have their onset at some point during adolescence. There is something about adolescence that means this period of life is a window of vulnerability to these illnesses.

My aim 12 years ago was to map out the development of the adolescent brain and then move on to brain development in teenagers who become schizophrenic. I've just made just a tiny input into the former, and I'm not anywhere near doing work on schizophrenia yet, but other labs around the world have started to do that. I've mentioned the NIH study particularly, and there are lots of labs, but I have a collaboration with Jay Giedd at the NIH and he is one of the pioneers of this area, and they have done research on longitudinal studies looking at brain development in kids who then go on to develop some mental illness, or developmental disorder. The data is quite new, it needs to be replicated. It's interesting, but it's quite varied.

The one take-home message is that it is critical to look at development in these disorders. Rather than taking a snapshot of what the brain looks like in an 18-year-old with depression compared to 18-year-olds who don't have depression—the brain might look similar by that age—what's critical is the way it gets there; there can be different developmental trajectories that end up at more or less the same point. The analogy I'd give is, you might use the motorways, the freeways, or the A roads, ending up at the same point, but you take a different route to get there. That seems to be what's critical in a lot of these developmental disorders and mental illnesses. Looking at development, and not just taking a snapshot in time is important.

Finally, I wanted to talk about prevention of mental illness because, like I said, adolescence might give us a window of opportunity, not just for things like education and learning, but for intervention. There's a dogma in social policy and educational policy that the first three years of life are the critical window where you have to get in and intervene. What this research on the brain is suggesting is that the brain continues to develop; it is plastic, but in a heightened way, right throughout the teenage years, so it's not too late during the teenage years to intervene in the cases where people might need some extra help.

One of the interventions that is important is an intervention that prevents, or at least reduces the onset of mental illness in people who are susceptible to it. That's something that a lot of people are thinking about at the moment. How do you do that? Is a universal approach better than a targeted approach? One of the areas that we are about to start working on is looking at whether mindfulness meditation, as a universal treatment in schools, has any affect on wellbeing and lowering anxiety and stress, but also reducing the onset of mental illness in teenagers. There is some promising data on that from other people's labs. We haven't begun our studies on that yet, but we are about to next year.

One of the things that I have learned over the last 10 or 12 years of researching in this area is that it's critical to include your research subjects in every aspect of your experimental design. This is not something that I had done previously. There's a tendency for adults to think that they know best for teenagers when teenagers know probably a lot more than we do about what's best for them in terms of their education, in terms of their social environment, what they want to do.

It applies to experiments as well. These days we always include teenagers in the designs of our experiments, the designs of our stimuli, and I'll give you a couple of trivial examples where that's helped. Firstly, talking to teenagers about actual phenomena that they experience has led us to design experiments, or apply for grants to research those phenomena. We probably wouldn't have remembered what it was like when we were teenagers, and also things have changed. Each generation is different.

There was one experiment we did where we had a stimuli with a load of objects on it, and one of those objects was a tape—a cassette—and involving teenagers at a very early stage of our research made us realize that no teenagers know what a tape is. They don't recognize it, so we changed that. We're interested in peer influence, and we found from involving teenagers in the designs of our studies that what matters to them is not being observed by a peer, but having a peer monitor their behavior. Being told that this friend of yours is going to sit behind you, and after you've done this task, they're going to fill in a questionnaire about how you did, that's what matters to them. Again, we couldn't have guessed that from the adult literature, because there's no indication that that is a significant factor in adults. That's what we do now on the basis of suggestions by teenagers.

THE REALITY CLUB

JOSHUA KNOBE: I like the point that you were making about how these behaviors on the part of teenagers that might seem irrational can be seen as rational. I was just wondering why that same argument doesn't also apply to parents. These kids are in a situation where they could do the seemingly safe thing, or they could do the seemingly risk-seeking thing. If you think about it, the actual risk of the safe thing is great, because the risk is a reputational risk. You can see why adolescents would evolve to do these seemingly risk-seeking things that are actually the safe thing because they avoid having reputational punishments.

Why do we, as parents, not evolve in the same way? Why is it that if I saw my daughter doing a safe thing I wouldn't pressure her to do a more badass thing so that she would avoid the possible reputational punishments? Why is it that, as parents, we don't have that exact same mentality: "You must seek the risk-seeking behavior, or else you're going to suffer these reputational punishments?”

BLAKEMORE: We're not very good at taking the perspective of people who are different from us, and we've forgotten how important it was to impress your friends, or not to be ostracized by them. Not all teenagers are like this, and again, I want to emphasize that there are individual differences. Some teenagers never take any risks, some teenagers aren't susceptible to peer influence. I bet, around the room, certainly for me, if you think about your teenage years you did take risks when you were with your friends, not when you were on your own. Now you probably wouldn't. You just wouldn't. If a group of your colleagues were smoking cigarettes outside and they offered you one, you wouldn't mind saying, "No thanks, I don't smoke," as an adult. It's easy to forget the social pressure of not being excluded by your group, and the importance of what that feels like.

KNOBE: You're thinking of it as happening it at the proximate level, not at this ultimate level. The question I was asking was why didn't we just evolve as parents to pressure our kids to engage in risky behavior?

BLAKEMORE: Well, I guess risk taking has to be constrained. Risk taking is a good thing. If we didn't take risks, where would we be? On the other hand, it can be dangerous and can result in accidents or even death. If you look at mortality rates across the life span, the number one cause of death in adolescents is from risk taking; it's from accidents. That's not true at any other period of life. Risk taking is a good thing unless it goes too far. You need some constraints over risk taking, and that's where parents come in. That's probably evolutionarily important as well.

HUGO MERCIER: I have a quick question about the evolutionary history of the theory. I guess you could have two thoughts. One is that throughout our evolutionary history, there was this significant period of our lives during which we had to form cliques to find partners—sexual partners, cooperation partners—so it was important, as you are saying, to get along with them and to give in to peer pressure. Because it was always around the same time that that happened throughout our history, our brains have taken that in, and now they reflect that history. Another possibility—both are evolutionarily consistent—would be that throughout our history, there had been times in people's lives in which they found themselves in these situations in which they have to make new friends, or new partners—that can happen when you're 40, it can happen when you're 15, it can happen at any age—so instead of having this maturational period of adolescence, our brains would be equipped to behave in the way that adolescents behave now, in any context in which it's the best thing to do. Could you think of people who are conscripted in the army, or start a new job, or move to a new country, and you find yourself in a situation that is similar to that adolescence, and would you think that they can revert, or they can become more adolescent-like?

BLAKEMORE: I don't know of any research looking at that. It probably exists, but I don't know any data. My guess would be that the brain is plastic throughout life. You can change, you can revert, you can behave differently in different contexts, but the large amount of brain development going on during adolescence, although it's protracted in some areas and continues right throughout the 20s and 30s, it is stabilizing around then. I suppose if you're going to attribute these changes in behavior during adolescence to changes in the brain structure and function, then those changes are not going to be as profound in adults, even if you find yourself in a situation which might demand that.

MERCIER: Could it be an artifact of the fact that most of the adults we've scanned happened to have stable lives? If you look only at adults who need to change plans often because they're moving often, or they don't have a stable partner, maybe you find that their brain keeps being adolescent-like for longer.

JENNIFER JACQUET: I have the same inverse question, then. Imagine a 17-year-old who has been a kidnapped bride, and now has two children, and isn't going through the same peer pressure. Would their brain show the level of stability then? Are there cross-cultural studies?

BLAKEMORE: That's an excellent question. Adolescence is often defined as the period of life between puberty and the end of adolescence is defined at the age at which you attain a stable, independent role in society. With that definition—which I like, a lot of people use it—that includes variations between cultures. In our culture it's normal to be an adolescent—using that definition—right into the 20s, even in the 30s we may not feel particularly stable and independent yet. In other cultures, like the cultures you are mentioning, kids are expected to become financially independent, get jobs, earn money as soon as they go through sexual maturity, or even before. Girls are expected to get married, have babies as soon as they can. Some people have argued that adolescence doesn't exist; it's a Western invention about 100 years ago.

There are three reasons why that's not a completely watertight argument. Firstly, if you look at cross-cultural studies you can see, even in cultures that vastly differ, increases in risk taking, and increases in peer influence, and self-consciousness in those different cultures. Secondly, there are studies in animals showing that even animals undergo a period of heightened risk taking and heightened socialization during their adolescence, post-puberty. If you look at historical descriptions of adolescents, even from thousands of years ago, or in Shakespeare 400 years ago, you see similar descriptions of this age group as the way we stereotypically describe today as taking risks, making bad decisions, and being particularly influenced by their peers.

I'm not saying that culture does not influence the development of the brain, of course it does. Not very much is known about that at all—how brain development looks in these different cultures—but I'm sure it will be subtly different, and people will measure that when they start these studies. There is a lot of overlap between cultures.

MICHAEL MCCULLOUGH: I've been thinking a lot about changes of state in biological systems. Something like the skull, there are so many functions contained in it and yet it has to grow from a small size to something much larger over time. To get all of those functions to coordinate, to create a change from the infant skull to an older skull, the genetic architecture that enables that to happen without devastating trauma is mindboggling. When it doesn't happen, it is truly devastating.

The same with something like sleep, going from a waking state to a sleeping state. We think of it as just unplugging the cord, right? It turns out, for people with chronic sleep disorders who cannot fall asleep there are many things that have to go right. We usually take for granted that they will go right. Or even the assignment of the primary sex characteristics during development, so many things have to go right.

I'm thinking about something like schizophrenia, which is different from lots of other mental disorders inasmuch as it is a truly devastating disease, where evidently, many things that have to go right, at least one of them hasn't gone right. If that premise is true, why does it take the shape it does? It's a menu you pick your features from, but why those features and not others? How do those features so reliably come out of a developmental failure—failure in a change of state—that leads most children to another social way of being in the world? Why does this generic failure, if everything would go right, lead one way, and what does that tell us about what the actual target is normatively?

BLAKEMORE: It's an interesting question, and that's what I was interested in when I was studying schizophrenia back in the day. I was interested in delusions of control and auditory hallucinations. This is where patients think that their movements are being controlled by someone else, or a machine, or they're hearing voices inside their head. My question wasn't why, it was how come that doesn't happen for all of us? When I move and pick this up, how do I know that's my own movement, and that was a movement caused by my intention? That's amazing. How does that happen? Why doesn't that go wrong in all of us? That's exactly the question I was interested in, but not from a development point of view at that point, from a phenomenological point of view. Also, from a mechanistic point of view, how do we achieve that?

All these things are on a spectrum, and yes, people with schizophrenia, there is a clear cut set of symptoms, and it's severe, but each one of those symptoms, most of us will have experienced to a tiny extent at some point. You only have to look at the effects of psychotropic drugs to just push the brain over into temporary paranoia or hallucinations to see it's all a fragile state, and it's not a black and white qualitative difference. It's a quantitative difference. All of these things are on a spectrum, and you can measure that with things like schizotypy questionnaires. The question is why are some people pushed over the edge into this, sometimes permanent, situation of constantly experiencing delusions and hallucinations?

MCCULLOUGH: Or command hallucinations, for example. It's so tempting to try to draw a parallel between those and where the normal, typical pattern of development should take you, which is appropriate approval and respect from your peers. Having people that you can influence, and that can influence you in ways that are going to be adaptive through the rest of adulthood.

BLAKEMORE: One interesting theory of adolescence, back in the '50s, by Peter Elkind, was this idea that teenagers have an imaginary audience. They think they're being watched and judged by other people much more than they are. That is similar to the state that people with schizophrenia describe how their life feels, that they're constantly being watched and observed. I'm not saying there is an overlap or there is a similarity, but there is something that is similar, and the question is how do most teenagers not go over the edge into paranoia?

Lawrence Ian Reed: "The Face Of Emotion"

What can we tell from the face? There're mixed data, but some show a pretty strong coherence between what is felt and what's expressed on the face. Happiness, sadness, disgust, contempt, fear, anger, all have prototypic or characteristic facial expressions. In addition to that, you can tell whether two emotions are blended together. You can tell the difference between surprise and happiness, and surprise and anger, or surprise and sadness. You can also tell the strength of an emotion. There seems to be a relationship between the strength of the emotion and the strength of the contraction of the associated facial muscles.

LAWRENCE IAN REED is a Visiting Assistant Professor of Psychology, Skidmore College.

THE FACE OF EMOTION

My name is Lawrence Ian Reed. I'm a Clinical and Evolutionary Psychologist over at Skidmore College. Today I want to talk about facial expression of emotion, and a question that's been gnawing at me for probably six or seven years. We've got some answers, and I'm excited to talk to you guys about what they are.

The first questions that I asked about facial expression were "how" questions: How do our facial expressions change when we're feeling depressed or when we've got bipolar disorder, or when we're being deceptive? I don't ask those questions any more for a couple reasons. One is that the questions I'm asking now are much more interesting. The other reason is that I felt satisfied with a lot of the answers. I'm going to review some of those questions and talk about how they led up to the questions that I'm asking now, and we'll see what you guys think about what I have to say.

What can we tell from the face? There're mixed data, but some show a pretty strong coherence between what is felt and what's expressed on the face. Happiness, sadness, disgust, contempt, fear, anger, all have prototypic or characteristic facial expressions. In addition to that, you can tell whether two emotions are blended together. You can tell the difference between surprise and happiness, and surprise and anger, or surprise and sadness. You can also tell the strength of an emotion.

There seems to be a relationship between the strength of the emotion and the strength of the contraction of the associated facial muscles.

The thing that I find fascinating about facial expression of emotion is our limitation in the ability to control them. They come in two separate forms. The first is in the distinction between spontaneously and deliberately induced facial expressions of emotion. What I mean by spontaneously is, those facial expressions that happen as a result of some stimulus intended to elicit emotion of some valance. Deliberately induced facial expressions are those that happen as a result of a directed facial action task—if someone says, "I want you to smile on command, or frown on command."

It turns out that the spontaneously induced and deliberately induced facial actions emanate from separate upper motor neuron pathways. The deliberately induced facial expressions come from the cortical motor strip, whereas the spontaneously expressed emotions emanate from the phylogenetically older, extrapyramidal motor system. What this means is that if you’ve got damage to the cortical motor strip, you lose the ability to make facial expressions on command but retain the ability to make them in response to some emotion-eliciting stimulus. On the other hand, if you’ve got damage to the extrapyramidal motor system, you lose the ability to respond emotionally to a stimulus, but you’ll be able to make a facial expression on command. That, I found very fascinating.

There’s another limitation that’s been found by Ekman’s group. What he has found is that a certain number of facial expressions can be thought of as reliable. Reliable facial muscles have two properties. Less than 10 to 15 percent of individuals can do two things with these reliable muscles. The first thing is to create them in the absence of that associated emotion. The second thing is to stop them from happening in the presence of the associated emotion.

Those two findings are the ones that made me jettison all the previous questions that I was asking about how we make facial expressions of emotion, because this struck me as extremely odd. Why would we express something as private and as subjective as our motivational and emotional states somewhere as conspicuous as the face? Wouldn't the smell of fear just give our opponents some advantage? Wouldn't a poker face be better? I went from asking these "how" questions to these "why" questions. There are three classes of answers to these questions. I’ll get to the one that I’ve been focusing on recently, last.

The first class comes from a physiological function. This comes from Darwin's *Expression of Emotion in Man and Animals*. He stated that facial expressions have a physiological function. That is, they allow us to respond to recurrent stimuli that happen within our environment. For example, the widened eyes that you see in fear expressions allow us to get more information from our periphery when we need that; the scrunching of the nose and the protrusion of the tongue allow us to expel noxious stimuli from our nose and mouth. There's good evidence to corroborate this suggestion.

None of these classes are mutually exclusive, not by any means. The second class is a communicative class of functions. Darwin didn’t talk about this communicative class of functioning for a very specific reason. At the time that he was writing *Expression of Emotion in Man and Animals*, a lot of creationist folks, including Sir Charles Bell, thought that facial expressions were a God-given way of communicating emotion. Darwin, for obvious reasons, wanted to distance himself from those conceptualizations because he had the whole evolution by natural selection thing that he wanted to propagate.

Darwin didn’t talk about any communicative function of facial expressions at all, and that’s why these questions have lagged a bit behind. To his credit, he did imply a communicative function with his antithesis principle. The antithesis principle states that expressions have the form that they do because that form opposes the form of opposite expressions. For example, the upward lip corner turning in smiling differs from the downward turning in sadness, the outward turning in fear, and the dimpling action in contempt. That would be Darwin’s explanation for something like a smile.

That's the second class—the communicative class of functioning. This makes a lot of intuitive sense, but it still didn’t answer the question that any adaptationist would ask: What is the benefit to the signaler of expressing something on their face? The benefit to the receiver seems more clear. It allows the receiver to predict what that person’s emotional state is and what they’re going to do next. The benefit to the signaler wasn’t so clear to me.

That takes me to the third class of functioning, which is related to the communicative functioning, and what I’m going to call the commitment function of facial expression. What I’m about to say is very ubiquitous; it spans economics, biology, psychology, and it’s been put forth by Robert Frank and Jack Hirshleifer, and previously Thomas Schelling in the Sixties. The idea is that we’ve got sets of emotions that Adam Smith called moral sentiments. These moral sentiments function by competing with calculations that stem from rational self-interest.

If you take someone that has strong guilt feelings or is capable of strong guilt feelings, this person’s not going to cheat on their spouse, not because they’re afraid of getting caught, but because they don’t want to; they know that the guilt feelings are going to be aversive and it’s going to outweigh whatever benefit they’d get from cheating on their spouse. Similarly, take someone who is capable of great acts of revenge. They’re not going to need a formal contract to get revenge, they’re going to get revenge because they want to, because it’s going to feel good and that good feeling is going to outweigh the negative consequences of seeking revenge.

This is still problematic because the good feelings that we get when we remain faithful to our spouse and the good feelings that we get when we exact revenge are in and of themselves very real rewards. But we are living in a material world of material payoffs and for these incentives to be viable, they must have incurred some material benefit. That is to say, it’s no good for me to say, "Look, you should have trusted me when we could have cooperated before, because I would have," and it's no good for me to say, "Look, you should not have harmed me a while ago," when the harm is already done. The idea is that for these incentives to give material payoffs, they need to be honestly communicated beforehand.

From this view, this answers the question that has been gnawing at me of what’s the benefit to the signaler? The benefit to the signaler is that they can now honestly express threats and promises. From this viewpoint, what I categorized before as limitations in our ability to control facial expressions now becomes a necessity—a defining feature of the conceptualization of facial expressions. This is the idea that I’ve taken, that I’ve tried to find empirical evidence for. I’ve done a couple things so far, so I’ll talk to you about those just briefly, and then I’ll talk to you about what I’ve been doing very recently—the data that I’ve collected just last week—and some of the questions that I have moving forward.

The idea is that facial expressions should enhance the credibility of our threats and promises. In a study we had out a couple years ago, we had people play a one-shot prisoner’s dilemma with an acquaintance period beforehand. What we found is that most people gave verbal promises to cooperate during this prisoner’s dilemma, probably 75 to 80 percent of those individuals. Some of those folks promised with genuine, difficult to fake smiles—Duchenne smiles. The verdict’s out on how difficult it is to fake a Duchenne smile. I could do it right now, but you could argue that it’s a costly signal.

What we found was a couple things. Those individuals that had their verbal promises paired with Duchenne smiles were more likely to cooperate with their partners. Furthermore, their partners predicted that they would be more likely to cooperate with them, which suggests that there’s some encoding of these promises in facial expression among signalers, and some decoding of these promises among receivers.

The other side of that coin is looking at threats and promises. What we did next is we had people play an ultimatum game. An ultimatum game, I’m sure a lot of you are familiar with, but quickly, you’ve got a proposer and a responder. The proposer has control of how much he’s going to divide the pie between the two of those individuals. The responder can accept that proposal, in which case both individuals get what the proposer decides. Or the responder can reject it, in which both individuals get nothing.

What we did is we had threats to reject incredible offers—high offers, offers that responders wouldn’t normally reject. We paired them with either neutral or angry, difficult to fake angry facial expressions. What we found was that those threats that were paired with difficult to fake anger expressions resulted in high proposer offers. What we think that suggests is that facial expressions, specifically difficult to fake anger expressions, add credibility to our threats.

What we want to do now is to see what other facial expressions might have this commitment function. I don’t think that all of them would, but all the facial expressions would have some communicative function. In addition to threats and promises, we’re also looking at requests. Again, this is the data that we’ve just got this week, and we’re still writing it up. The idea, and some people have positioned this, particularly Ed Hagen, and Randy Nesse, that sadness functions as a bargaining tool that would allow us to extort—I’m hesitant to use that word because it has a negative connotation, but I don’t mean it that way—resources from con specifics that would have some stake in our fitness.

We just did an MTurk sample in which we had two people playing a two-person threshold public goods game. We had one of the individuals state a request saying, "I can’t contribute my fair share in order for us to get the threshold for the payout. Can you give me more?" What we found was that this verbal statement, when paired with difficult to fake sadness expressions—the reliable facial muscles of the triangularis, which lowers the lip corner and the medial frontalis, which gives you that triangulation of the brows in worry and sadness expressions. Think of the Woody Allen expression. He’s one of those 10 to 15 percent that can do that at will. But now that he can it means absolutely nothing, right? You don’t infer any concern from him. That goes back to the point that Robert Frank made a long time ago that if the facial expressions that are characteristic of specific emotions were easily and readably fake, they would no longer be characteristic of those specific emotions.

We've looked at threats. We’ve looked at promises. Now we feel we’ve got some preliminary data that there’s a communicative function for sadness. The other emotions, I’m not so sure that they’ve got a commitment function, but some of the things that we’d like to do are to look at disgust. Disgust is hypothesized that it functions physiologically for the signaler, but the receiver benefits because it lets them know, "Look, whatever this person is exposed to, I might be exposed to as well."

Often times you’d imagine the EEA, in the Pleistocene era, that people are sharing meals. Such an adaptation that will allow us to figure out whether someone might be infected would be very beneficial. One of the things I’m thinking about doing—you’ll have to forgive me but this is in its planning stages—is of having some shared meal between participants. I was thinking about having Jelly Belly jellybeans, one of the nasty flavors put in there, and see how the recipient or the receiver responds when the person eats that disgusting one.

What I would anticipate is that there might be some contagion effects. If you saw an individual with disgust, you’d find that in the other participant as well. You might find the same with something like surprise. Surprise, again, has this physiological benefit to the signaler, but to the receiver, it tells them, okay, this person is watching out for something dangerous, and I am in their proximity. Maybe there’s something I’ve got to watch out for as well. That might be a way to look at some of the surprise expressions.

Finally, there’s this idea that facial expressions, insofar as though they are self-conscious, might form as a way of communicating common knowledge between other individuals. What I mean by common knowledge is the knowledge that *I know that you know that I know that you know* something, ad infinitum. It turns out that if you have just, *I think that you think that I think that I think*, that the recursive nature explodes that doubt into something that’s very different than common knowledge. There are some shortcuts to it, like direct eye gaze. I think Hugo was talking about this earlier. When you look at someone directly in the eye, it’s very off-putting sometimes. It’s often used in sexual come-ons and threats. But there’s no denying that you’ve made eye contact with that individual.

Another one could be tearing. Tearing is hypothesized to be a handicap that might serve as a cue for common knowledge because the other individual can be absolutely certain that you’re crying. They can be absolutely certain that you know that they're crying because it impairs your vision in a specific way. Some people have argued, Provine specifically, that tearing functions as a handicap to make it difficult for us to make attacks on other individuals and defend because of the blurred vision. Also laughing. Laughing is an intuitively strange thing to make that response in response to sometimes seemingly innocuous comments, but that could also be a signal for common knowledge. Any of the motions could be insofar as the only requisite is that they are self-conscious and the other person knows that they are self-conscious. Things like crying, laughing, direct eye contact, blushing are good candidates for that.

That is the question. Those are the putative answers. I'm still trying to find more evidence for this commitment function of facial expression, and asking questions about why we express emotions on our face. It's very satisfying, because sometimes you have questions, and then they never get answered. But here I feel as though we're making some headway, which is extremely satisfying.

THE REALITY CLUB

HUGO MERCIER: This is fascinating work, but there is one thing on commitment and emotions that some people get slightly confused about. I'm not saying you're doing this, but some people are. It's the idea that as soon as emotions are not under voluntary control, then that solves the evolutionary problem of how they can be stable. We know that's not the case because, evolutionarily, the question that's relevant is not whether they are under voluntary control or not but whether we could evolve in such a way that they could be expressed in different contexts. Once you have shown that someone who displays anger can reliably express this in behavior, you're still only going to have half of the picture.

If you want to know how the communication of any given emotion remains stable, not in the sense of why people cannot take advantage of this in terms of controlling their own emotions, but why we didn't evolve in such a way that we could take advantage of that. One of the solutions we put forward in the paper with Guillaume Dezecache and Thom Scott-Phillips is that the receivers of the emotion are vigilant towards the emotions that aren't being displayed. If you think the source isn't as worthy, or the emotion doesn't feel right in the context in which it is being expressed, you're not going to respond to it in the way that the sender intends it to. I'm curious if you have done any experiments, or can you think of experiments in which someone displays an emotion and they don't act in the way that is expected of them? Then the next time they display the emotion they’re not taken as seriously as they were, which would maintain the stability of the communication.

REED: That’s a good point. I would expect that there would be some limits on our ability to control our facial expressions, and then some limits on receivers’ ability to decipher how accurate they are, based on exactly what you said, because there’s going to be an arms race between the two. Why haven’t we evolved to completely lie? There are limitations, but the limitations need to be either that everyone has a limited ability, or that only certain people can do it. As long as it’s a heuristic that works more than half of the time, then we’ll have evolved for that. Then we’ll have adaptations for it.

There are studies in which they have displayed film clips, or stimuli intended to elicit various emotions, and then instructed people either to display a different emotion or not display any emotion at all just to see how good people are at controlling them. Forgive me, I don’t know, it’s mixed for all the emotions, but it does say that we have some limited ability to control them.

Then the other part of your question is how does that affect receivers if they know people have the ability to control them?

MERCIER: Even if people had no ability to control emotions at all. If you look at signaling in animals, at least the vast majority of it is not under any voluntary control, we still have the question of how can it remain stable. Why don’t animals evolve to send signals when it’s advantageous to them and not to the receivers? The emotional signals could be completely impossible to fake, and that still would tell us nothing about how they remain stable. You still need the receiver to be able to tell when she should respond to the signal. In a way then you could nearly see how controllable the signals are as being nearly orthogonal to the evolutionary question. Answering the controllability question doesn’t buy you the evolutionary stability that you’re looking for.

FIERY CUSHMAN: I was going to jump in and say that one way to make it not quite orthogonal is to turn the question around, say, given that we’re going to have to construct an honest signal, would it be a viable approach to put it under voluntary control? It's interesting to think that if you were to hand the keys to the car over to voluntary control, voluntary control would surely crash the car so quickly that it wouldn’t be a useful, honest signal.

MERCIER: At least the vast majority of our communication is under voluntary control and it’s mostly honest. So it can be done.

CUSHMAN: Fair enough. Yes.

DAVID RAND: But not in situations where it conflicts it, where conflicting interests exist. Right?

MERCIER: If it’s purely a non-zero sum game you’re playing then there can be no communication, whether it's emotional, voluntary, or whatever. The format of the communication is irrelevant to the evolutionary question. It's not completely irrelevant, but it’s not a necessary condition.

RAND: It seems to me that a point is “What keeps the signal honest?” One answer, which you were suggesting but I think you could test in a straightforward way, is that when someone gives us a signal that they will be honest and then they betray you, you’re more mad than if they had just betrayed you without giving the honest signal.

MERCIER: I’m going to do that. If nobody could steal that, it would be great.

JOSHUA KNOBE: I was thinking about the evolution of involuntariness. It seems like you’re saying that there’s some adaptational advantage that we have from having it not be under voluntary control —that we can send these honest signals. So, if you brought in pairs of people, some who have it under voluntary control and some who don’t, you should expect the people who don’t have it under voluntary control to show certain advantages. That is to say, if I can’t fake sadness but Hugo can, then with people who know me and people who know Hugo, I should be able to get what I want in some cases where he can’t.

It would be an easy way to test this hypothesis with this advantage because we can actually observe people who are doing it.

REED: I completely agree with you, and I am stealing that idea.

RAND: Specifically it would be that the person that can fake it would be at an advantage when interacting with strangers, but the person that can’t fake it would be at an advantage when interacting with friends.

DAVID PIZARRO: This is why reputation is generally spoken of in this context. Robert Frank goes way out of his way to say reputational effects—like communicating, "Hugo is a horrible person"—tend to dominate. There is something about signaling in emotions that is different from the signaling that we often see in the animal world, which is, the emotional signaling is directed at one person. You being angry when you’re talking to me, means something more than just you being angry. Both of those are showing the same facial expression. Is there a way in which humans are better at decoding angry? Angry directed at me means you’re going to screw me over. Angry directed at everybody else means: "This is my buddy. Don’t mess with us." That's just how complex we are. The signal itself isn’t telling everybody the same thing.

REED: That’s another wonderful, empirical question that you could do with eye gaze or something along those lines.

PIZARRO: I feel like somebody has done angry expressions looking at and looking away from, and people have very different reactions to them.

LAURIE SANTOS: Yes, this is true even in monkeys where you get different neural responses for direct gaze facial expressions versus sideways gaze facial expressions. The idea is the same.

Molly Crockett: "The Neuroscience of Moral Decision Making"

Imagine we could develop a precise drug that amplifies people's aversion to harming others; on this drug you won't hurt a fly, everyone taking it becomes like Buddhist monks. Who should take this drug? Only convicted criminals—people who have committed violent crimes? Should we put it in the water supply? These are normative questions. These are questions about what should be done. I feel grossly unprepared to answer these questions with the training that I have, but these are important conversations to have between disciplines. Psychologists and neuroscientists need to be talking to philosophers about this. These are conversations that we need to have because we don't want to get to the point where we have the technology but haven't had this conversation, because then terrible things could happen.

MOLLY CROCKETT is Associate Professor, Department of Experimental Psychology, University of Oxford; Wellcome Trust Postdoctoral Fellow, Wellcome Trust Centre for Neuroimaging.

THE NEUROSCIENCE OF MORAL DECISION MAKING

I'm a neuroscientist at the University of Oxford in the UK. I'm interested in decision making, specifically decisions that involve tradeoffs—for example, tradeoffs between my own self-interest and the interests of other people, or tradeoffs between my present desires and my future goals.

One thing that's always fascinated me, specifically about human decision making, is the fact that we have multiple conflicting motives in our decision process. And not only do we have these forces pulling us in different directions, but we can reflect on this fact. We can witness the tug of war that happens when we're trying to make a difficult decision. One thing that is great about our ability to reflect on this process is that it suggests that we can intervene somehow in our decisions. We can make better decisions—more self-controlled decisions, or more moral decisions.

The reason I've become interested in the neuroscience of decision making is because I have this sense that pulling apart the different moving parts of this process and looking under the hood will give us clues about where we might be able to intervene and shape our own decisions.

One case study for this is moral decision making. When we can see that there's a selfish option and we can see that there's an altruistic or a cooperative option, we can reason our way through the decision, but there are also gut feelings about what's right and what's wrong. I've studied the neurobiology of moral decision making, specifically how different chemicals in our brains—neuromodulators—can shape the process of making moral decisions and push us one way or another when we're reasoning and deciding.

Neuromodulators are chemicals in the brain. There are a bunch of different neuromodulator systems that serve different functions. Events out in the world activate these systems and then they perfuse into different regions of the brain and influence the way that information is processed in those regions. All of you have experience with neuromodulators. Some of you are drinking cups of coffee right now. Many of you probably had wine with dinner last night. Maybe some of you have other experiences that are a little more interesting.

But you don't need to take drugs or alcohol to influence your neurochemistry. You can also influence your neurochemistry through natural events: Stress influences your neurochemistry, sex, exercise, changing your diet. There are all these things out in the world that feed into our brains through these chemical systems. I've become interested in studying if we change these chemicals in the lab, can we cause changes in people's behavior and their decision making?

One thing to keep in mind about the effects of these different chemicals on our behavior is that the effects here are subtle. The effect sizes are really small. This has two consequences for doing research in this area. The first is because the effect sizes are so small, the published literature on this is likely to be underpowered. There are probably a lot of false positives out there. We heard earlier that there is a lot of thought on this in science, not just in psychology but in all of science about how we can do better powered experiments, and how we can create a set of data that will tell us what's going on.

The other thing—and this is what I've been interested in—is because the effects of neuromodulators are so subtle, we need precise measures in the lab of the behaviors and decision processes that we're interested in. It's only with precise measures that we're going to be able to pick up these subtle effects of brain chemistry, which maybe at the individual level aren't going to make a dramatic difference in someone's personality, but at the aggregate level, in collective behaviors like cooperation and public goods problems, these might become important on a global scale.

How can we measure moral decision making in the lab in a precise way, and also in a way that we can agree is actually moral? This is an important point. One big challenge in this area is there's a lot of disagreement about what constitutes a moral behavior. What is moral? We heard earlier about cooperation— maybe some people think that's a moral decision but maybe other people don't. That's a real issue for getting people to cooperate.

First we have to pick a behavior that we can all agree is moral, and secondly we need to measure it in a way that tells us something about the mechanism. We want to have these rich sets of data that tell us about these different moving parts—these different pieces of the puzzle—and then we can see how they map onto different parts of the brain and different chemical systems.

What I'm going to do over the next 20 minutes is take you through my thought process over the past several years. I tried a bunch of different ways of measuring the effects of neurochemistry on what at one point I think is moral decision making, but then turns out maybe is not the best way to measure morality. And I'll show you how I tried to zoom in on more advanced and sophisticated ways of measuring the cognitions and emotions that we care about in this context.

When I started this work several years ago, I was interested in punishment and economic games that you can use to measure punishment—if someone treats you unfairly then you can spend a bit of money to take money away from them. I was interested specifically in the effects of a brain chemical called serotonin on punishment. The issues that I'll talk about here aren't specific to serotonin but apply to this bigger question of how can we change moral decision making.

When I started this work the prevailing view about punishment was that punishment was a moral behavior—a moralistic or altruistic punishment where you're suffering a cost to enforce a social norm for the greater good. It turned out that serotonin was an interesting chemical to be studying in this context because serotonin has this long tradition of being associated with prosocial behavior. If you boost serotonin function, this makes people more prosocial. If you deplete or impair serotonin function, this makes people antisocial. If you go by the logic that punishment is a moral thing to do, then if you enhance serotonin, that should increase punishment. What we actually see in the lab is the opposite effect. If you increase serotonin people punish less, and if you decrease serotonin people punish more.

That throws a bit of a spanner in the works of the idea that punishment is this exclusively prosocially minded act. And this makes sense if you just introspect into the kinds of motivations that you go through if someone treats you unfairly and you punish them. I don't know about you, but when that happens to me I'm not thinking about enforcing a social norm or the greater good, I just want that guy to suffer; I just want him to feel bad because he made me feel bad.

The neurochemistry adds an interesting layer to this bigger question of whether punishment is prosocially motivated, because in some ways it's a more objective way to look at it. Serotonin doesn't have a research agenda; it's just a chemical. We had all this data and we started thinking differently about the motivations of so-called altruistic punishment. That inspired a purely behavioral study where we give people the opportunity to punish those who behave unfairly towards them, but we do it in two conditions. One is a standard case where someone behaves unfairly to someone else and then that person can punish them. Everyone has full information, and the guy who's unfair knows that he's being punished.

Then we added another condition, where we give people the opportunity to punish in secret— hidden punishment. You can punish someone without them knowing that they've been punished. They still suffer a loss financially, but because we obscure the size of the stake, the guy who's being punished doesn't know he's being punished. The punisher gets the satisfaction of knowing that the bad guy is getting less money, but there's no social norm being enforced.

What we find is that people still punish a lot in the hidden punishment condition. Even though people will punish a little bit more when they know the guy who's being punished will know that he's being punished -- people do care about norm enforcement – a lot of punishment behavior can be explained by a desire for the norm violator to have a lower payoff in the end. This suggests that punishment is potentially a bad way to study morality because the motivations behind punishment are, in large part, spiteful.

Another set of methods that we've used to look at morality in the lab and how it's shaped by neurochemistry is trolley problems—the bread and butter of moral psychology research. These are hypothetical scenarios where people are asked whether it's morally acceptable to harm one person in order to save many others.

We do find effects of neuromodulators on these scenarios and they're very interesting in their own right. But I've found this tool unsatisfying for the question that I'm interested in, which is: How do people make moral decisions with real consequences in real time, rather than in some hypothetical situation? I'm equally unsatisfied with economic games as a tool for studying moral decision making because it's not clear that there's a salient moral norm in something like cooperation in a public goods game, or charitable giving in a dictator game. It's not clear that people feel guilty if they choose the selfish option in these cases.

After all this I've gone back to the drawing board and thought about what is the essence of morality? There's been some work on this in recent years. One wonderful paper by Kurt Gray, Liane Young, and Adam Waytz argues that the essence of morality is harm, specifically intentional interpersonal harm—an agent harming a patient. Of course morality is more than this; absolutely morality is more than this. It will be hard to find a moral code that doesn't include some prohibition against harming someone else unless you have a good reason.

What I wanted to do was create a measure in the lab that can precisely quantify how much people dislike causing interpersonal harms. What we came up with was getting people to make tradeoffs between personal profits—money—and pain in the form of electric shocks that are given to another person.

What we can do with this method is calculate, in monetary terms, how much people dislike harming others. And we can fit computational models to their decision process that give us a rich picture of how people make these decisions -- not just how much harm they're willing to deliver or not -- but what is the precise value they place on the harm of others relative to, for example, harm to themselves? What is the relative certainty or uncertainty with which they're making those decisions? How noisy are their choices? If we're dealing with monetary gains or losses, how does loss aversion factor into this?

We can get a more detailed picture of the data and of the decision process from using methods like these, which are largely inspired by work on non-social decision making and computational neuroscience where a lot of progress has been made in recent years. For example, in foraging environments how do people decide whether to go left or right when there are fluctuating reward contingencies in the environment?

What we're doing is importing those methods to the study of moral decision making and a lot of interesting stuff has come out of it. As you might expect there is individual variation in decision making in this setting. Some people care about avoiding harm to others and other people are like, "Just show me the money, I don't care about the other person." I even had one subject who was almost certainly on the psychopathy scale. When I explained to him what he had to do he said, "Wait, you're going to pay me to shock people? This is the best experiment ever!" Whereas other people are uncomfortable and are even distressed by this. This is capturing something real about moral decision making.

One thing that we're seeing in the data is that people who seem to be more averse to harming others are slower when they're making their decisions. This is an interesting contrast to Dave's work where the more prosocial people are faster. Of course there are issues that we need to work out about correlation versus causation in response times and decision making, but there are some questions here in thinking about the differences between a harm context and helping context. It may be that the heuristics that play out in a helping context come from learning about what is good and latch onto neurobiological systems that approach rewards and get invigorated when there are awards around, in contrast to neurobiological systems that avoid punishments and slow down or freeze when there are punishments around.

In the context of tradeoffs between profit for myself and pain for someone else, it makes sense that people who are maximizing the profit for themselves are going to be faster because if you're considering the harm to someone else, that's an extra computational step you have to take. If you're going to factor in someone else's suffering—the negative externality of your decisions—you have to do that computation and that's going to take a little time.

In this broader question of the time course of moral decision making, there might be a sweet spot where on the one hand you have an established heuristic of helping that's going to make you faster, but at the same time considering others is also a step that requires some extra processing. This makes sense.

When I was developing this work in London I was walking down the street one day checking my phone, as we all do, and this kid on a bike in a hoodie came by and tried to steal my phone. He luckily didn't get it, it just crashed to the floor -- he was an incompetent thief. In thinking about what his thought process was during that time, he wasn't thinking about me at all. He had his eye on the prize. He had his eye on the phone, he was thinking about his reward. He wasn't thinking about the suffering that I would feel if I lost my phone. That's a broader question to think about in terms of the input of mentalizing to moral decision making.

Another observation is that people who are nicer in this setting seem to be more uncertain in their decision making. If you look at the parameters that describe uncertainty, you can see that people who are nicer seem to be more noisy around their indifference point. They waver more in these difficult decisions.

So I've been thinking about uncertainty and its relationship to altruism and social decision making, more generally. One potentially fruitful line of thought is that social decisions—decisions that affect other people—always have this inherent element of uncertainty. Even if we're a good mentalizer, even if we're the best possible mentalizer, we're never going to fully know what it is like to be someone else and how another person is going to experience the effects of our actions on them.

One thing that it might make sense to do if we want to co-exist peacefully with others is we simulate how our behavior is going to effect others, but we err on the side of caution. We don't want to impose an unbearable cost on someone else so we think, "Well, I might dislike this outcome a certain amount but maybe my interaction partner is going to dislike it a little more so I'm just going to add a little extra safety—a margin of error—that's going to move me in the prosocial direction." We're seeing this in the context of pain but this could apply to any cost—risk or time cost.

Imagine that you have a friend who is trying to decide between two medical procedures. One procedure produces the most desirable outcome, but it also has a high complication or a high mortality rate. Another procedure doesn't achieve as good of an outcome but it's much safer. Suppose your friend says to you, "I want you to choose which procedure I'm going to have. I want you to choose for me." First of all, most of us would be very uncomfortable making that decision for someone else. Second, my intuition is that I would definitely go for the safer option because if something bad happened in the risky decision, I would feel terrible.

This idea that we can't access directly someone else's utility function is a rather old idea and it goes back to the 1950s with the work of John Harsanyi, who did some work on what he called interpersonal utility comparisons. How do you compare one person's utility to another person's utility? This problem is important, particularly in utilitarian ethics, because if you want to maximize the greatest good for the greatest number, you have to have some way of measuring the greatest good for each of those numbers.

The challenge of doing this was recognized by the father of utilitarianism, Jeremy Bentham, who said, "'Tis vain to talk of adding quantities which after the addition will continue to be as distinct as they were before; one man's happiness will never be another man's happiness: a gain to one man is no gain to another: you might as well pretend to add 20 apples to 20 pears."

This problem has still not been solved. Harsanyi has done a lot of great work on this but what he ended up with—his final solution—was still an approximation that assumes that people have perfect empathy, which we know is not the case. There's still room in this area for exploration.

The other thing about uncertainty is that, on one hand it could lead us towards prosocial behavior, but on the other hand there's evidence that uncertainty about outcomes and about how other people react to those outcomes can license selfish behavior. Uncertainty can also be exploited for personal gain for self-serving interests.

Imagine you're the CEO of a company. You're trying to decide whether to lay off some workers in order to increase shareholder value. If you want to do the cost benefit analysis, you have to calculate what's the negative utility for the workers of losing their jobs and how does that compare to the positive utility of the shareholders for getting these profits? Because you can't directly access how the workers are going to feel, and how the shareholders are going to feel, there's space for self-interest to creep in, particularly if there are personal incentives to push you one direction or the other.

There's some nice work that has been done on this by Roberto Weber and Jason Dana who have shown that if you put people in situations where outcomes are ambiguous, people will use this to their advantage to make the selfish decision but still preserve their self-image as being a moral person. This is going to be an important question to address. When does uncertainty lead to prosocial behavior because we don't want to impose an unbearable cost on someone else? And when does it lead to selfish behavior because we can convince ourselves that it's not going to be that bad?

These are things we want to be able to measure in the lab and to map different brain processes— different neurochemical systems—onto these different parameters that all feed into decisions. We're going to see progress over the next several years because in this non-social computational neuroscience there are smart people who are mapping how basic decisions work. All people like me have to do is import those methods to studying more complex social decisions. There's going to be a lot of low-hanging fruit in this area over the next few years.

Once we figure out how all this works—and I do think it's going to be a while -- I've been misquoted sometimes about saying morality pills are just around the corner, and I assure you that this is not the case. It's going to be a very long time before we're able to intervene in moral behavior and that day may never even come. The reason why this is such a complicated problem is because working out how the brain does this is the easy part. The hard part is what to do with that. This is a philosophical question. If we figure out how all the moving parts work, then the question is should we intervene and if so how should we intervene?

Imagine we could develop a precise drug that amplifies people's aversion to harming others; on this drug you won't hurt a fly, everyone taking it becomes like Buddhist monks. Who should take this drug? Only convicted criminals—people who have committed violent crimes? Should we put it in the water supply? These are normative questions. These are questions about what should be done. I feel grossly unprepared to answer these questions with the training that I have, but these are important conversations to have between disciplines. Psychologists and neuroscientists need to be talking to philosophers about this. These are conversations that we need to have because we don't want to get to the point where we have the technology but haven't had this conversation, because then terrible things could happen.

The last thing that I'll say is it's also interesting to think about the implications of this work, the fact that we can shift around people's morals by giving them drugs. What are the implications of this data for our understanding of what morality is?

There's increasing evidence now that if you give people testosterone or influence their serotonin or oxytocin, this is going to shift the way they make moral decisions. Not in a dramatic way, but in a subtle yet significant way. And because the levels and function of our neuromodulators are changing all the time in response to events in our environment, that means that external circumstances can play a role in what you think is right and what you think is wrong.

Many people may find this to be deeply uncomfortable because we like to think of our morals as being core to who we are and one of the most stable things about us. We like to think of them as being written in stone. If this is not the case, then what are the implications for our understanding of who we are and what we should think about in terms of enforcing norms in society? Maybe you might think the solution is we should just try to make our moral judgments from a neutral stance, like the placebo condition of life. That doesn't exist. Our brain chemistry is shifting all the time so it's this very unsteady ground that we can't find our footing on.

At the end of the day that's how I try to avoid being an arrogant scientist who's like, "I can measure morality in the lab." I have deep respect for the instability of these things and these are conversations that I find deeply fascinating.

THE REALITY CLUB

L.A. PAUL: I had a question about how you want to think about these philosophical issues. Sometimes they get described as autonomy. You said if we could discover some chemical that would improve people's moral capacities, do we put it in the water? The question I have is a little bit related to imaginability. In other words, the guy who tried to steal your phone. The thought was: If he were somehow better able to imagine how I would respond, he would somehow make maybe a better moral judgment. There's an interesting normative versus descriptive question there because on the one hand, it might be easier to justify putting the drug in the water if it made people better at grasping true moral facts.

What if it just made them better at imagining various scenarios so that they acted in a morally better way, but in fact it had no connection at all to reality, it just made their behavior better. It seems like it's important to make that distinction even with the work that you're doing. Namely, are you focusing on how people actually act or are you focusing on the psychological facts? Which one are we prioritizing and which one are we using to justify whatever kinds of policy implications?

CROCKETT: This goes back to the question of do we want to be psychologists or economists if we're confronted with a worldly, all-powerful being. I am falling squarely in the psychologist camp in that it's so important to understand the motivations behind why people do the things they do -- because if you change the context, then people might behave differently. If you're just observing behavior and you don't know why that behavior occurs, then you could make incorrect predictions.

Back to your question, one thought that pops up is it's potentially less controversial to enhance capabilities that people think about as giving them more competence in the world.

PAUL: There's interesting work on organ donors in particular. When people are recruiting possible organ donors and they're looking at the families who have to make the decision, it turns out that that you get better results by encouraging the families of potential donors to imagine that the daughter was killed in a car accident, the recipient of the organ will be 17 and also loves horses. It could just be some dude with a drug problem who's going to get the organ, but the measured results of the donating family are much better if that family engages in this fictitious imagining even though it has no connection at all to the truth. It's not always simple. In other words, the moral questions sometimes come apart from the desired empirical result.

CROCKETT: One way that psychologists and neuroscientists can contribute to this discussion is to be as specific and precise as possible in understanding how to shape motivation versus how to shape choices. I don't have a good answer about the right thing to do in this case, but I agree that it is an important question.

DAVID PIZARRO: I have a good answer. This theme was something that was emerging at the end with Dave's talk about promoting behavior versus understanding the mechanisms. There is—even if you are a psychologist and you have an interest in this—a way in which, in the mechanisms, you could say, "I'm going to take B.F. Skinner's learning approach and say what I care about is essentially the frequency of the behavior. What are the things that I have to do to promote the behavior that I want to promote?"

You can get these nice, manipulated contingencies in the environment between reward and punishment. Does reward work better than punishment?

I want to propose that we have two very good intuitions, one, which should be discarded when we're being social scientists, is what do we want our kids to be like? I want my kid to be good for the right reasons. In other words, I want her to develop a character that I can be proud of and that she can be proud of. I want her to donate to charity not because she's afraid that if she doesn't people will judge her poorly but because she genuinely cares about other people.

When I'm looking at society, and the more and more work that we do that might have implications for society, we should set aside those concerns. That is, we should be comfortable saying that there is one question about what the right reasons are and what the right motivations are in a moral sense. There's another question that should ask from a public policy perspective: what will maximize the welfare of my society? I don't give a rat's ass why people are doing it!

It shouldn't make a difference if you’re doing it because you’re ashamed (like Jennifer might be talking about later): "I want to sign up for the energy program because I will get mocked by my peers," or if you’re doing it because you realize this is a calling that God gave to you—to insert this little temperature reducer during California summers. That "by any means necessary" approach that seems so inhuman to us as individuals is a perfectly appropriate strategy to use when we’re making decisions for the public.

THE REALITY CLUB

CROCKETT: Yes, that makes sense and it's a satisficing approach rather than a maximizing approach. One reason why we care about the first intuition so much is because in the context in which we evolved, which was small group interactions, someone who does a good thing for the right reasons is going to be more reliable and more trustworthy over time than someone who does it for an externally incentivized reason.

PIZARRO: And it may not be true. Right? It may turn out to be wrong.

DAVID RAND: That's right, but I think it’s still true that it’s not just about when you were in a small group—hunter-gatherer—but in general: if you believe something for the right reason, then you’ll do it even if no one is watching. That creates a more socially optimal outcome than if you only do it when someone is watching.

PIZARRO: It’s an empirical question though. I don't know if it’s been answered. For instance, the fear of punishment...

RAND: We have data, of a flavor. If you look at people that cooperate in repeated prisoners dilemmas, they’re no more or less likely to cooperate in one shot, or they’re no more likely to give in a dictator game. When the rule is in place, everybody cooperates regardless of whether they’re selfish or not. When no incentive is there, selfish people go back to being selfish.

SARAH-JAYNE BLAKEMORE: There’s also data that in newsagents in the UK, where sometimes you can take a newspaper and put money in the slot, and if you put a couple of eyes above the money slot, people are more likely to pay their dues than if you don’t put any eyes there.

PIZARRO: That’s certainly not acting for the right reason. That can’t be the right reason.

RAND: You were bringing up the issue of thinking about the consequences for yourself versus the other person. When we're thinking about how these decisions get made, there are two stages that are distinct but get lumped together a lot conceptually and measurement-wise. You have to understand what the options are, and then once you know what the options are, you have to choose which one you prefer. It seems to me that automatic versus deliberative processing has opposite roles in those two domains. Obviously to understand the problem you have to think about it. If you're selfish, you don't need to spend time to think about the decision because it's obviously what to do. We try to separate those things by explaining the decision beforehand when you're not constrained. Then when it comes time to make the decision, you put people under time pressure.

CROCKETT: That can explain what's going on and that's a good point because these ideas about uncertainty and moral wiggle room, those are going to play the biggest role in the first part—in the construing of the problem. Is this a moral decision or is this not a moral decision? Potentially also playing the biggest role is this idea you were talking about earlier about how do people internalize what is the right thing to do? How do you establish that this is the right thing to do?

We should talk more about this because, methodologically, this is important to separate out.

HUGO MERCIER: Can I say something about this issue of mentalizing? You're right in drawing attention to the importance of mentalizing in making moral decisions or moral judgments. It seems that the data indicates that we're not very good at it, that we have biases and we tend to not be very good when we think about what might have caused other people's behavior.

The reason is that in everyday life, as contrasted with many experimental settings, we can talk to people. If you do something that I think is bad, and we know from data about how people explain themselves, that spontaneously you're going to tell me why you did this and you're going to try to justify yourself. I don't have to do the work of trying to figure out why you did this, what kind of excuse you might have had because you're going to do it for me. Then we set up these experiments in which you don't have this feedback and it's just weird. It's not irrelevant because there are many situations in which that happens as well, but we still have to keep in mind that it is unnatural. In most of these games and most of these experiments, if you could just let people talk, they would find a good solution. The thing with the shocks, if the people could talk with each other, you could say "Well I'm happy to take the shock if you want to share the money." Then again I’m not saying it's not interesting to do the experiments at all, but we have to keep in mind that it’s kind of weird.

CROCKETT: That's true to a certain extent. A lot of moral decisions, particularly in the cooperation domain out in the real world, do usually involve some sort of communication. Increasingly, however, a lot of moral decisions are individual in the sense they involve someone that's not there. If you're deciding whether to buy a product that is fair trade or not, or if you're a politician making a decision about a health policy that's going to affect hundreds, thousands of people, millions of people who are not there. Some of the most wide-reaching moral decisions are made by an individual who does not see those who are going to bear the consequences of that decision. It's important to study both.

MERCIER: Maybe by realizing that the context in which these mechanisms of mentalizing evolved was one in which you had a huge amount of feedback can help us to better understand what happens when we don’t have this feedback.

CROCKETT: Maybe that's why we see selfish behavior is that we're used to having an opportunity to justify it when now there are many cases in which you don't have to justify it.

FIERY CUSHMAN: One of the things that’s unique and cool about your research is the focus on neuromodulators, whereas most research on how the brain processes morality has been on neural computation. Obviously, those things are inter-related. I guess I’ve always been, I don't know if confused is the right word, about what neuromodulators are for. It seems like neural computation can be incredibly precise. You can get a Seurat or a Vermeer out of neural computation, whereas neuromodulators give you Rothkos and Pollocks.

Why does the brain have such blunt tools? How does thinking about neuromodulators in particularly as a very blunt tool but also a very wide ranging one, inform your thought about their role in moral judgment as opposed again to neural computation?

CROCKETT: It's important to distinguish between the tools we have as researchers for manipulating neuromodulators, which are incredibly blunt, versus the way that these systems work in the brain, which are extremely precise. The serotonin system, for example, has at least 17 different kinds of receptors. Those receptors do different things and they're distributed differentially in the brain. Some types of receptors are only found subcortically and other receptors have their highest concentration in the medial prefrontal cortex. There's a high degree of precision in how these chemicals can influence brain processing in more local circuits.

To answer the first part of your question, the function of these systems is because cognition is not a one-size-fits-all kind of program. Sometimes you want to be more focused on local details at the exclusion of the bigger picture. Other times you want to be able to look at the bigger picture at the exclusion of small details. Whether you want to be processing in one way or the other is going to depend profoundly on the environmental context.

If you're in a very stressful situation, you want to be focusing your attention on how to get out of that situation. You don't want to be thinking about what you're going to have for breakfast tomorrow. Conversely if things are chilled out, that's the time when you can engage in long-term planning. There's evidence that things like stress, environmental events, events that have some important consequence for the survival of the organism are going to activate these systems which then shape cognition in such a way that's adaptive. That's the way that I think about neuromodulators.

Serotonin is interesting in this context because it's one of the least well understood in terms of how this works. The stress example that I was talking about, noradrenaline and cortisol and those neuromodulators are understood fairly well. Noradrenaline is stimulated by stress and it increases the signal to noise ratio in the prefrontal cortex and it focuses your attention.

Serotonin does tons of different things but it is one of the very few, if not the only major neuromodulator that can only be synthesized if you continually have nutritional input. You make serotonin from tryptophan, which is an amino acid that you can only get from the diet. You can only get it from eating foods that have tryptophan, which is most foods, but especially high protein foods. If you're in a famine, you're not going to be making as much serotonin.

This is interesting in an evolutionary context because when does it make sense to cooperate and care about the welfare of your fellow beings? When resources are abundant, then that's when you should be building relationships. When resources are scarce, maybe you want to be looking out for yourself, although there are some interesting wrinkles in there that Dave and I have talked about before where there could be an inverted U-shaped function where cooperation is critical in times of stress.

Perhaps one function of serotonin is to shape our social preferences in such a way that's adaptive to the current environmental context.

Hugo Mercier: "Toward The Seamless Integration Of The Sciences"

One of the great things about cognitive science is that it allowed us to continue that seamless integration of the sciences, from physics, to chemistry, to biology, and then to the mind sciences, and it's been quite successful at doing this in a relatively short time. But on the whole, I feel there's still a failure to continue this thing towards some of the social sciences such as, anthropology, to some extent, and sociology or history that still remain very much shut off from what some would see as progress, and as further integration.

HUGO MERCIER, a Cognitive Scientist, is an Ambizione Fellow at the Cognitive Science Center at the University of Neuchâtel.

TOWARD THE SEAMLESS INTEGRATION OF THE SCIENCES

I am Hugo Mercier. I'm a cognitive scientist, and I currently work at the University of Neuchâtel, in Switzerland, in the Cognitive Science Center. Today I want to talk about the integration of the cognitive and the social sciences, and in particular how the work of Dan Sperber can help us further that integration between the cognitive and the social sciences.

One of the great things about cognitive science is that it allowed us to continue that seamless integration of the sciences, from physics, to chemistry, to biology, and then to the mind sciences, and it's been quite successful at doing this in a relatively short time. But on the whole, I feel there's still a failure to continue this thing towards some of the social sciences such as, anthropology, to some extent, and sociology or history that still remain very much shut off from what some would see as progress, and as further integration.

There are several issues. Some of them are just purely sociological, but some of them are more substantial. Two of the issues I would suggest are that maybe we don't necessarily have the right tools to help people in the social sciences see how they can use the cognitive sciences, and the other is that in some cases we don't have very good models of high-level cognition. Even if they could integrate what we know about cognition with what they want to explain in the social sciences, we just wouldn't be able to provide them with the right mechanisms to tinker with. Some of Sperber's work can help us solve both of these issues.

On the first front, which is to have conceptual tools to integrate cognition and culture, to make cognitive and social sciences shorter. Just to give you a bit of background, Sperber trained as an anthropologist, but he very quickly realized the potential of the cognitive sciences to help us better understand cultural phenomena. One of the main things he brought about was the importance of communication. He saw communication as a way of bridging the cognitive level with the more social and cultural level because, obviously, most of culture is transmitted through communication.

One of the many things that his studies of communication have revealed is that—people kind of knew all along, but they hadn't really fully realized it, I guess—communication is extremely noisy. For instance, what I'm saying today, let's imagine there was no transcript. Whatever memory you will have will be extremely different from what I have in my mind, and then if you were to repeat that to someone else it would, again, be extremely different. That creates a big problem for culture, which is that, given that the transmission process is so noisy, culture—in the sense of having the same elements that you can identify, in many, many different people—should not even exist at all. Basically if you have one guy who has an idea, and he says it to someone else, and then that person says it to someone else, basically after a few steps you can't really recognize the original idea any more. If that's what happens you shouldn't have culture at all. Basically you should just have a bunch of ideas that maybe somewhat resemble each other, but that are too different to really be called cultural.

If you look at how communication works it raises the issue of this transmission noise that jeopardizes the very existence of culture, but it also provides some of the answers as to why culture can exist, and then as to what is more likely to become a cultural phenomenon.

When you look at how communication, especially ostensive communication, in humans works, it's a very rich inferential process that we don't really see usually. When someone tells you even something as trivial as, "It's raining," we think, there is this content, someone tells you something about the rain and you just have to understand what raining means and you're done. But, in fact, what the work of people like Grice, Sperber, and Wilson have revealed, is that you have this very rich process of inferring what the person actually means when she says it's raining. Even something as trivial as you have to understand that it's raining now, and that it's raining here, which is not said in the sentence. Usually even "it's raining" will mean many more things.

For instance, if Josh this morning had told me that it was raining, I would have inferred that he wanted me to understand that it would be complicated logistically today rather than just saying, "It's raining." That's what makes communication kind of noisy, but that's also what can make communication and this inferential process help culture stabilize, in that some cases it can correct communication that would have failed otherwise.

One of the examples that Sperber often takes is that of a tale. Imagine, for the first time in your life you're being told the tale of the *Little Red Riding Hood*. At the end the wolf's belly is opened and the Little Red Riding Hood is taken out, but the grandmother is not taken out. The person who tells you the story forgets to mention that the grandmother is taken out of the belly. What's going to happen, or what is likely to happen anyway, is that when you tell the story in your turn you will add that element back. You will correct the story because you might not realize that there had been a problem in the first place, but when you have to recreate it, the version in which both characters are taken out of the belly is in some ways more felicitous.

That creates what Sperber has called an attractor. The version of the story in which both characters are taken out is an attractor, so that other versions of the story that deviate from that are going to revert to that one in the process of transmission. That is what creates cultural stability. In that model what creates stability is not that transmission is faithful, it is that even though transmission is noisy, the noise is not pure noise; it's not undirected noise. It tends to go in some directions rather than others.

That's the general idea. Now with Nicolas Claidière and Thom Scott-Phillips, they have nice mathematical formulations of that that can be helpful for the model-oriented people, but I'm just going to give you a few examples of studies that are more or less recent, but most of them recent, that have been done using that concept to flesh it out.

The most famous studies that have been done using that concept were probably those of Pascal Boyer, who is an anthropologist/psychologist, who attempted to explain traditional religious beliefs, such as beliefs in ancestors, as being attractors. And in that case, what would make them attractive, is the fact that they are, as he calls it, minimally counterintuitive. In some ways they are counterintuitive because you have dead people that still are able to do things to some extent and to have thoughts and everything, but they are minimally counterintuitive in the sense that these dead people are very much like other people.

Basically, we can recruit all of our mechanisms of mind reading and theory of mind that we use to understand live people, to understand the desires, and the intentions, and the beliefs that dead people have. That makes most of these ideas intuitive, but they're still in some ways counterintuitive because the guy is dead, and that makes them more relevant. You have the right mix of being understandable, you can draw inferences, it's kind of interesting, but it's also extra interesting because it's not just run of the mill, well, someone has a belief, and someone dead who has a belief. That makes it more relevant.

Now, Pascal Boyer, with some other colleagues—Nicolas Baumard and Coralie Chevallier—are applying this to the spread of moral religions, the things that happened around the Axial Age when you had Jesus, and Buddha, and Confucius, and a whole different type of religion emerged. They're trying to explain that in terms of the changing psychology of the people that made some types of religion more attractive at different times.

That's one of the best worked out examples, but there are a few more recent that have been fun lately. One of them, which is nice because it takes the phenomenon down to its essence, is a recent study by Nicolas Claidière and his colleagues, which looked at transmission in baboons. What they do in these experiments is: the baboons play on the computer screen, in which there is a 10-by-10 grid of small squares. On that grid you have four squares that light up, and then they disappear very quickly. The baboons have to touch the screen where the four squares will light up.

What happens is that sometimes the baboon will make a mistake. Whatever the baboons manage to do will be transmitted to the next baboon. You have the initial thing that is random—you have four squares anywhere on the grid, and then the baboon does the thing. Whether succeeds or he fails, that will be transmitted to the next baboon. And you repeat that process many, many, many, many times.

What you see appear is basically Tetris, so you see the forms, the shapes of Tetris—the square, the line, and the S thing—appear because the mistakes the baboons are making are not random. They are making some mistakes, but they're not just going to tap any square. They're more likely to tap a square that is closer to the square they had originally tapped, and once you have a form that is an attractor, even though baboons will deviate from them from time to time, when they make a mistake from that deviation, they're more likely to revert back to the attractor than to go into some other direction.

You have these shapes that are extremely stable, so that once they're there they really stick around, whereas any other pattern will be much more volatile. It's a good example of attraction in that what is creating the stability is not that the baboons are really very good at repeating some shapes, because they're quite good but they're not perfect, but that there is this systematic bias to always pull in the same direction.

The other three quick examples will illustrate the different fields to which you can apply this. One is from the history of art. We know that humans have this mechanism that is extremely ingrained of paying attention to the gaze of other humans. We really pay attention. That's likely why we have the white in our eyes—so that we can really see where people are looking. In particular, something that is very salient is direct gaze. When someone is looking directly at you, especially for a long time, and they're not talking to you, it's really a signal of a strong emotion, whether it be lust, or aggression, or something else.

What this predicts is that cultural representations of faces that look directly at you should be in some way more attractive. They should be seen as more relevant. Okay? This is more interesting than a face that doesn't look at you. What Olivier Morin did is look at the portraits; he looked at two cases. One was 16th century Europe, the other was a span of 7th century in Korean portraiture.

He looked at two things. One was the evolution of the gaze of the portraits through the generations, and the other was which of these portraits were picked up by contemporary art books, and he found that in both cases the attractor hypothesis predicted what was happening. In both Korea and in Europe at that time you have a shift from portraits that looked away from the viewer to portraits that look towards the viewer. It turns out that the portraits that are selected in the art books now are more likely to be portraits that look directly at the viewer rather than portraits that look away. It seems as if this very fundamental communicative mechanism can explain a small part, obviously, of this cultural phenomenon.

The other example bears on that Leibniz-Newton dispute mentioned earlier. Just to give you a bit of background, you know they both invented—more or less at the same time—differential calculus. What is clear, though, is that Leibniz published his findings much earlier than Newton. He had a very strong head start, in particular in France. In England Newton was Newton, basically anything he said was fine, but in France Leibniz was published earlier, and he had this advantage from the start. You can see that, for instance, his notation—the *dx* and the big S for the integral—were kept.

However, when Newton was introduced in France it is his concept of the infinitesimal that won. It's kind of surprising because historically you think well, basically Leibniz was there before, he was hugely influential and had this huge prestige, and yet it's Newton's idea that ended up being used by all mathematicians. Not all physicists I'm told, but most mathematicians anyway. The idea is that the Leibnizian formulation of the infinitesimal treats the *dx*—that infinitesimal quantity—as an entity, precisely as something that exists. The claim that Christophe Heintz has made is that as soon as you start talking about entities that is going to recruit some of our numbers sense intuitions, or mathematical intuitions that will treat that as a little object, and that is going to make very hard to process the idea that *x* plus *dx* equals *x*. Basically we have this strong intuition that if you are dealing with an entity, if you add that entity to another entity, then you get something else. You don't just get the entity you had to start with, whereas Newton's formulation did not have this issue because it treated the infinitesimal more like a limit—it was not quite the concept of limit, it gave rise to the concept of limit later, but it didn't have this issue.

Even though in purely mathematical terms both concepts could have worked out, the fact that one of them jived better with our number sense, or rather that one of them didn't work out so well might explain why one of them was more successful. I mean this number sense that we're talking about is something fundamental—you find it in nonhuman animals—and the hope is that it can explain some of the most complex cultural entities ever.

The last example in that line of work is some work that Nicolas Claidière, Helena Miton, and myself have been doing on bloodletting. As some Americans might know, in the winter of 1799 George Washington got ill. The best doctors in the country were brought to his bedside, where they proceeded to bleed him of about four liters of blood. I guess that's less than a gallon, or something like that. That was not a good idea. That's about half of someone's blood. Then he proceeded to die. There's still some dispute about whether that killed him or whether he would have died anyway, but people are pretty sure that that did not help.

Bloodletting is this puzzling phenomenon that was the main therapy throughout I guess the 17th, 18th, early 19th century in Europe and in North America. For us, we don't think that it works; it's extremely puzzling that people would do this. You know, why on earth would you do this? It's the best doctors who are doing it to the most important people—they're also doing it to everyone else—and it seems very puzzling. The answer that historians in particular usually suggest is that it was mostly a matter of authority and prestige. You have these extremely prestigious ancient physicians, such as Galen in particular, who exerted a huge influence throughout the history of medicine in the West, and basically people accepted his humoral theory. You can derive bloodletting from this, and that is why people are doing bloodletting.

As a side note—I can't help but mention this—Galen was so much into bloodletting that he thought it was a good idea to do it in case of hemorrhage, which is kind of funny. Galen was a great guy, I'm sure. The standard explanation is that you go back in time, you have Galen, you can go back to the Hippocratic writers, and then you can even go back to the Egyptians, who have their own great story about how bloodletting was born, but it's mostly a story of prestige and authority. You have these guys who are unusually influential, and basically they could have developed any other theory and people would have accepted that. That's the common idea about bloodletting, and other forms of therapies used at the time.

Then you can think well, whatever cognitive mechanisms people have doesn't really matter, all they need is some kind of bias to listen to prestigious individuals. In order to test whether that was the right explanation, we did two things. The first was to look at the anthropological data to see if people who had not been influenced by these early Western physicians also practiced bloodletting. It turns out that that is the case, many, many, many cultures in the world practice bloodletting; In North America, the Native Americans used to do it with cute little bows and arrows; In Australia it's done with stones most of the time; In Africa it's done in many ways, including using horns. It seems as if many cultures throughout the world have found the idea of bloodletting rather intuitive. That shows that it's not just a fluke that these guys in Greece and in Rome just got that idea, and that spread to us. It seems as if there's something that makes bloodletting a rather intuitive form of therapy.

In order to confirm that we did some experiments with American participants who we checked don't believe that bloodletting works, most of them anyway. What we did is that we gave them stories that involved something that looks like bloodletting, in the context of an Amazonian tribe, so it's plausible that they could do something like bloodletting. We give them that story, then we distract them for a little while, and then they have to recreate the story, and we take that story and we give it to someone else. It's like the baboon thing earlier, except that it's done with stories instead of that grid.

When you do that for a number of generations, what you see is that the stories tend to converge towards bloodletting, to some extent. In the case that is the most striking we had some of the stories starting with something like, "This guy in an Amazonian tribe has a headache that stops him from hunting a bird he's supposed to hunt for some ritual reason. The guy has a headache, and at some point he cuts himself with a stone." We specify that it's an accident. "The day after, his headache gets better." (As all headaches do. That's what they do.)

What we see is that after several generations, in some cases, the thing that was accidental starts to become intentional and it starts to cause the recovery. In the end, in some cases you have full-blown bloodletting. You have people who say, "Well, the guy had a headache, he took a stone, he cut himself, and that made him feel better." I'm not saying that the people believed in it, but it's more intuitively appealing than just having this story about the guy who has a headache and cut himself. I mean you can see how it could explain, to some extent, the emergence of the phenomenon, because someone being sick, cutting himself, and then getting better is bound to have happened very often. Then you tell that story, and in the process of transmission you can see how you have the story about a guy who cuts himself intentionally, and that makes him better. It's been fun working on this.

The first part was this set of methodological and conceptual tools that can link what we know about cognitive mechanisms with cultural phenomena. One of the other things that Sperber brought is a better understanding of some cognitive mechanisms that are really important to understand many cultural phenomena.

The first mechanism, and I mentioned that earlier, is communication. What they brought with Deirdre Wilson in this theory that is called Relevance Theory is this idea building on Grice that we have, as I mentioned earlier, all these levels of intentional mind reading or theory of mind that are involved communication. For instance, if Josh tells me it's raining what I'm really processing is something like Josh wants me to know that he wants me to believe that it's raining. And that's kind of counterintuitive because it doesn't feel as if we're doing this whole work, but we can see that that's what's happening, if you look at other means that Josh might have to make me believe it's raining.

For instance, maybe he wants to play a practical joke on me, and he turns on the sprinkler in order to make me believe it's raining. Now you have Josh who wants to make me believe it's raining. Now if I see him doing this, I know that Josh wants to make me believe it's raining, but that's not going to help him make me believe it's raining. On the contrary. Then when you get at ostensive communication, if I see Josh telling me, "I'm trying to play a practical joke on you," in which case I understand that by turning on the sprinkler he wanted to make me believe that it was raining. And that's where you have full-blown communication. And if you don't have all of this, human communication doesn't work.

Not only is it counterintuitive, but for a long time people thought that it was not plausible for two reasons. One was that children were thought to be really bad at doing theory of mind and mind reading (children below three, basically, even though they obviously communicate fairly well). The other was that adults were thought to be really bad at doing many levels of mind reading. People thought that if you do four or five, it saturates your cognitive abilities. More recent experiments have shown that infants—I don't know when the youngest experiments are, but at least ten-month olds—can do some rather sophisticated mind reading. New experiments by Thom Scott-Phillips are showing that adults can do up to at least seven levels of mind reading without any issue at all if you do that in the right context. Some of the issues about this have been dismissed, and now we can come back to this idea that we're doing all of this work when we're communicating.

Just to give you an example of how that can be used to understand some cultural phenomena, Alessandro Pignocchi, another colleague from Paris, is doing great work trying to understand how art is processed as ostensive communication. When you see a painting or when you see a movie, your brain treats this as the artist trying to tell you something, and you attribute intentions to the artist. When you see a painting of, let's say, a sunset, it's extremely different from just seeing the sunset because, even though you're not doing that consciously, your brain is figuring out why did the artist depict the sunset this way and that way, et cetera. That helps Alessandro integrate some of the findings about the cognitive science and how we treat communication with how people understand art, and what art is more successful.

Another thing that has been very important that Dan has been doing related to communication is stressing the importance of the mechanisms that allow us not to understand communication, but to evaluate communication. Most of the cognitive sciences that have looked at communication have focused on how we understand communication, basically linguistics, pragmatics, semantics, etc., and not how we evaluate it. It's implicitly taken for granted that most of communication is going to be honest, and that people just have to understand, and then you're fine, whereas, in fact, from an evolutionary point of view, communication can be used to mislead, to lie, to deceive, and we have to be careful about what other people tell us.

The idea of epistemic vigilance is that we would have a whole set of mechanisms that would be devoted to evaluating other people's communication in order to make sure that we don't get deceived, too often anyway. One of the most exciting developments related to that has been a huge amount of work on children, showing how good children are at telling who they should believe, and who they should not believe, research done by Paul Harris, in Harvard, by Kathleen Corriveau by Fabrice Clément, by Olivier Mascaro and many other people. They have these great results showing that in some cases children—infants, 12-month olds—will be able to integrate their own prior beliefs with information that is communicated to them, and even to discriminate between experts and non-experts. It's a really precocious thing. And adults, we don't realize it, but we're extremely good at doing this.

One of the interesting consequences of that, or one of the interesting applications for cultural phenomena, is that it flies in the face of beliefs we have about the efficacy of propaganda, of advertising, of political campaigns, and we tend to think that people are rather gullible. Many of us, and even some of our professional colleagues, have written that we're quite gullible. We start by accepting information rather than being careful from the start, and so the work that I'm doing at the moment is trying to show that, in fact, all of these cultural phenomena—propaganda, political campaigns, the news, advertising—in fact, are much, much, much less powerful than people usually take them to be, and that whatever influence they have on people is fully consistent with people being extremely vigilant with communicating information.

The very last bit, which is really the one that if there are any questions I'd rather they be on the last bit, because it's really the one I know something about. Dan and I have been working on this theory of reasoning, which is related to epistemic vigilance. The broad framework of epistemic vigilance is that we have to have a bunch of mechanisms that protect us from potentially misleading information, and that basically allow communication to work smoothly, and to remain mostly honest.

What we have suggested is that one of these mechanisms is reasoning. People used to think of reasoning as a mostly individual skill—you reason to make better decisions, you want to make sure that you have sound reasons for your decisions, or for your beliefs. What we've been claiming is that instead reasoning is something that is done for argumentation. That is, people reason so that they can produce arguments to convince others, and so that we can evaluate other people's arguments. We have a bunch of empirical results on this that serves this theory, but I'm not really going to talk about that now.

What I'm going to talk about is briefly how that can be used to explain some cultural phenomena by relating low-level communicative mechanisms with very complex cultural phenomena, such as complex religious beliefs or complex scientific beliefs.

In the case of science, we've been doing a little of that with Christophe Heintz, and in the case of religion Helen De Cruz has been doing a bit of that. The idea is that when you're arguing, you're recruiting people's intuitions to make something intuitive that was not previously intuitive. You're shifting around their intuitions so that just at the moment when you're setting up the argument, they say, "Oh, yes. Right." I can recruit this intuition to modify a belief I had before. If you repeat that process many, many times, you can see how you can arrive at beliefs or decisions that are extremely counterintuitive, and that seem, well, completely unrelated to what we know about most of cognition, because it is counterintuitive, but in fact, you can trace a chain that each step is relatively intuitive. Take an example like a standard mathematical proof, each step is supposed to be relatively intuitive, at least for the people who have the right skills, but then if you just take the axioms and the theorem no one can really have the intuition that the two will fit. Each step is intuitive, but you have to go through the steps.

One very last piece I want to mention to illustrate how these concepts can illuminate some cultural phenomena is a nice piece of anthropological work by Radu Umbres, an anthropologist who has used this concept of epistemic vigilance. You send people on a snipe hunt, which is this animal that's supposed to be impossible to catch, and you send a novice hunter, and you say, "Well, you have to catch this animal." You make up all kinds of stories, the idea is to show that novices are gullible; they don't know anything. What's interesting is that it reveals how, in most cases, these are really the exceptions. I mean what makes them kind of funny and interesting is that they are exceptional. Most people usually do not get taken in by these things, and even when they do get taken in it is in a context in which it makes sense, because you are a novice and you assume that everything else he tells you is stuff that you didn't know about before, and that turned out to be right. It's not unreasonable to trust him in this case as well.

This is my last example; it's not at all my domain, but it's interesting stuff that's going on now, and more of it will happen soon.

THE REALITY CLUB

MICHAEL MCCULLOUGH: I was interested in your bloodletting example, because it clearly is an example of something that people must have some sort of deep intuition about, and then when you combine it with a prestige bias it leads to the repeating of this trope for millennia. At the same time, you know, we have medical anthropologists scouring the globe to eavesdrop on traditional hunter-gatherers and horticulturists who know about medicinal plants that do work for reasons they can't explain, but at least some of the time do empirically work.

Sometimes it seems like we recruit a prestige bias that ends up being a failure, colossal failure, and other times we seem to recruit a bias to copy the successful. This is all cultural evolution stuff that both of those sorts of heuristics might be at work in an individual mind. How can we make predictions about untested phenomena that can tell us when one of them is going to win and the other is going to lose, if those sorts of heuristics are at work simultaneously?

MERCIER: We're just trying to explain stuff, that's kind of hard enough already. You'd have to know a lot about the specifics of this situation in order to be able to make predictions. I mean, clearly, you can do predictions. Like, if you run an experiment you can say, well, my theory predicts that in this case the prestige bias will be trumped by something else, or vice-versa. If you look at real-life cases it's going to be hard in the foreseeable future, I guess, to make predictions regarding whether it's the most intuitive idea that will win or whether it's the idea that is defended by the most prestigious individual, even though it's counter-intuitive, that will win. It's going to be tricky to make predictions in the short run.

MCCULLOUGH: It would be great if we had some way of putting those biases on common scales, so we could set up horse races.

MERCIER: Yes. It's complicated, because then they can interact in complex ways. Let's hope that that happens.

IAN REED: The bloodletting interests me as well from a clinical perspective, because there is a subset of patients who feel better by cutting themselves and seeing their own blood.

I've had particular patients who actually call it bloodletting, so it might be that a specific intervention might be useful for one population and then just inadvertently extrapolated to be useful for entire population, and maybe that's how certain things …

MERCIER: Yes, that's a very good point. In the research we talk a little bit about this non-suicidal self-injury, as I believe they're called, and it is striking that even in cultures that obviously don't practice bloodletting you still have in some people this intuition. As you are saying it's an interesting possibility that these people brought about the thing in the first place. Then again, if that happens and that does make you feel better, then you say, well, maybe we can emulate that, so it makes sense. That's an interesting idea on how that could have emerged.

REED: And it happens with treatments for psychological disorders all the time.

MERCIER: That's a good point.

LAURIE SANTOS: Just to give a nod to my colleagues in other social sciences, it's all well and good for cognitive scientists to be here, and be like all the stuff we do explains these mysteries in culture, and so on. Are there any cases where you think the bigger social sciences can come back and tell us about intuitions we just didn't know were there in cognitive science?

We could go to art galleries and look around and be like, "Why is all this stuff here?" And they'd be like, "Wait! Maybe there's this intuitive bias that we completely misread."

MERCIER: Potentially, there are many, many cases like this. On the whole it's pretty dispiriting to see sometimes how little psychologists or cognitive scientists know of the population level phenomena that are very much involved in the mechanisms of what they're supposed to study. Just to give you an example, psychologists who study emotional contagion, not all of them, but some of them take it for granted that the view of crowds and panic, as you know, basically as soon as you have an emergency or a threat most crowds are going to panic. All hell is going to break loose, all the norms are going to be trampled, along with the actual people. What's interesting is that this is completely false. All the sociologists and social psychologists who have studied this know that that basically never happens, that people are extremely pro-social when there is an emergency and a threat.

You have this huge disconnect between the two, and I'm sure that the cognitive scientists could help the sociologists better understand this phenomena, but at the moment it's really mostly the cognitive scientists who would need to hear what the people are telling them, because look, this is just not what's happening at all. It's a case in which when they just have wrong beliefs.

DAVID RAND: In this attractor idea, it seems to me that then the key question is “What makes something an attractor versus other things not an attractor?” In some of these perceptual domains that you were talking about, I could see maybe it’s some fundamental aspect of cognition. But a lot of what I think about is social behavior and social norms, which are a similar flavor. Have you thought about that stuff in the social domain, and do you have thoughts on what makes some things particularly attractive?

MERCIER: I haven't, but some people have. I'm thinking, for instance, of the work that Nicolas Baumard is doing, another one of my colleagues, and he has the theory of where our sense of fairness comes from. This idea is that we have an intuitive sense of fairness that basically relies on default, you know, 50-50 distribution that can be influenced by the contribution of the partners and what not. It's kind of easy to see how norms can hitchhike on this, because this is an intuition. This is really an intuition you don't need an explicit norm to have this intuition.

RAND: Where does the intuition come from?

MERCIER: Well, you have a cognitive mechanism in your head that evolved. Basically, it's a proximal mechanism that fulfills the ultimate function of making you a good partner, making people think you're a good partner. The intuition, evolutionarily that's the story that you have. Partner selection makes it valuable for people to develop a reputation as good partners. One of the things that makes you a good partner is to be fair, and then basically that gets written in our cognition as this sense of fairness that, in some cases, you can recognize fairness and you can try to be fair when you think it's worth it for you. And then once you have this thing it's quite easy to see how an explicit norm that taps into this intuition would be more successful than when it doesn't.

RAND: But that doesn't explain variation in norms, right? That seems like the interesting question: why some things are attractors in some contexts?

MERCIER: Very good point. One thing we need to bring in is that in different environments, like the cost and the benefits will be different, and then that will predict differences in the norms that will apply.

Jennifer Jacquet: "Shaming At Scale"

Shaming, in this case, was a fairly low-cost form of punishment that had high reputational impact on the U.S. government, and led to a change in behavior. It worked at scale—one group of people using it against another group of people at the group level. This is the kind of scale that interests me. And the other thing that it points to, which is interesting, is the question of when shaming works. In part, it's when there's an absence of any other option. Shaming is a little bit like antibiotics. We can overuse it and actually dilute its effectiveness, because it's linked to attention, and attention is finite. With punishment, in general, using it sparingly is best. But in the international arena, and in cases in which there is no other option, there is no formalized institution, or no formal legislation, shaming might be the only tool that we have, and that's why it interests me.

JENNIFER JACQUET is Assistant Professor of Environmental Studies, NYU; Researching cooperation and the tragedy of the commons; Author, *Is Shame Necessary?* **Jennifer Jacquet's *Edge* Bio Page**

SHAMING AT SCALE

My name is Jennifer Jacquet. I'm an assistant professor in environmental studies at NYU and I'm interested in large-scale cooperation dilemmas. A lot of those are environmental in nature. I wonder about what it's going to take to leave 1,700 billion barrels of oil in the ground, or half the fish in the ocean, or to remove nitrous oxide from the atmosphere so that we don't deplete the ozone.

The interesting thing about conservation, and science in general, is that it's moving into the social sciences and into questions about human nature. You would say, especially someone like Josh Knobe might say, well, that's not that interesting because a lot of fields, including philosophy, are doing more empirical work. Gender studies are also moving more into the social domain and empirical data collection. The same pertains to African-American studies. But the interesting thing here is that with conservation science it was epistemologically and institutionally a discipline in the natural sciences, rather than the humanities.

I find this move interesting and also challenging for a lot of hardcore biologists and ecologists who have traditionally dominated the field to recognize that the most important interrelationship is not between the plants and animals, or the animals in the ecosystem, but between humans and the environment. I view there being a big wave of environmental social science coming on board, and I'm part of that wave.

My supervisor for my PhD was a fisheries biologist. He did important work early on in basic science on oxygen and growth level, population ecology, and then only later in his career, realizing the problems out there in the ocean, turned more toward social questions, about subsidies, and the affects of marine reserves, and had this more human-dominated view. That's one exciting wave in social science, at least in the realm of conservation.

One of the first disciplines to get on board with this has been psychology, so we now have conservation psychology and environmental psychology. They've done important work. The American Psychological Association released a report in 2009 about the effects of climate change on psychology and vice-versa. But I don't think psychology is going to be the only social science we need in this pursuit, because of the focus on the individual, and I'll get to that in a moment.

My reason for turning to social science from the natural sciences is because I became interested in guilt as a motivation for changing one’s behavior, for changing one’s decision-making, and I saw guilt being prevalent in environmental issues, such as over-fishing, climate change. That interest in guilt led to shame, to the point where people started describing me as somebody who worked on fish and shame, which was just a little bit weird, but I hope to tie that all together for you today.

Backing up to some of that more basic or even humanities type research on guilt, it's argued that guilt is a relatively new phenomenon, in general. It's primarily a Western phenomenon, linked to the rise of individualism as a characteristic, and more prevalent in the West than the East. In some Eastern cultures they don't even have a word for guilt, and there are others who argue and philosophers who argue that it's also linked to the rise of abstract thinking.

My own view on guilt is that it's highly dependent on how much time you get to spend alone. I think that when you have zero chance of spending any time alone in your society, you're very unlikely to have strong feelings of guilt every day, in part because I view guilt as defined—and there are lots of arguments, and you all know these better than I do, definitions about what guilt or shame really mean—but guilt is internalized, and the only person you're answering to is your own self. I view guilt as the cheapest form of punishment there is. It's self-punishment, and you prevent the group from having to punish you by either cutting yourself off from doing the act, to begin with, or paying some sort of penance afterward.

Shakespeare used the word "guilt" only 33 times. He used the word "shame" 344 times. So when we start thinking that it's just a Western thing, we should also note that it's even more modern than just being a Western phenomenon. My own particular interest is in environmental guilt, which I see this rising a lot, basically beginning in the 1980s, and I tie this to a switch from a system that was focused on changing a supply chain and production of chemicals or bad products, to more a demand-focused side strategy.

With that demand focus strategy, the focus on the individual, guilt was an easy low-hanging way of getting people to engage with the issues. Of course, there's a big threshold problem there. Because it's linked to a switch from the focus on supply to the focus on demand, it means that its power is very limited.

If you ask does this behavior scale, I would argue, no it doesn't scale. Does the U.S. feel guilty for doing something? Does BP feel guilty for the Gulf Oil spill? By the very definition of what guilt is—an internal regulation of one's own conscience—it implies, at least to me, that it does not scale to the group level; although, you have these trends, like survivor guilt or collective guilt, that call this into question.

I am interested in social problems, so maybe we should focus on the types of social emotions that might scale, and not just social emotions, but social tools, and that's why I got interested in shame as a tool, which is separate from shame as an emotion. We could all disagree here about what shame is as an emotion. A lot of people agree that it requires some sort of audience, but some don't. Some people argue it's a sense of your whole self, or as guilt is just based on the transgression itself. But I want to focus on shame as a tool, as a punishment, and situate it within a larger body of punishment.

I would like to distinguish shame, starting off, from transparency. A lot of people confuse them in the popular media, thinking that they're the same thing. Transparency exposes everyone in a population, regardless of their behavior, whereas shame exposes only a minority of players, and this is an important distinction. Both shame and transparency are obviously only interesting if the distribution is not uniform. So we have to have some variability in there; otherwise, we're really not interested in the behavior. I want to argue, too, and one of the points I make in some recent work, is that shame is more effective the larger those gaps are, not just between existing behaviors, but between what we think should happen and what is actually happening.

Shame and transparency are different. Shame, in fact, even though we all have a knee jerk reaction to thinking it's a terrible tool, can be more protective than transparency. We have these groups like the Sunlight Foundation that say we want to expose all politicians' behavior, and actually, I argue that maybe only exposing the worst of the players' behavior can be more beneficial than exposing everyone.

The way that I first started working on shame was to look at public goods games, these cooperative dilemmas that have already been talked about, where you can either donate money or not, and then that money is doubled or tripled and redistributed evenly to all players, so there's that tension between the individual and the group. What's interesting about punishment in these games is that the way it's been operationalized is entirely monetarily, so you play a little bit of money, and you extract a bigger amount of money from the person you're punishing. This is a very one-dimensional way of looking at punishment.

We should get a little more creative about the way that we operationalize punishment in these games. Of course, it's very hard. Maybe we could shock people. We can't put them in prison. We can't kill them. That wouldn’t get past an ethics review. But these are the forms. We have these kinds of deprivations: we can either remove life, liberty, physical safety, resources, or reputation.

In our experiment we told players at the start of the game that the two players that donated least out of six would be exposed at the end of the game. That was the only threat of punishment they received, this burden of social exposure, what we would call shame, again, as a tool, rather than an emotion. We didn't even measure how they felt. We didn't care how they felt, frankly. We were caring about the behavior that manifested, and, again, that's because I don't come from a psychology background. I'm interested in a more economic side of things, especially with the scaling effects.

What's interesting is that we have these forms of punishment monetarily, and no one would ever question those ethically. But after our experiments came out, I was blindsided by the fact that some people didn't agree with the experiments from an ethical perspective: that we would expose these two players to the group for having contributed the least—despite them having known at the beginning of the game that this was the case. It's interesting to me that people would say this isn't ethical, because we use this form of punishment all the time in society in all different sorts of ways. If we view it unethically in the academic environment, and yet we're using it a lot in society, it makes it very hard to make any sort of empirical statements about what we should or shouldn't do in the future. Either we should get rid of social exposure entirely in society, or we should test it in a lab and see how it works, but we can't continue this disjuncture between having lots of it in real life and nothing in the lab. What we were able to show is that the threat of shame increased cooperation by about 50 percent. You don't get a lot of movement because there's a binary choice between whether or not they want to give one dollar or not in these games, but I find that that's interesting, because unlike the other forms of punishment in these games it was just reputation. It was only that they had to stand up in front of the crowd at the end. Nothing more. But we did recruit students, and this would be hard, again, in the replication. They would have to read very closely who came from the same class, so they did know that they would see one another again in the future, which matters, of course, to ultimately maybe more resources or reputational effects down the line.

One way in which social science can also push the field is to look at these different forms of punishment in cooperative dilemmas or otherwise, in part, because they do scale, and also because, again, we have them widely used in society, but not actually looked at empirically.

In my work I have tried to look for studies that used shame in various ways, maybe not in a lab experiments, but there were studies that used it in natural field experiments, especially with voting behavior, and they were interesting because they find that shaming actually leads to the greatest change. The experiments involved sending letters that said here's your voting record and here's your neighbor's voting record, and after the next election we're going to send you an updated version of this. They were basically threatening exposure to your neighbors of your own voting behavior, and this single piece of mail actually increased voting by eight or nine percent, and normally no piece of mail works better than one percent.

On the other hand, with one of the shaming conditions, so many people called and were angry about it that ultimately they decided not to do part of the experiment, which in this case was to publish the names in the newspaper. It shows that shaming might be very effective as a tool, but very impermissible as a means of changing society. This is what I'm looking at now, ways in which you can make shaming more effective as a tool, which, again, you have to just be completely agnostic about. And then there's the bigger question that is more on the normative side of things, which is just because it's effective does that mean it's acceptable in society? Does that mean that we want to use it?

A real-world case, in which shaming was very effective, was when Amnesty International went after the U.S. in the international media for executing juveniles. Until the law was changed in 2005, we were one of the only countries, certainly the only Western country that is executing juveniles. Amnesty International conducted a large-scale shaming campaign that worked, which points to a few things.

Shaming, in this case, was a fairly low-cost form of punishment that had high reputational impact on the U.S. government, and led to a change in behavior. It worked at scale, one group of people using it against another group of people at the group level. This is the kind of scale that interests me. And the other thing that it points to, which is interesting, is the question of when shaming works. In part, it's when there's an absence of any other option. Shaming is a little bit like antibiotics. We can overuse it and actually dilute its effectiveness, because it's linked to attention, attention is finite. With punishment, in general, using it sparingly is best. But in the international arena, and in cases in which there is no other option, there is no formalized institution, or no formal legislation, shaming might be the only tool that we have, and that's why it interests me.

It makes sense to me that evolutionarily speaking we would need harsher forms of punishment, because if shaming was perfectly effective, if shaming was the ultimate tool, we wouldn't throw people in prison, behead them, or anything like that. It is, I would argue, a hit or miss type tool. It can really work, or not work well, depending on a lot of things that I'm trying to explore in doing research which looks at what makes shaming more or less acceptable, and what makes it more or less effective.

THE REALITY CLUB

DAVID PIZARRO: The features that make shame a handy tool also in some ways are what make people very afraid of it. For instance, in order to make something legal there are usually procedures you have to go through, where you can get most of the members of society to agree that this ought to be outlawed. Shame can be used as a tool to enforce something that a wide segment of people might not believe, the person themselves might actually not believe, so unlike with guilt you don't have to endorse that you did something wrong to experience shame. And it is, as you pointed out, powerful. So we get things like slut shaming, right?

JACQUET: Sure.

PIZARRO: And just because any group of people can decide that you ought to feel shame, they can exert this social pressure that could be—as you rightly point out because it's such a fundamental part of our human nature—arbitrary, capricious, and in the hands of people who could just decide on a whim that they don't like you. The tradeoff and why many people may be against it and another problem is you could have, for instance, is that with the Internet you have this vigilantism of shame where it's such low cost to shame somebody else that the fear is of tyranny, this particular kind of emotional tyranny that doesn't seem … like, strangers can't make me feel that guilty, but they can sure make me feel shame. Is there anything to the thought that as effective as it might be, we should just get rid of it?

JACQUET: Get rid of it? Yes. Absolutely. The main argument here is that shaming undermines human dignity. The role of the state, or one if its goals should be to protect human dignity, and, therefore, we should outlaw shaming for individuals entirely, something I am not that opposed to.

PIZARRO: Well, not laws. I'm saying as like good people, like a law … in fact, without laws …

JACQUET: You're saying why not just have a law instead of the shaming?

PIZARRO: No. I'm saying that as human beings we should frown on the use of shame. I don't know whether laws can change people.

JACQUET: Imagine the type of law that you would need, especially in the digital space, and how it would be operational. How it would affect what we would argue would be free speech, or activism, or because it’s just words. That's the crazy power of shame. I am in favor of the state not being involved in shaming individuals. I don't know about other people. It's very hard to control behavior. There's some level of decency that should speak to this. That's why I'm interested in using shame at the group level.

There are cases, in fact, with individuals in which it's warranted. We had talked about civil servants earlier, and things like that, and certainly there are very strong arguments. The argument I find even more compelling is the cost on the audience. The reason why shame works is that there's an audience that you imagine is endorsing this position, and for shame to really work there has to be an audience that's heckling. This can be some kind of imaginary audience, such as a bunch of fake Twitter names**.** Again, all of us, as in the antibiotic case, would benefit from shame at least being used very, very sparingly, even for our own just audience-member perspective, because shame demands our attention, let alone the fact that we could eventually become its victim. But while shame is absolutely the last resort, I don't think we should take it off the table entirely.

DAVID RAND: As an opposite of Dave's question, I had a thought on how to do it better, more effectively, that is.

LAURIE SANTOS: Good David and Bad David.

RAND: Like the case with the voting study, which is cool, and what we've been doing in our observability field studies, with the blackout prevention. They both have potential for the same thing, because basically, while you can question whether it's shame or a celebration of people doing good things, either way it's reducing privacy and making people's behavior observable to people around them, right? It's a huge thing that if it's obvious that the reason that you're doing it is to shame them, it pisses people off. But there's ways that you can provide exactly the same information with a different purported purpose, and it can be effective without pissing people off.

In the case of the blackout prevention study, it wasn't a case of putting your name and address on the signup sheet so that the other people will see you, it's just that when you have a signup sheet, it seems like a natural thing that you're going to have to put your name and address on it. I'm sure that for people in the observability condition, it never occurred to them that that was being manipulated. Or, with other kinds scaling versions of that, i.e., if you want to have something where you advertise on Facebook that you did some energy-improving remodeling of your house, or something like that, we can't say, "Put this badge on your Facebook page to show off to people how good of a person you are," because people will be unhappy. But instead we can say “People don't know about this program, so if you post this badge other people will know, and that will help spread the word.” I guess it's a little bit harder with shaming. Perhaps that's another question. What do you think about shaming versus celebrating?

JACQUET: The observability stuff, even the eyes example, are cool and interesting experiments, but it turns out the effect wears off after like 12 months, without any real punishment coming in, even CCTV effects. Regarding observability, I would argue you can get these spikes in improvement, but for the long term it doesn't necessarily play out in the same way. And also, with the blackout stuff, the cooperative dilemma is quite different.

Then, for instance, let's look at the tax delinquent problem. There are tax delinquents in every state. If you can imagine some system of observability or transparency it would be unfair if you paid your taxes on time—if there was a list that said: paid taxes on time and listed the amount. A lot of us don't want people to know our salaries, or our taxes. But in the case of California they say: we're going to publish online the names of tax delinquents, and we're going to actually send them a letter in advance, so they have six months to avoid being posted, which is also key. The threat of shame is more effective than the act of shame, because once you've shamed the delinquent you create a reputational effect where people think the damage is done, so why not continue to defy the norm.

But on top of that, California only exposes the top 500 worst. It's actually a protective. It's not transparency. It's not every tax delinquent. It's just the very worst. I would argue observability in this case to the general population would be very undesirable, but the shaming option is perfectly acceptable. That's where with the blackout system there's less variability in behavior overall—the tax example is almost like a power law distribution.

RAND: I mean to me what you just said *is* observability. All observability means is making some information about what people do salient or available to other people.

JACQUET: Sure. I'm just using observability as a synonym for transparency, where you expose everyone's behavior, and shaming as a targeted exposure on the minority of bad actors, which is a key distinction in the way the policies play out. It's really different.

RAND: Right. But then whether shaming the bad actors or celebrating the good actors is going to be more effective probably depends on which is more common. In the blackout study we had about five percent of people signing up.

JACQUET: But imagine celebrating the good actors in the tax issue. You would never do this, right?

RAND: Exactly, because that's a situation where that's the very common …

JACQUET: It's not that it's very common. It's that …

PIZARRO: That's what Molly was saying about above and beyond.

MOLLY CROCKETT: Yes. Well, not just that, but this very question has been worked out theoretically by Roland Benabou who has looked at, in a theoretical, mathematical sense, which is more effective, celebrating or shaming, and it turns out that it really depends on the base rate. He gives the example of hybrid cars. Initially, hybrid cars were quite rare, and so incentives worked really well and celebrating worked really well, because you get a free pass to drive in the carpool lane, so on and so forth, and that encourages take-up of that behavior. But then once it becomes widespread then the optimal strategy shifts towards shaming those who violate the norm.

JACQUET: But it depends on more than just the distribution curves as well. You can imagine that with climate change, we could honor the countries that are actually doing something about it. But without everybody on board it doesn't matter, because it's a threshold collective risk dilemma. Therefore, the type of problem actually defines which one you go for rather than just the distribution. That's why the two probably intersect in an interesting theoretical way.

FIERY CUSHMAN: We've addressed several times today this question of do you want to just achieve a behavior, or does it matter what the psychology underlying the behavior is, and we've talked about that in terms of the agency of the individual and what counts as truly moral behavior, but there's also just utility consequences.

I could keep myself awake by drinking a delicious latte, or by poking myself in the eye continuously. Maybe poking myself in the eye is a better way of keeping myself awake, but the latte tastes better. And one might also think that being in a society that focuses on celebration to the maximum extent possible and uses shame as little as possible is just going to make for a happier citizenry, even if there's a slight hit that you take in terms of the behavior that you want to maximize.

JACQUET: That's where the issue of human dignity gets interesting, because yes, at the psychological level that's probably true, and shame is this horrible, terrible thing that we hope to all avoid, but at that group level, where you could argue that corporations don't have the same level of human dignity, or Congress doesn't have human dignity, or the U.S., or Yemen, or whoever, then things get a little more interesting, because then you're saying this is just about reputation, and we don't mind hurting their dignity, because the people can come or go from that group as they please, and because of that you can say it's an effective tool, and we might have taken it off the table for these psychological reasons that don't apply at scale.

JOSHUA KNOBE: I'm following-up on Fiery's question about what is it that makes us so resistant to the use of shame as a method. Consider your case of tax evaders. Suppose someone said: “We have two options. One is to put tax evaders in prison, and the other is to put billboards on the street -- for instance, a big picture of the tax evader.” And then suppose that the tax evaders all prefer the billboard; they would much rather have a billboard with their picture saying that they're a tax evader than to go to prison. I would still feel terrible about it. It still seems so demeaning to have the billboard. You don't feel like there's something drawing us away from shame that goes beyond …?

RAND: Prison is pretty bad.

KNOBE: In a sense, as a society, we don't want to be the society that has the billboards?

JACQUET: I do feel that way, in part because again, as the audience, I'm asked to be part of that, I'm asked to be complicit in the punishment, and a good democracy was based on the idea that the state has the only authority to punish at severe levels.

SARAH-JAYNE BLAKEMORE: What about the relationship between religion and shame? Because that seems like it hasn't been mentioned, but some religions seem like institutionalized shame.

JACQUET: Sure. *The Scarlet Letter* was a product of the church itself. Maybe the government was more in favor of what God thought was right than what the majority thought was right, but it hurt human dignity, and there are long-term consequences that aren't great for individual psychology.

BLAKEMORE: They still embraced religion.

JACQUET: This is what happens when you lack formalized punishment. It's what Foucault argued in *Discipline and Punish*, with prisons. When you could send someone away to prison this made a lot of shaming punishments go away, because there was another formalized mechanism for punishment that we didn't have before. We didn't have that liberty option for deprivation. Reputation was one of the only forms aside from life, from cutting somebody's arm off, or taking their house away, etc.

And this is why shaming is used in the tax case. At the federal level you can go to prison, so there is no shaming option there, but at the state level all they can do is confiscate second homes and luxury vehicles, and in the time it would take for them to get that legislation changed, because there's so much resistance to anything related to legislation and taxes, anything, they said here's our stopgap. We're going to use reputation, and we're going to get back $340 million, which makes it totally worth their while, as it only costs them $180,000 a year to run.

In our society we ideally want these punishments formalized, and we want due process in all these things, but in the interim, it's an interesting sort of group punishment that's accessible to everyone, which is the scary thing about it.

HUGO MERCIER: I have a quick question about this scaling that you mentioned, how it works exactly, in that I couldn't imagine how shaming would work if no single individual feels ashamed. That is, Congress can't feel shame obviously, so if any individual congressman or congresswoman doesn't feel shame, then why would they do anything?

JACQUET: I'm not even certain that shaming's effectiveness is because of an emotion. It could just be linked to reputation and the fear of losing resources down the road. There are a lot of psychopathic people who would still respond to shame out of the fear of them being ostracized, or out of fear of something harsher later.

Simone Schnall: "Moral Intuitions, Replication, and the Scientific Study of Human Nature"

In the end, it's about admissible evidence and ultimately, we need to hold all scientific evidence to the same high standard. Right now we're using a lower standard for the replications involving negative findings when in fact this standard needs to be higher. To establish the absence of an effect is much more difficult than the presence of an effect.

SIMONE SCHNALL is a University Senior Lecturer and Director of the Cambridge Embodied Cognition and Emotion Laboratory at Cambridge University.

[NOTE: For my talk at HeadCon '14, I explored my personal experience with a replication project. In the talk, I shared some reflections on this experience, and on how replication efforts are currently carried out. In this regard, I was talking to a group of colleagues in my field who are mostly acquainted with the relevant scientific issues*. Edge* readers unfamiliar with some of the discussion points can find further details on what some have called "replication bullying" in an article in *Science*. —SS]

MORAL INTUITIONS, REPLICATION, AND THE SCIENTIFIC STUDY OF HUMAN NATURE

I'm Simone Schnall. I'm a social psychologist at the University of Cambridge and I study all kinds of judgments, namely how people make judgments about the physical world but also about the social world. One thing I'm particularly interested in is moral judgments and how people's intuitions and feelings shape their moral judgments. At the moment, the field of social psychology is an interesting context in which to study people's judgments. There are all kinds of discussions going on, in particular about replication.

What's a replication? It sounds simple enough. You do the same study again and you see if you get the same result again. But when it comes to social psychology it's a little more complicated because what we usually do is to test a specific question with various different experiments. For example we've done some work with Jon Haidt, Jerry Clore and Alex Jordan to look at the influence of physical disgust on moral disgust. For example, we induce physical disgust by a disgusting smell and then look at participant's moral judgment and we find it makes them more severe.

Then we run other studies with other manipulations such as having participants sit at a disgusting table or watch a disgusting film clip and in each of those studies we find the same thing, that people make harsher moral judgments. These are called conceptual replications and in psychology, social psychology, we do them all the time. Usually we report them in the same paper.

Our entire literature is built on those conceptual replications, but those are not the ones that people are now discussing. They are different. They're called direct replications. The idea there is that you take an experiment in exactly the same way and repeat it with that precise method. A direct method, for example, would be to take that same study with a dirty desk and then again have participants complete a moral questionnaire.

That's different. That's what some people consider more valid in a way. They say it's similar to clinical trials in medicine or it's more similar to the hard sciences. But then of course if you think of the hard sciences, what they do is very different from what we do in social psychology because for example, they have a specific pill like 50 milligrams of Lipitor, and they look at the outcome in terms of people's blood lipid levels. It's very clear what needs to be measured: the pill and the outcome.

Whereas for social psychology, our outcomes and also our manipulations often are more complicated. There are many ways to induce disgust and there are many types of moral outcomes one can look at. And indeed people have looked at all kinds of factors when it comes to disgust and moral judgment and there's an entire literature based on those conceptual replications, even though nobody's ever done any given study twice. It's a bit of a different interpretation of what's considered a replication.
Intuitively it sounds like one would have to find the same result if one had an original finding, if the finding was true. But it turns out that's not necessarily the case at all and that's very counter-intuitive. This is a complicated story but there's a very good paper by David Stanley and Jeffrey Spence where they talk about the expectations for replications and they run computer simulations where they do experiments thousands of times under perfect conditions with nothing but measurement error. And even then one gets a great variability of results. The conclusion is that any one given study is not that

conclusive. That's why normally we do lots of studies to see if there is a general pattern.

One thing, though, with the direct replications, is that now there can be findings where one gets a negative result, and that's something we haven't had in the literature so far, where one does a study and then it doesn't match the earlier finding. Some people have said that well, that is not something that should be taken personally by the researcher who did the original work, it's just science. These are usually people outside of social psychology because our literature shows that there are two core dimensions when we judge a person's character. One is competence—how good are they at whatever they're doing. And the second is warmth or morality—how much do I like the person and is it somebody I can trust.

These are very relevant dimensions. Somebody's work is clearly relevant to how they're judged and how they perceive themselves. It's interesting to look at these direct replications and how they've been evaluated among colleagues and in the literature. It's an interesting situation because it points to the fact that people often use these intuitions that it seems like it's a really scientific way of confirming a previous finding when in fact that's not necessarily the case. In this context it's useful to think of how evidence is used in other contexts.

There was a really important paper in 1964 by Herbert Packer, a law professor. He made the distinction between two types of law. One is the due process model of law, the other is the crime control model of law. Due process is where the burden of proof always has to be very high. For example before you point the finger at anybody, you have to have some evidence to make an accusation. The law recognizes that if you were to just accuse somebody of something without any proof, that would be a crime. And the burden of proof also has to be really high before we as a society make a conviction. We usually consider any criminal innocent until proven guilty, so we're very careful how we assemble the evidence, what we examine and so on. And if we cannot make a conclusive judgment, we say that the person walks.

It's a very labor-intensive process and Packer calls it an "obstacle course" that ensures that we figure out the truth and don't convict an innocent person. In science we do something similar. We have an obstacle course where we consider lots of data, we run various controls, checks, we do all kinds of things and then I suppose our version of what's in the law the Fourth Amendment, is our publication ethics. These laws are our way of ensuring that editors cannot just decide on their own what they want to print but we have independent experts confirm the validity of the findings. This is our peer review system. The idea is that what we consider a verdict also has undergone due process.

In contrast to the idea of due process there is the crime control model of law. That's now very different. First of all, the burden of proof is much lower. That's the case when it comes to suspicions, where it's all about trying to look for suspicious activity and often it's not even clear what it is. For example in the former East Germany there was a system in place where each citizen had a file, their neighbors were spying on them, and everything was subject to monitoring.

It's all based on suspicions and there's a low burden of proof for the suspicion and then also a very low burden of proof when it comes to conviction. In that system, a person is assumed guilty until proven innocent, and the burden of proof is not very high. The goal is to convict people very quickly because there are so many suspicious people. Packer calls it an "assembly-line conveyor belt." The goal is the suppression of crime at any cost, to make sure that not a single criminal slips through. Of course that leads to some errors, and some innocent people get convicted but then, that's acceptable. The usual phrase is that "the innocent have nothing to fear", but, in reality, they have a lot to fear.

Crime control comes from crisis. In the United States, for example, some of these measures were put in place by the government in the form of the U.S. Patriot Act where various civil liberties were curtailed, where citizens were encouraged to be on the alert to look for any suspicious activity. It wasn't quite clear what to look for but one had to be careful anyway. It was because of that crisis, because of that horrible event that had happened that there was this unbelievable betrayal by somebody right within the community, in fact, several people who turned out to be terrorists and we had no idea.

In social psychology we had a problem a few years ago where one highly prominent psychologist turned out to have defrauded and betrayed us on an unprecedented scale. Diederik Stapel had fabricated data and then some 60-something papers were retracted. Everybody was in shock. In a way it led the field into a mindset to do with crime control, the sense that times are different now, we need to do things differently from what we used to do and we need to be more careful. We need to look for the fraudsters; we need to look for the false positives. And that has led to a different way of looking at evidence. This is also when this idea of direct replications was developed for the first time where people suggested that to be really scientific we should do what the clinical trials do rather our regular way of replication that we've always done.

Let's look at how replications are currently done, how these direct replications are carried out. First of all, there is no clear system in terms of how findings are selected for replication. At the moment, the only criterion is that a study has to be feasible, that is easy to conduct and that it's important, or rather the finding is important.

But then it's very hard to define what's important. In that sense, anything could be important and anything could be suspicious. What has happened is that some people have singled out certain findings that they find counterintuitive and often it's people who don't work in the research area, who wouldn't have any background on the literature or on the methods, but who nevertheless have a strong opinion that the findings somehow don't seem plausible.

There's been a disproportional number of studies that have been singled out simply because they're easy to conduct and the results are surprising to some people outside of the literature. It's unfortunate because there is not necessarily a scientific reason to investigate those findings, and at the same time, we know there may be some findings that we should go after more systematically but we really don't know which ones they are. There is no systematic effort to target specific findings.

There are also some issues with the quality of some of these replication projects. They're set up for very efficient data collection, so sometimes it really resembles that idea of an "assembly line conveyor belt" that Packer described for the crime control model where there's lots of data that's being collected even though it's not necessarily done all that carefully.

For example, there was a large-scale project called the "Many Labs" Replication Project. They went around the world and had various labs participate and rerun earlier studies. There was one original study conducted in the United States where participants had been presented with an American flag and they were asked about their attitudes about President Obama. Then Many Labs went around the world and presented participants in Italy, in Poland, in Malaysia, in Brazil and many other countries with an American flag and asked them about their views on President Obama. This is taking that idea of direct replication very literally. There have been other examples like this as well where it's not clear whether the kind of psychological process we're trying to capture is realized in that experiment.
The conclusions are also interesting and, again, it relates to Packer's idea of this quick processing of evidence where it's all about making the verdict and that verdict has to be final. Often the way these replications are interpreted is as if one single experiment disproves everything that has come before. That's a bit surprising, especially when a finding is negative, if an effect was not confirmed. We don't usually do that with positive findings. We don't say this now proves once and for all that such and such effect is real. It probably perhaps

comes with that idea that it intuitively seems like this is the real study because we repeated exactly what had been done before.

There's a number of problems with how these replications are done, but at the same time, some people feel very strongly that they are the only right way to basically confirm whether the effect exists. The studies usually are not so much about whether the effect is confirmed; it's more about whether that particular method got a significant finding. It's just that one example rather than the whole body of literature that's available. Now, one reason why some people feel so strongly about these direct replications is that perhaps they've taken on a moral connotation.

Linda Skitka has talked about moral conviction where people feel like they have a moral mandate where something is so important that it just by default has to be right: it's just a better way of doing things and that's basically the end of it. And when that happens, Skitka has shown that people feel like the regular rules don't apply. That has recently happened where there was a journal special issue with 15 replication papers covering 27 earlier reported effects. That issue went in print without having undergone any peer review.

It may not seem like a big deal but peer review is one of our laws; these are our publication ethics to ensure that whatever we declare as truth is unbiased. I took issue with the fact that there was no peer review and one of my findings was reported to not be replicated by some researchers. I looked at their data, looked at their paper and I found what I consider a statistical problem. What was really interesting though, was that when I alerted the editors, they were not very interested. They were not interested at all. In fact, they denied me the right to a published response. I had to fight tooth and nail to get that response.

And at every step of the way I was made to feel like whatever I could possibly say must be wrong. Mind you, that was without that paper nor any of the other papers, having gone through peer review. When that whole thing became public, it was interesting to observe people because one thing I pointed to was this idea of replication bullying, that now if a finding doesn't replicate, people take to social media and declare that they "disproved" an effect, and make inappropriate statements that go well beyond the data.

But that was not the main point about the bullying. The much more serious issue was what happened in the publication process, because again this is the published truth. Of course it's easy to say, and some people did say that peer review is not always accurate. Some reviewers make mistakes and anyway, maybe it wasn't such a big deal that there was no peer review.

But again let's think about it in the legal context. This is to declare a verdict on people's work, on the quality of people's work, without a judge and without having given the people whose work is concerned any right to even look at the verdicts, never mind to defend themselves.

Interestingly, people didn't see it that way. When I raised the issue, some people said yes, well, it's too bad she felt bullied but it's not personal and why can't scientists live up to the truth when their finding doesn't replicate? It was quite interesting just to see how people arrived at those judgments because it's ultimately a judgment of wrongdoing because it is personal. If my finding is wrong, there are two possibilities. Either I didn't do enough work and/or reported it prematurely when it wasn't solid enough or I did something unethical. It's also about allegations of wrongdoing. It's quite interesting how quickly people made those allegations.

People really didn't fully appreciate what it means that there was no peer review. Some people raised various general points such as "we have all these findings in the literature that don't replicate, so that's why we must do replications". All that is true but then of course I don't know how I can help with that because so far I don't know of a single person who failed to replicate that particular finding that concerned the effect of physical cleanliness and moral cleanliness. In fact, in my lab, we've done some direct replications, not conceptual replications, so repeating the same method. That's been done in my lab, that's been done in a different lab in Switzerland, in Germany, in the United States and in Hong Kong; all direct replications. As far as I can tell it is a solid effect. But people just repeat the mantra of well, it's important to do direct replications, we need to do them more often and so on. They go by the intuition that if there was a study with a large sample that repeated exactly the same method, it must be the right study, the ultimate study, when that is not necessarily the case.

What happened on social media was also interesting because the whole thing played out quite publicly and there were various heated discussions and at some point some people said oh, but what we really need to look at the data. I had been required to make all my raw data available so people were crunching numbers and there were blogs with all kinds of colorful pictures. At the end all the blogs concluded Schnall is definitely wrong. She is definitely wrong about that claim that there's a concern about her replication finding, no, there absolutely is not.

That was then called "post publication review" by the editors when in reality those self-appointed reviewers neglected to do the main part of their assignment, which was to evaluate the quality of the replication paper; in particular, the rejoinder where I am in print accused of hunting for artifacts. In terms of what was considered due process, nine months after I raised the concern that there was no peer review and although I found a technical problem in the replication, there still hasn't been any independent review. And that's not just for myself but that's for a total of 44 colleagues who all now have the label of "failure to replicate." There was no independent verification; there was no judge for that verdict.

Such judgments are made quickly nowadays in social psychology and definitively. There are now news reports about so and so many findings replicated, so many findings did not replicate when in reality it's usually a single experiment and nobody mentions all the conceptual replications which are part of the literature. When one looks at how these replications are done, they have a lot of the features of the crime control mindset, so there are no clear criteria for what's suspicious. We don't know what a false positive looks like or what we're looking for.

Then the quality criteria are oftentimes not nearly as high as for the original work. The people who are running them sometimes have motivations to not necessarily want to find an effect as it appears. We now have all these findings that didn't go through any peer review and yet there are exaggerated claims of what they can tell us.

When crime control is implemented by governments, it's a means of control, and it creates fear. And it is used in times of crisis. It's the kind of situation where people just aren't sure what's happening and they worry that they may become a suspect because anybody can become a suspect. That's the most worrisome thing that I learned throughout this whole experience where after I raised these concerns about the special issue, I put them on a blog, thinking I would just put a few thoughts out there.

That blog had some 17,000 hits within a few days. I was flooded with e-mails from the community, people writing to me to say things like "I'm so glad that finally somebody's saying something." I even received one e-mail from somebody writing to me anonymously, expressing support but not wanting to reveal their name. Each and every time I said: "Thank you for your support. Please also speak out. Please say something because we need more people to speak out openly."

Almost no one did so. They all kept quiet and they say they can't afford to speak out; they can't afford to question the replication movement because they don't have tenure yet, they don't have jobs yet and they can't afford to become a target. That's really the worrisome thing here, that we have created a system where there's just so much uncertainty or so much variability regarding what's done that people probably are not as much afraid that their findings don't replicate, as they're afraid of the fact that there's absolutely no due process. Anybody could be singled out at any point and there are no clear criteria for how the verdict is handed down.

That's a real problem and of course one could think that well, when it comes to governments that implement crime control, sometimes in times of crisis it can be useful. For example, if you have to be so sure that a particular person doesn't blow up a building or an airplane and you have good reason to believe that that might happen, it may still be useful to detain them even if it's wrong, if it's an error, to do so just to be on the safe side. When it comes to crime control it can be good to be on the safe side as far as criminals are concerned. But if we have that kind of crime control mindset when science is concerned, that's never a good thing because it comes with errors. We'll have errors across the board. We have them regarding our false positives and our false negatives. We'll just have a bunch of errors. And now we already have them in the literature in that particular special issue. Even the so-called "successfully replicated" findings have errors.

It's a problem for the accuracy of the published record because those are our verdicts. Those are what researchers build on. The whole idea was to increase the credibility of the published record. It's also a problem for all the people who put in the hard work running the replication studies and doing exactly what was expected of them and they now end up with publications that are not very valuable on a scientific level.

What social psychology needs to do as a field is to consider our intuitions about how we make judgments, about evidence, about colleagues, because some of us have been singled out again and again and again. And we've been put under suspicion; whole areas of research topics such as embodied cognition and priming have been singled out by people who don't work on the topics. False claims have been made about replication findings that in fact are not as conclusive as they seem. As a field we have to set aside our intuitions and move ahead with due process when we evaluate negative findings.

If junior people in the community aren't comfortable joining the discussion, then we have a real problem. If they're too afraid of being targeted for replication simply because it's not clear what's going to happen once their findings are under scrutiny, we really need to be careful. I appeal to colleagues to say, look, we often use intuitions. We do it all the time. But we know from the research that that's not the way to make a good decision, a good judgment. And we should treat our findings and our colleagues with at least the same respect that we give to murder suspects. We hear them out, we let them talk, we look at the evidence and then we make a decision.

THE REALITY CLUB

FIERY CUSHMAN: One of the things I really appreciated that you brought up is what is the appropriate analogy between science and law and the kind of standards and due process that get used in law. The analogy that you invoke is to criminal law where at least in the United States the standard of evidence is beyond a reasonable doubt, which is a very high standard of evidence. And criminal law not entirely, but mostly deals with intentional harms, what would be the equivalent in science would be intentional fraud.

For instance if I were to accidentally spill my coffee on Laurie, that would be handled through torte law where a different standard of evidence is applied. It's a preponderance of the evidence so … the idea is Laurie's got a claim, I've got a claim and 51 percent in favor of either one of us is going to decide the matter. Another interesting analogy would be to the area of libel law.

There's a concept in U.S. law that different people who occupy different roles in society are held to different standards in terms of when they can make a claim of libel against somebody else. As a private citizen, there's a fairly low standard. I can make a claim that someone's libeled me or slandered me in a broader array of circumstances. If I'm a public official, especially an elected official, then it's incredibly hard for me to make a successful claim of slander or libel. And the reason is because the … legal scholars and justices have interpreted the Constitution to imply that if you put yourself out in a public arena, then you're opening yourself up to criticism and the existence of that criticism is vital to a well-functioning democracy.

SCHNALL: Sure, that's right.

CUSHMAN: I'm curious to hear more from you about what are the appropriate analogies within the law and what kind of standard of due process are you envisioning? I think you brought up one issue that is whether or not a replication paper should be subject to peer review. I wouldn't be surprised at all if every person in this tent right now feels strongly that any publication in the literature needs to be subject to peer review. Are there elements of due process that go beyond that where extra scrutiny is required for replication …?

SCHNALL: Well, I will say this. We know how easy it is for any study to fail. There is almost no limit to the reasons for a given experiment to fail and sometimes you figure out what the problem was, you made an error, there was something that you didn't anticipate. Sometimes you don't figure it out. There are always many reasons for a study to go wrong and everything would have to go right to get the effect.

We have to apply a really high standard before we infer that there is no effect. In a way, before you declare that there definitely is no effect, the burden of proof has to be really high.

CUSHMAN: And do you think that burden of proof is most appropriately applied by the action editor and reviewers or by the readership?

SCHNALL: Based on the paper by Stanley and Spence I mentioned earlier, the conclusion is that there's very little you can say based on a single study. Practically all the large-scale replication projects that are being conducted now such as the Reproducibility Project, they will say very little about the robustness of the effect because it's just a one-off experiment. It's practically impossible to read much into that one experiment. And we usually don't do that, either. That's why we usually have a line of work rather than one single one that we consider conclusive.

It's about doing lots of studies as we've always done and getting at that effect from different angles rather than putting the weight into that one study. Just because it's the exact repetition of an earlier method and just because it has a large sample doesn't mean it's the conclusive study. In a way, that's really a misconception at the moment where people think that's the best kind of study to run when in fact it's not.

DAVID PIZARRO: There is a way, in which as you point out, social psychologists are well suited to see the problems in the scientific process because as you say, we know very well that people use evidence in very different ways depending on what their motivations are. And you rightfully point out a lot of the issues with people motivated in this way. It's not … if I wanted to show that say somebody's studies were wrong I could just do a poor job. I could claim to replicate Laurie's monkey studies and since I don't know anything about doing it, that's problematic.

But at the same time so much of what we know is that we have equal errors on the side of trying to find what you're convinced about. This has been problematic and there are reasons why this might be magnified in psychology right now or maybe across all sciences right now because of the ease with which we communicate. This has always been problematic. There's this way in which you can have an ideal answer where you say that … science corrects itself. You're motivated to find this, I'm motivated to find that, and at the end of the day it'll work itself out because there will be this body of evidence.

But the truth of the matter is that people get trampled in the process. I can imagine that … I don't know, if Newton and Leibniz had Facebook there would be flame wars and people would take sides. And there's a way in which this is just extra-problematic now. But it's not a new problem, at least in principle. I don't know if there are any good ideas aside from, say, just being more rigorous about publishing, about how to go about fixing given the ability to smear reputations in the modern world.

SCHNALL: Sure. Well, one key thing is to select, to really go after phenomena or methods for which we have evidence that they're in some way not as reliable as we hoped they were.

PIZARRO: But how do you get that evidence?

SCHNALL: Well, that would have to be something that the field as a whole decides; right?

PIZARRO: It's a problematic first step …

SCHNALL: Well, the way right now is that you can go to a website and anonymously nominate a replication target. One doesn't have to give any evidence of why it makes a good replication target except that it needs to be easy to run and important. Basically there are no criteria. That's the problematic thing because if we want to go in depth into a specific phenomenon, we need to do that, rather than just covering lots of different things and doing a one-off study that will tell us very little.

That's really a key thing for the field to decide, how to select replication targets because it does come at a huge reputational cost. There is no question about it but at the same time it needs to be done. We need to go after those potential false positives.

About your earlier point about people's expectations, one can always have biases this way or the other way. I would imagine just considering how easy it is for any given study to go wrong, that it's easier to get a study not to work than for it to work just based on bias. That's just a hunch.

LAURIE SANTOS: Let me follow up on that. One of the things I haven't liked in following the replication crisis is the fact that these effects are seen as either/or. Like either having a dirty table is going to cause moral evaluations or it won't. And in psychology the depth of our effects and the amount that they're going to stick varies.

I could have all kinds of preconceived notions that Müller-Lyer is a fake effect and it doesn't work. I'm going to bring 20 people in and they're going to see it. Social psychology is not as profound as perception that is why it's probably a lot more interesting. But it raises the question of can you use these null effects to see the boundary conditions on these things?

You brought up the case of the many labs doing, looking at the American flag in Brazil. Probably a boundary condition on the American flag effect is you have to be American. When we don't see replications there, we learn something about the effect. That it’s bounded—we see it only in Americans.

Is there a way to move the debate closer to that, that we can learn something important and scientific about the boundary conditions of different effects through non-replications? And I see the issue if somebody is doing what Dave was going to do in my monkey study, he's just actively and intentionally doing a crappy job. I'm not sure any of the cases are like that. I would like to believe that most of the cases are people who are curious about whether these effects….

SCHNALL: Yes, that's right. The issue is with these one-off experiments where a single experiment is taken to be representative of the effect as a whole as opposed to just one particular method. And that's what we normally do anyway with the conceptual replications. We do a line of studies. If one wanted to really go after specific false positives by using direct replications, one would have to use a series of direct replications rather than a one-off across a large number of phenomena.

SANTOS: But do we still learn something from the direct … I run a study. Somebody directly replicates it and they don't see the same result.

SCHNALL: Again, that paper by Stanley and Spence, that's an excellent paper. It's quite stunning just how much variability one can get. For example, they did computer simulations with perfect testing conditions, thousands of simulations, nothing but just measurement error. If you have a known correlation coefficient of 0.3 with a known reliability of 0.7, what you get as a correlation coefficient can range from 0 to 0.5. It can be much, much smaller and much, much larger than the real thing, 0.3. And there's just such a big range that any one given study tells us very little.

L.A. PAUL: It seems to me that it might be productive to distinguish a couple of things. It's my job as a philosopher anyway. And I heard you talk about a couple of things that I wanted to sort of separate. One issue involves the evidential standard, namely; what's the standard that our evidence has to reach before we draw a conclusion? Another thing though that I was hearing you talk about is what counts as evidence? And then embedded in that is also a question about, if we establish a particular evidential standard, do we … are we keeping it constant as we move from context to context? And so one reason why it's helpful to distinguish these things is because I don't think that you want to argue, that's ok if my evidential standard is low and other people's evidential standard is high.

SCHNALL: Sure.

PAUL: Rather, what I hear you talking about are two problems. One is that we're not being careful enough about what counts as evidence and so the quality of the replication studies must be looked at to understand whether or not we should even judge these results as evidence. And then the second problem is that it seems like we are holding different people and different groups to different standards. And for high quality scientific research and inquiry, we need to have a constant standard. It's a little bit related to what Fiery was raising.

SCHNALL: Yes. That's exactly my point. First of all, we consider: Is it admissible evidence? For example, I have had people write to me, ask me for my experimental materials. They take materials that were done in a lab. They run an online study then they don't find the effect and they make a big deal out of it on social media and blogs and so on, failure to replicate an effect. Well, is that something that should be put out there because it in no way repeated the original study? It was a lab study as opposed to an online study. It's obvious why they're not as highly controlled in each case.

What is considered admissible evidence? And the line is very easily blurred for two reasons when it comes to social psychology. One is that there are all these discussions on social media. The reason why they're everywhere is because social psychology seems so intuitive to people. Everybody has an opinion: Do I believe that finding, yes or no? As opposed to string theory where people will accept that they just don't know enough about it.

It's a real problem that people feel they know enough about it to say that yes, it's probably true or not. It's just an intuitive judgment as opposed to a scientific judgment and that's a big problem now with social psychology where everybody feels like they can be a social psychologist and make conclusions and put up studies online or … especially now with some of these replication efforts that don't require any expertise, so in a way they propagate that image that anybody can sign up and run a study.

In the end, it's about admissible evidence and ultimately, we need to hold all scientific evidence to the same high standard. Right now we're using a lower standard for the replications involving negative findings when in fact this standard needs to be higher. To establish the absence of an effect is much more difficult than the presence of an effect.

David Rand: "How Do You Change People's Minds About What Is Right And Wrong?"

There are often future consequences for your current behavior. You can't just do whatever you want because if you are selfish now, it'll come back to bite you. In order for any of that to work, though, it relies on people caring about you being cooperative. There has to be a norm of cooperation. The important question then, in terms of trying to understand how we get people to cooperate and how we increase social welfare, is this: Where do these norms come from and how can they be changed? And since I spend all my time thinking about how to maximize social welfare, it also makes me stop and ask, "To what extent is the way that I am acting consistent with trying to maximize social welfare?"

DAVID RAND is Assistant Professor of Psychology, Economics, and Management at Yale University, and Director of Yale University’s Human Cooperation Laboratory.

HOW DO YOU CHANGE PEOPLE'S MINDS ABOUT WHAT IS RIGHT AND WRONG?

I'm a professor of psychology, economics and management at Yale. The thing that I'm interested in, and that I spend pretty much all of my time thinking about, is cooperation—situations where people have the chance to help others at a cost to themselves. The questions that I'm interested in are how do we explain the fact that, by and large, people are quite cooperative, and even more importantly, what can we do to get people to be *more* cooperative, to be more willing to make sacrifices for the collective good?

There's been a lot of work on cooperation in different fields, and certain basic themes have emerged, what you might call mechanisms for promoting cooperation: ways that you can structure interactions so that people learn to cooperate. In general, if you imagine that most people in a group are doing the cooperative thing, paying costs to help the group as a whole, but there's some subset that's decided "Oh, we don't feel like it; we're just going to look out for ourselves," the selfish people will be better off. Then, either through an evolutionary process or an imitation process, that selfish behavior will spread.

The question that has preoccupied people for a long time is "How do you stop that from happening?" There are a lot of good answers. For example, if you interact repeatedly with the same person, then that changes things. If the other person has a strategy where they'll only cooperate with you tomorrow if you cooperate with them today, it becomes in your self-interest to cooperate. Or, if people can observe what you're doing, you'll get a reputation for being a cooperator or a non-cooperator. And if people are more inclined to cooperate with people that have cooperated in the past, then that also creates an incentive to cooperate. Or there is partner choice—if people are choosing who they want to work with, who they want to interact with, then if they're more likely to choose cooperative partners, that creates an incentive to cooperate.

What all these different mechanisms boil down to is the idea that there are often future consequences for your current behavior. You can't just do whatever you want because if you are selfish now, it'll come back to bite you. I should say that there are mathematical and computational models, lab experiments, and also real-world field experiments that show the power of these forms of accountability for getting people to cooperate.

For example, we did an experiment with a utility company in California. We were trying to get people to sign up for a blackout prevention program, where they let the utility company turn down their air conditioners a couple of degrees on really hot days so there's not a big spike in energy use which can cause blackouts. It's a great program, but nobody signs up because it's a pain: You have to be there when the guy comes to install the device and so on. We found that if we made the sheet where you signed up to be part of the program public, so that you had to write down your name and your unit number on the signup sheet instead of just a random code number, it tripled signups. This was a field study with over a thousand Californians. These effects matter in the real world. They're powerful. There's no question that these reputational effects can be powerful motivators of cooperation.

In order for any of that to work, it relies on people *caring* about you being cooperative; people have to care that you do the right thing. There has to be a norm of cooperation where people think it is acceptable to do what's socially beneficial, and that it's not acceptable to do things that are not socially beneficial. These observability mechanisms don't work in situations where that's not true.

There are cross-cultural experiments, for example, where people play cooperation games with the option to punish other players. If you run that setup with Harvard or Yale undergrads it works great: Most people cooperate and punish those that don't cooperate, so everyone learns to cooperate and it's a nice, happy outcome. If you go to places like post-Soviet Russia or Eastern Europe, or the Middle East, and run these experiments, you get very different outcomes. People often don't support doing the cooperative behavior, the thing that is collectively beneficial; they wind up punishing the people that are being cooperative. In those places, it's worse to have punishment and accountability than to just let everybody do whatever they want anonymously.

The important question then, in terms of trying to understand how we get people to cooperate and how we increase social welfare, is this: Where do these norms come from and how can they be changed? By norm, I mean a person’s internalized sense of what's appropriate, what's acceptable behavior and what's unacceptable behavior. That is, your moral values: what you believe inside is right, rather than what you do because you're forced to do it under threat of punishment or exclusion

There are certainly examples in recent history where deeply-held norms have changed dramatically—attitudes toward smoking, driving while drunk, gay rights—these are cases where we've seen massive shifts in the U.S., for example, in people’s opinions about what is right and wrong.

It's not at all clear, however, how that happens or where that sense of right and wrong comes from. This is what I've gotten interested in, and what I’ve started to spend a lot of time trying to unpack. Where does our sense of right and wrong come from? A general framework for thinking about these questions that I've been using comes from the study of judgment and decision making: the idea of heuristics.

Rather than carefully thinking about all the details every time we're confronted with a situation, and then asking ourselves, "All right, what is optimal here?" we sometimes use rules of thumb. If we’ve been in a similar situation before, then the behavior that typically works well in that situation can pop into your head. This heuristic rule of thumb can often work out, but sometimes isn’t perfectly matched to the current situation you’re facing. It may be that if you stop and think more carefully, you might be like, "Oh, my heuristic doesn't work that well in this specific setting, I should do something different." This process is talked about a lot in the domain of individual choice, like risking taking and impulsivity. But it seems to me that it is equally important in the domain of social interaction.

What this perspective would suggest is that the thing that you internalize, that you get used to as your way of being in the world, is the thing that typically works well for you in your daily life. If most of the time you live in a setting where you're rewarded for being cooperative and you're punished for being selfish, you wind up getting in the habit of cooperating. You internalize cooperation as your default response. Then when you find yourself in a situation where you could exploit someone without any consequence, your first impulse is to keep treating it as if it was daily life, where you shouldn't exploit the person or you're going to get in trouble. But if you stop and think about it, you might be like, "This situation is different. My first impulse to cooperate isn't optimal for me in this particular situation."

We've done a lot of experiments to try and provide evidence for this. Because I come out of a behavioral economics/experimental economics background, I like to use economic game experiments where you give people actual money and then let them choose how much to keep for themselves and how much to contribute to something that benefits other people (rather than just hypothetical surveys). We've done a number of experiments that find for most of our subjects—who are usually American and come from relatively safe daily lives where they trust other people, people don't exploit them, and it's a good idea to be cooperative—their default response is cooperation. If you make them stop and think about it, they get less inclined to spend money to help other people.

If this idea about social heuristics is right, it's not a story about our evolutionary past where we had to take care of each other so we devolved some gene that makes us, by default, be cooperative. It's a story about learning and culture. If you come up in a place where it's not a good idea to be cooperative, either because the norms of the people around you are bad or the institutions are corrupt, then you should wind up internalizing something different as your default—not cooperation but selfishness. We find evidence of this: What is intuitive to people varies depending on their experience of the world. People that experience the world as a trustworthy place are often intuitively cooperative, but people that experience the world as an untrustworthy place are intuitively selfish. It means that the way you experience the world has broad implications for your behavior.

We've done other experiments to try and directly demonstrate this effect of experience. We first have people interact for 20 minutes either under a set of rules that makes it a good idea to cooperate, so that they learn to cooperate and spend 20 minutes cooperating; or under a set of rules where it's a good idea to be selfish, so that they learn to be selfish and spend 20 minutes not cooperating. Then we have everyone do an identical battery of one-shot anonymous interactions where in some cases they can pay to help other people, or in others they can pay to punish people for being selfish.

What you see is huge spillover effects, where the habit that gets established in first part where we manipulated the rules then spills over to the subsequent anonymous interactions. If you get people used to not cooperating, they wind up being much less altruistic, less trusting, less trustworthy, less optimistic about the behavior of others, and less inclined to sanction other people for being selfish.

What this means is that when you think about what people have as their sense of right and wrong, where these values come from, I would argue that at least a good chunk of it comes from what they are experiencing in general. That means that you can have a hard equilibrium problem. If you're in a setting where the norms favor selfishness, that's reinforcing default selfishness. And then the opposite, if you're already in a good situation where people have good norms and are cooperating, it reinforces cooperative defaults.

In terms of what you can do to change the norms in a setting where they're not good, although I don't have an awesome answer, our research suggests that top-down institutional rules can play an important part. Say you're managing an organization. If you set up rules within your organization to reward people that behave cooperatively or punish people that behave selfishly, that can change what is optimal behavior in the context of your organization. That can generate a culture—an institutional culture—within that institution. Potentially, that can also have a spillover effect, where it not only affects people's behavior in this particular context but also affects people's behavior more generally.

In this vein, I like the idea of Paul Romer's model cities. You go to a place where the institutions or the norms are bad and you say, "In this one setting the institutions are going to be good. There are going to be rules that incentivize good behavior and only come and join this if you want to play by those rules." You can create a new culture there that, hopefully, people take with them when they leave, and also inspire other people who see the benefits.

The general idea I'm arguing for is that the top-down rules you establish can have a profound impact not only on the way people behave when the rules are watching them, but also on what they internalize and what they carry with them when they're interacting outside of the rules.

If you think about it in an institutional setting—in the context of a company for example—you can have incentives to get people to be cooperative and reward people for the outcome of the team as well as their individual outcome. Not only do you get them to behave well in those contexts, but it also creates a general culture or set of norms where people are more likely to help each other out even in settings that are not explicitly going to get rewarded by the company. If you get people in the habit of having positive, constructive, cooperative interactions with others, they're more likely to do those cooperative actions in settings that are not governed by the official rules.

There could also be a broader generalization where if you get used to operating that way in the context of your organization, you also carry that with you to some extent when you go out into the world. This is potentially a tool for public policy and institution design but it's also something to think about when we think about certain organizations or industries that are explicitly built around self-interest as their cornerstone. That given of institutional culture can have consequences for how employees behave more generally. There are all these interconnections between institutions and norms that are important for considering.

That's not to say that institutions are the only thing that matter for what we think is right and wrong. It's also important to think about bottom-up change, like the examples about attitudes toward smoking or drunk driving or gay rights. It's not at all clear to what extent those changes were the result of big national advertising campaigns and attempts to change people's understanding from a top-down perspective, and how much it was just some organic process of change occurring among individuals convincing each other that things should be different. We're also interested in this bottom-up change and how you can be a moral exemplar to people around you.

How do you get people around you to be more pro-social, to do what they know is right? Maybe by inspiring them. The other bigger question is how do you change people's minds about *what* is right in the first place? It's one thing to say, "I know I should be doing this but I don't feel like it. Oh, look, Laurie is a great role model. She's super-moral so she could inspire me to be moral like that, too." But it’s a bigger question to change people's minds about what's right in the first place? That is the most important question in any of this work related to cooperation and pro-sociality. It's something that there is not that much known about. That is, you can push people's behavior around but the question is how do you change their sense of what's right and wrong? I am hopeful that this is something that there will be a lot of interdisciplinary effort around.

Since I spend all my time thinking about how to maximize social welfare, it also makes me stop and ask: "To what extent is the way that I am acting consistent with trying to maximize social welfare?" As an academic, my life is fun and I get to do interesting, cool things all the time. But I don't know to what extent a lot of what I'm doing is working to improve social welfare.

This is potentially an opportunity to use the things that we're always studying, how to motivate people to maximize social welfare, to try and change norms within academia itself. It would be great if more people were like, "It is an important and valuable thing to do things that matter, that have some impact on the world and on trying to make the world better." There is, in general, a lot of looking down at research that is applied, and that is not socially optimal.

In some sense, it's exactly the same problem of what I was talking about before where there is a bad norm in place. It's pretty easy to observe the extent to which people are doing things which have some real application and impact on the world, versus not. But there's not a norm in place that says that's something that should be valued and rewarded. Even though it’s observable, there's no incentive to do relevant, applied work. In fact, if anything there's a disincentive to a substantial degree because it's looked down on.

This is an important thing for us to try and change, both because applied work has an important impact on the world, and that is a good, and also because doing work that feels meaningful is important to people. Purpose comes from feeling like you're doing something that matters and that helps people. You can argue about whether it's true or not, but companies like Facebook and Google have this as part of their pitch when they try and get smart, competent people that are finishing PhDs to go and work for them that rather than staying in science. Part of the pitch is, "Here's a way that you can do something real, something that interacts with the world." That puts an additional market-type pressure on academia to try and satisfy this dimension of people's lives in order to keep the smartest people in academia.

Cooperation is good; we can get people to cooperate if there are norms in place that support that cooperation. We should try and do some of that ourselves.

THE REALITY CLUB

FIERY CUSHMAN: You introduced yourself as an economist and a psychologist and both of those themes are reflected in your talk. The economist is taking a top-down approach, where you design an institution and don't worry about the psychological mechanisms. The institutional pressures and the cost-benefit structure are going to determine behavior by hook or by crook in the long run. The psychologist is going to be the bottom-up person who wants to understand the human mind so that you can see norms that are going to prosper and you can get people to take up norms in the most efficient way possible.

You ended with the idea that we want to change the world. If someone walked into this room and said, "I'm in possession of worldly power. I can only talk to Dave the economist or Dave the psychologist," who are you going to be? Just from what you said, Dave the economist is a much, much more attractive option because we can be formal about analyzing the institution designs from a game theoretic perspective, because we don't need to do a lot of hard empirical work to try to figure out how a human brain works. The guy with worldly power can just by fiat establish the institution, and then all of the people take care of themselves.

As someone who is a psychologist, it makes me question this other theme that you ended up with: should we, as psychologists, be trying to change the world or should we be doing basic science? Even if the economists completely dominate on the problem of making the world a better place, maybe irrationally, I'm still committed to the view that it would be exciting to learn how the human mind works.

RAND: I also, obviously, share that desire, which is why I'm doing what I'm doing. There are two parts to the answer. One is that I love figuring things out and the joy of discovery, and that is a lot of what motivates me to do what I do. That's basically everything that motivates me to do what I do. But I feel guilty about it in some sense, that I'm living my life doing this awesome, fun thing. Is this optimal? In terms of trying to make things better, understanding psychology is important.

There's also this whole literature on crowding out. If I give you an extrinsic motive to do something it can destroy your intrinsic motivation to do it. Then when I take away the threat of punishment, you're like, "Well, I'm not going to do that anymore," even if before I started threatening you to get you to do it, you did like it. That's an important psychological question of when do you get habit formation and when do you get crowding out? You need to understand a lot about cognition and about the human mind in order to sort those things out.

In addition to that blackout prevention project, we're doing a bunch of different experiments with the Department of Energy and different utility companies. Doing this, you realize that there are interesting psychological questions: how do you take this thing that works in the lab, in this cut and dry situation where people come into the lab and play a game with two options and two different payoffs, and you do this simple manipulation and it works very nicely, how do you translate that into the world in a way that works? In order to do that you have to know a lot about the psychology of people. Another way of saying it is, in the process of trying to figure that out you learn a lot about psychology and about what people care about and how these different factors interact with each other.

My goal is not to have part of my time spent doing the search for truth because it's fun, and then a different part, like, "Well, let's try and do something practical" but to find things that combine those elements that are interesting and that reveal things about how the mind works but in the service of trying to do something that has a real application.

MICHAEL MCCULLOUGH: An interesting case study in this that I've been fascinated with for a couple of years is the cultural evolution of social insurance—Social Security, guaranteed income. That happened at a place in time, it started in Germany. There seems to be two things that happened and a third thing that moved toward France and England. One was that there was a real concern in the government that socialism was so appealing that we have to do something about the glittering prizes that socialism offers or else that's going to, from the bottom-up, become what people want here in Germany. It was like, "Right, let's take away that tool and let's create a guaranteed social insurance. If people lose their jobs they'll have something to eat." Right? There was a strategic benefit to shifting the norm there.

RAND: But hold on. In that context, what you're saying is that the people running the institution made a specific decision in an effort to change or to control the way people's understanding of right and wrong was evolving.

MCCULLOUGH: Absolutely. You get the Leviathan to do the work for you. But there was another part of that, particularly in England, which involved a lot of debate and also coverage by newspapers. People could read the debate and process argument. It took years to get social insurance. It was very, very hotly contested. Once they do it, you roll the tape forward 100 years, 120 years, and now the norms are so strong among the rank and file, people gladly pay their taxes in order to provide certain benefits for everybody in society.

It's so powerful now that you can have a comedian like Rich Lowry, this prominent comedian in the U.K., who was engaging in legal tax dodges—finding ways to shelter his income from income tax. He was so browbeaten in the public eye for doing perfectly legal things. He had to come back and apologize for taking absolutely legal loopholes that he was entitled to. You can compare that against the norms in the United States where Romney was taking equally legal, equally plausible tax loopholes to shelter his income and he said, "I'm not at all apologetic," and he didn't need to be apologetic because there wasn't this groundswell of disgust in response to it.

RAND: Norms is a messy word that is used to mean a lot of different things in a lot of different settings. There is clearly a contextual element of it where there are certain things where this seems right—behavior like hugging each person you meet is acceptable or not depends on the context. There are some contexts where that would be weird and counter-normative and there are other contexts where it would be inappropriate to not do it.

There are other things that are more fundamental. Your basic values for example, the extent to which it is important to care about others or to do things that are good for society versus good for you as an individual. For one person you apply those kinds of basic morals across a range of settings and the contextually specific norms are implemented in light of these more basic moral principles.

Both of them are interesting but I am particularly interested in the latter. I don't think it stops science, because if all you did is said, "Oh, well, the reason it happens is because it's a norm," that is the end of the discussion. What I'm interested in is trying to understand where do they come from? What are they? And how do you change them? It could be a conversation-ender to say, "Well, it's the norm, so we do it." To me, that says that the interesting question is "What is going on there and how do you unpack that?"

SIMONE SCHNALL: You were asking the question earlier about how to get people to do the right thing, or how to change their norms. I was thinking of how to do that and I guess the biggest problem is opportunity costs. You do one right thing and you might not do another. Or you may have to decide how much time you spend doing one good thing or good things in general, let's say, positive things.

I was thinking about it when you described your job. You said, "Well, I have all this fun doing my job but I'm not doing anything good for the world." Maybe that's exactly the right way to think about it. You think you're having fun when in reality you're studying cooperation, which, if you make good progress, will have real life outcomes. As long as you think you're not doing anything morally good, you're just having fun, you'll be driven to do other morally good things. Do you see what I mean?

A "moral licensing" effect, for example, like, "Oh, I'm having so much fun, I'd better do something good as well." Perhaps one has to, in a way, restructure how people think about what is good or what is just having fun. Is it pleasant, unpleasant, this or that? As opposed to the knowledge that people only will do one good thing, or you have a limited amount of time, of energy, and whatsoever, so that perhaps one can repackage the things that they're doing and get them to do more of the good things.

RAND: Yes. You take things which are societally beneficial, good things that people are doing and get them to think of those things as fun things.

SCHNALL: You turn the morally good things into fun things so they keep doing those because they think it's fun. Then you get them to do additional morally good things on top of that.

RAND: Certainly it seems like being able to get socially beneficial behavior to be a thing that seems fun is a good idea. Like I was saying before, if you think about the two separate questions, where one is: how do you get people to do the thing that most people agree is the right thing to do? We have a pretty good handle on that in terms of all these reputation observability, reciprocity things. To me, the more challenging question is how do you change what most people think is the right thing?

MOLLY CROCKETT: I wonder if some of the challenge in establishing norms, particularly for cooperation, has to do with this distinction between getting people to do something versus getting people not to do something.

What I think about lately is harm and there are very strong norms against harming people and that's an easy norm to think about. Whereas, you harm people indirectly by not cooperating but because we have this distinction between actions and omissions it seems a lot harder to establish getting people to do something as a norm than prohibiting something.

RAND: That's interesting. It interacts a little bit with the question that John was asking, in terms of if you have this deeply-held underlying value that it is wrong to harm people, then when you're thinking about a individual specific context and asking, "Oh, what's acceptable or not acceptable here?", things that are functionally equivalent could provoke a much stronger reaction if they're cast as harm, because you have this underlying principle that affects how you structure or interpret these more contextual norms. Yes, that's interesting. The implication of that would be "Let's frame things as harm."

CROCKETT: Exactly. If you could somehow frame not cooperating as something that's harmful as an act, rather than an omission, that that would be a more powerful way to get people to internalize these sentiments.

RAND: Yes. Think about all the people we're harming by not doing applied science.

L.A. Paul: "The Transformative Experience"

We're going to pretend that modern-day vampires don't drink the blood of humans; they're vegetarian vampires, which means they only drink the blood of humanely farmed animals. You have a one-time-only chance to become a modern-day vampire. You think, "This is a pretty amazing opportunity, do I want to gain immortality, amazing speed, strength, and power? But do I want to become undead, become an immortal monster and have to drink blood? It's a tough call." Then you go around asking people for their advice and you discover that all of your friends and family members have already become vampires. They tell you, "It is amazing. It is the best thing ever. It's absolutely fabulous. It's incredible. You get these new sensory capacities. You should definitely become a vampire." Then you say, " Can you tell me a little more about it?" And they say, "You have to become a vampire to know what it's like. You can't, as a mere human, understand what it's like to become a vampire just by hearing me talk about it. Until you're a vampire, you're just not going to know what it's going to be like."

L.A. PAUL is Professor of Philosophy at the University of North Carolina at Chapel Hill, and Professorial Fellow in the Arché Research Centre at the University of St. Andrews. **L.A. Paul's *Edge* Bio page**

THE TRANSFORMATIVE EXPERIENCE

My name is Laurie Paul, and I'm a professor of philosophy at the University of North Carolina at Chapel Hill. I'm a metaphysician. I'm especially interested in metaphysics and philosophy of mind. I have been developing what I think of as formal phenomenology. In other words, I'm especially interested in looking at formal techniques engaging with the nature of experience, and I've paid special attention to temporal experience. One thing I've been thinking a lot about lately is the notion of transformative experience, which I'll tell you a little bit about today.

The questions that have been occupying me involve questions that come up when we as individuals think about making big life decisions. Metaphorically, it's when we think about making decisions when we're at life's crossroads. As we live our lives, all of us experience a series of these crossroad-style big decisions.

The worries or puzzles that I've been thinking about and exploring come from drawing together a number of strands in philosophy that haven't been drawn together before. The first strand involves a relatively new area of inquiry in philosophy that goes by the name of formal epistemology. It's an interesting and engaging new development, and formal epistemologists are interested in the way that individuals make decisions. They're interested in looking at formal decision theory, but they're interested in doing this within the context of epistemic questions. The thought is to explore how we can make rational decisions by taking agents to have psychologically real utilities or desires, by thinking in terms of particular degrees of beliefs or credences and also psychologically real preferences, and thinking about how we want, in an epistemic context, to think about individuals making decisions so that these individuals can know how they should act. The "should" is important; we're exploring these questions from a normative perspective.

I'm interested in normative decision theory, as opposed to behavioral decision theory. I'm interested in what the epistemic gold standard is that we as individuals should be aspiring to reach when we make decisions. In particular, what I'll talk about a little bit more concerns the normative gold standard for when we make important decisions.

The formal epistemologist usually thinks of the individual in a third-personal sense. Namely, it's as though we're observing individuals, and thinking about their epistemic states and how they're making their decisions. But there's another perspective that's also important and draws in another strand from philosophy, a strand of work that's been important over the last 30 or 40 years in philosophy. People like Dave Chalmers have made important contributions to philosophy involving the notion of consciousness and trying to understand what consciousness is.

What I want to look at closely is what philosophers have learned about the value of experience—how we've learned about what experience teaches us. A lot of times this discussion occurs in the context of worrying about the mind-body problem, or questions about physicalism. That's not my focus. I want to get a better understanding and think about how important it is, in some contexts, that we have certain experiences in order to know or understand certain information. There are disputes in the philosophical community about whether experience is required to know certain facts, or what exactly it is that experience teaches. I'm not worried about that dispute. I just want us to be able to see that sometimes experience is important. It's necessary, at the very least, for us to have conceptual or imaginative abilities in order to grasp certain kinds of imaginative content.

If we draw the strands together—formal epistemology, normative decision theory, consciousness with a focus on what experience teaches—then we get a different perspective on how we need to think about how we make big decisions. What I'm going to say is going to connect a little bit to what Molly talked about earlier today and some of the things that Josh Knobe talked about the last time we had this session. When each of you thinks about how you make a big decision, you need to consider how you—your current self—wants to perform some act or decide what to do in order to maximize the utility for your future self. The choices I'm especially interested in are ones that are life-changing decisions. As I said, they're high-stakes—the things that we care about very much.

What we ordinarily do is imagine ourselves into different possible scenarios: "Maybe I could do A, maybe I could do B, maybe I could do C. What should I do? How do I want to live my life? What kind of person do I want to be?" You can think of this philosophically as what kind of future self do I want to become? What kind of future do I want to occupy? I care about what it's going to be like to be me after I undergo this central experience that's part of this big decision. That's the question I'm interested in.

As a philosopher, it's kind of funny to tell a bunch of scientists about a fictional example. The reason why it's important to look at these fictional examples, or at least the one about vampires I'm giving you, is because the structure is present in a number of real-life cases. It's important to get the structure out there so we can understand it. We're not worried about questions about morality here. Obviously those questions are important, but as a metaphysician I don't think about morality, I lack the relevant expertise.

We're going to pretend that modern-day vampires don't drink the blood of humans; they're vegetarian vampires, which means they only drink the blood of humanely farmed animals. You have a one-time-only chance to become a modern-day vampire. You think, "This is a pretty amazing opportunity, do I want to gain immortality, amazing speed, strength, and power? But do I want to become undead, become an immortal monster and have to drink blood? It's a tough call." Then you go around asking people for their advice and you discover that all of your friends and family members have already become vampires. They tell you, "It is amazing. It is the best thing ever. It's absolutely fabulous. It's incredible. You get these new sensory capacities. You should definitely become a vampire." Then you say, " Can you tell me a little more about it?" And they say, "You have to become a vampire to know what it's like. You can't, as a mere human, understand what it's like to become a vampire just by hearing me talk about it. Until you're a vampire, you're just not going to know what it's going to be like."

The question you need to ask yourself is how could you possibly make a rational decision about whether or not to become a vampire? You don't know, and you can't know what it's like. You can't know what you'd be choosing to do if you became a vampire, and you can't know what you're missing if you pass it up. This would be a problem if we faced these choices on a regular basis because what it suggests is that there is a principled, philosophical reason why, when faced with this big choice, we would be unable to reach our epistemic gold standard.

If that were the only case in which this situation arose, most of us probably wouldn't have to worry about it, but I don't think it's the only situation in which this kind of thing arises. Now I want to talk about a case that is different in important ways from the vampire case because it's a low-stakes case. It's a little closer to real life, so we can see how this philosophical problem is one that we grapple with even if we're not always recognizing that we're grappling with it on a regular basis.

I've never tried a durian fruit. If you've tried a durian before, then bear with me. You can probably remember back to before you'd tried durian, and for those of you who haven't tried durian, we're in the same epistemic boat. The thing to know about durian is it's an exotic Southeast Asian fruit; it's very distinctive. One important chef says, "The only way to describe its taste is 'indescribable.'" The thought is, until you've tasted a durian fruit, you can't know what it tastes like. There are various evocative descriptions people have: "Eating vanilla ice cream by a sewer" or "French kissing a dead rat." These evocative descriptions are interesting, but they're not going to give you the information that you might like to have, namely, what it's like to taste a durian. The only way that you can know what it's going to be like for you is to taste one.

It's not about being sophisticated or liking exotic things because, as I already mentioned, even those with sophisticated palates, like chefs, differ widely on how they respond. Some people find it absolutely repulsive; other people call it the king of fruits. Ambrosia would be the description. In this situation, when someone asks, "Do you like the taste of durian?" you don't know. You would have what I think of as an epistemic transformation if you tasted durian. Once you taste durian for the first time, you know what it's like.

The philosophical example in the literature that parallels this, about the value or what experience can teach you, is an example that was developed by Frank Jackson. He talks about black and white Mary. Mary, we suppose, has grown up in a black and white room. She's never seen color, she's just seen shades of gray and black and white. When she's finally let out of her black and white cell and sees a red fire engine, she learns something. She learns something that she couldn't have learned by reading all the literature about color science or about how we see or hearing testimony of other people. She learns what it's like to see red. The thought is that we can all recognize that there is something important that we gain by experience and by experience alone. We gain an ability to grasp a certain phenomenal concept. We gain a certain imaginative ability—the ability to imagine redness in various contexts.

This is important because if we think about what experience teaches us, then we can see how the puzzle that I was sketching with the vampire comes up again in the case of the durian fruit. Imagine that you're in Thailand. It's breakfast time, you're looking at the menu and you're trying to decide what you're going to have for breakfast. You have a choice between having some ripe pineapple for breakfast, or having some ripe durian. I'm going to assume we've all had ripe pineapple, and let's just assume you like pineapple, you think it's pretty good, but you've never had ripe durian.

The problem, when you're looking at your breakfast menu, is that you can't make a decision about what to have for breakfast based on which taste you prefer. Why? Because you've never had durian. You can't assign a value to the outcome of what it's like for you to taste durian. In a certain sense, the utility of that outcome is not defined. If that's the case, then there's no way to make sense of determining how best to maximize your utility, or how best to respect your preferences in terms of picking whatever you would like best to have for breakfast that day. Because you can't assign a value to what it's like for you to taste durian, you can't, in a sense, have a preference, at least based on the way that we're thinking of the options. You can't step back and think about what the epistemic gold standard would be for you, that you should apply to yourself, when you're thinking about how best to choose what you want to have for breakfast.

When we're in context where we face epistemically transformative experiences, there's a way to make the decision that's just not accessible to us because we lack certain information, or we lack a certain ability. Why does this matter? As I said before, one way in which we assess our different options is by imaginatively projecting ourselves forward into different possible scenarios: "There's me having durian for breakfast," or "There's me having ripe pineapple for breakfast." We decide which scenario meets our desires in a more satisfying way, which scenario we assign a higher utility, then act so as to maximize that utility. That's the gold standard route.

In a low-stakes case, like deciding what you want to have for breakfast, there are other things we might want to do. We might say, "I value discovery. I'll just flip a coin. I'll just try durian for the heck of it. It's not a big deal." That's just fine in low-stakes cases. What matters are high-stakes cases. The vampire case I was describing to you is a high-stakes case. What makes it high-stakes is that it's both epistemically transformative and personally transformative. It's the personal transformation, the fact that it's going to affect the rest of your life and your very being, that makes it important.

These high-stakes cases are the cases where we care most about meeting the epistemic gold standard, or at least we should care most, because the decision has big effects on you or maybe your loved ones. If any of those personally transformative decisions are also epistemically transformative, then the same problem we faced with the durian case resurfaces with the big decision case.

There are some real-life cases that have this structure. Let me just sketch two. The first case involves sensory capacities. Imagine a congenitally deaf person who has never been able to hear contemplating whether or not he should have a cochlear implant. Let's say he's built a lot of his life around being a member of the Deaf community. He's contemplating the possible outcome of getting a cochlear implant, and then presumably after he's learned how to interpret the signals from his implant, knowing what it's like to hear. Because he doesn't know what it's like to hear he can't, in principle, know what it's like to hear until he becomes a hearing person. There's a certain sense in which he can't assign a value to the outcome of what it's like to hear. It's a high-stakes case because, presumably, what it's like for him to hear is going to have a huge effect on the way he lives his life, and a lot of the features of his life.

It's not a matter of thinking more carefully or reflecting in a deliberate manner. For principled reasons, there's something that's epistemically inaccessible to him, and we can't expect him to make a decision based on information that he can't have access to. There has to be a different way to make that decision. Part of what I want to say is we need to recognize that agents can find themselves in that kind of epistemic situation.

There are lots of other cases involving disability and similar issues, but there's another case that is maybe a little more familiar to those of us who never had to face the possibility of having a cochlear implant: The choice of whether or not to have one's first child. Having one's first child is also an epistemically transformative experience. One of the most important and salient features of becoming a parent is what it's like to experience the attachment to the actual child that you produce—the loving, satisfying, attachment relation that you stand in to the child that you produce. In order for you to stand in this attachment relation, first you have to produce the child. Second, the character of that attachment relation is going to be highly defined by the particular characteristics of the actual child that you produce. Until you stand in that relation, you can't know what it's like. You might know some very general features, but it's the particular features that matter and that are going to have the biggest impact on your experience of being a parent.

When you make the choice or you think about whether or not you want to become a parent, and you cognitively evolve yourself forward and imagine holding your baby and what it would be like to be a mother or a father, performing that act might be an interesting exercise in imaginative fiction, but it's not going to give you information about what it's going to be like for you to become a parent. That means that the utility of that outcome is not defined for you. And of course, this is a high-stakes decision. Becoming a parent is one of the classic cases where people's preferences and other things about their situation change dramatically. Often people do take themselves to be a different person. Some people say they're less selfish, they care about different things, they don't party as much. There are lots of different things that happen.

This is another case, one that many people face, and that is when they think about whether or not they want to have their first child, there's something important that's epistemically inaccessible to them. It's the thing that we care about, and the thing that's going to personally transform you if you have a child. When you contemplate whether or not to have a child, if you want to do it by assessing what it would be like for you to be a parent then it's, in principle, not possible for you to make that decision while reaching the epistemic gold standard. In other words, by acting so as to maximize your utility in the way that you understand to be doing it.

It's important to recognize this philosophical issue, and to recognize how a natural and intuitive way that we want to deliberate and introspect and think about who we are might be in conflict with the thought that rational decision-making defines our epistemic gold standard. (Decision theory does define our epistemic gold standard, so there's a real tension here.) There's a lot of value in introspecting. It's important for us to try to think about who we are and who we want to become when we make these big decisions. Yet there might be an in-principle conflict between this desire we have to be authentic in this sense, and the desire we have to reach the epistemic gold standard.

I want to close with another problem that comes up because there are a cluster of issues here. The other problem, which stems from the conflict between authenticity and the epistemic gold standard is the following decision theoretic issue. A natural thing to do is to say, "Let's just do some empirical research. Let's look at this question from a scientific perspective." I'm in favor of that. Doing empirical research on these questions is absolutely the best way to go, but imagine that we find ourselves in the following situation. Imagine yourself as a child-free person, as somebody who takes themselves to be essentially someone who is child-free. You're a person who has no kids, you love your life the way it is, and you think of yourself as intrinsically child-free; you have no desire to have children. When you think about what it would be like to be a parent, you think, "I wouldn't be happy, that's just not the life that I want to live."

You go around and you talk to people. Let's pretend that all the empirical research out there tells you once you become a parent, the way that you're going to evaluate the quality of life as a parent, the utility of becoming a parent, is going to skyrocket. Once you become a parent, you're going to think that being a parent is fabulous, that it's the best way to live your life, far better than it would be to live your life child-free. Let's say that all the description and testimony—from your parents, your friends who have children—all say the same thing. Now you're in a situation where you value who you are as a child-free person, your preferences are to remain child-free. You also, in principle, cannot introspect into what it would be like for you to be a parent. If you were to be rational, you should replace your assessment, your imaginative projection, with this empirical information. All the empirical information and testimony that you have tells you that once you've undergone this experience, your preferences will change so that you'll be much happier—you'll be maximizing utility as a parent.

If we're to meet the epistemic gold standard, obviously we're supposed to be utility maximizers, right? If the way to do that is to listen to the empirical research in this case, then the right thing to do is to reject your current self and replace it with the future self that's a parent. There's a problem here. The way that I've set the case up, your current self assigns a reasonably low utility to becoming a parent, but because you undergo a transformative experience in virtue of becoming a parent, your future self as a parent assigns a very high utility to being a parent. Because you want to maximize utility, you should give up your present self and replace it with your future self.

That again illustrates the philosophical tension that comes out here. Some people think that to be rational, you need to respect your current preferences. To be authentic, you have to respect who you are now. It looks like if we want to meet the epistemic gold standard in this case, we have to violate the preferences of our current self—violate who we take ourselves to be now, who we take our current self to be—and replace it with a different self. That suggests that rationality can entail a kind of self-alienation that I find worth exploring.

THE REALITY CLUB

MOLLY CROCKETT: That was so fascinating and engaging and touches on a lot of things that I think about, both intellectually and personally. One thought that comes to mind is that there's a lot of evidence from psychology that not only do we choose the things that we prefer, but we also come to prefer the things we choose. I know Laurie Santos has done some cool work on this, so maybe you want to follow up after. I've never thought about this cognitive dissonance reduction stuff from a functional perspective, but I wonder if maybe one reason that we do this is to make up for the fact that we have to sometimes make these epistemically transformative choices. Maybe one interesting question for empirical research would be whether cognitive dissonance reduction and choice-induced preference change is stronger for these decisions where there is this epistemic transformation going on.

PAUL: Absolutely. One of the things that fascinates me is this notion of preference capture, where you're contemplating the possibility of changing your preferences, and where you can't forecast how they're going to evolve. That's an important component here. You don't know what's going to happen to you, but maybe you know, "Well, it's going to be the case that I'm going to change myself so that I'll be happier with the result." Philosophers need to think about this. You're right, there might be this interesting evolutionary or adaptive feature to this so there's a way to make sense of this and think about what the epistemic gold standard should be in that context.

It also raises interesting philosophical questions about how we want to think about authenticity in the light of that issue.

HUGO MERCIER: One question is, how different are these cases? In the case in which you don't know what it's going to be like to be a vampire, to any other source of uncertainty. You have to make a decision, and you just don't know why are they different. The other question is if you look at cases like the durian example, you might use a bunch of heuristics, such as how much you like novel food, how much you like novel fruits, if it's that you like something that some people hate. You can use a bunch of things to make it less blind.

PAUL: There are some subtleties here. Normally, uncertainty regards the probabilities or the credences involved with the situation. In the way that I'm thinking about decision theory—which is a very orthodox, natural way—we've got various probabilities that we would assign to states, and then the utilities of the outcomes. Normally, standard models involve uncertainty with respect to the probabilities. The standard assumption is that we know enough to be able to assign the values, but what we don't know are the probabilities, so that's where the uncertainty is. My problem is different from that because what we don't know is the value. Sometimes we might have probabilistic uncertainty, but sometimes we might have perfect certainty about the various likelihoods. We just don't know what the values are.

That said, there is a decision theoretic move that one could make. I've been exploring different ways of developing decision theoretic models to accommodate these issues. Interestingly, all of these decision theoretic models, to accommodate the problem, to force them into a problem, say, of uncertainty, means that you get even more problems on the authenticity. It's a dilemma, and you have to figure out which one you want to choose.

There is a move you could make, where instead of the utility being undefined, you say, "I'm just going to describe every possible utility." Then the decision involves a massive amount of uncertainty about which utility is going to come into play. That's a way of pushing it all into massive uncertainty. But as you can see, then the decision becomes horrible in a different sense.

MERCIER: In more realistic cases you can use heuristics to approximate how you feel.

PAUL: Yes, I definitely agree. In some of the work that I've been exploring, there are a couple of different issues that need to be separated here. One question is are we changing the way that we're making the decision in a way that takes us away from the way we want to? We might have to. As a philosopher, I want to say it might be the case that the natural way you want to think about should you become a parent just isn't the right way to do it. We need to think about how to replace that natural model with a better model. The next question is what other models should we use? Two of my favorite options are one, where we think we know that there are these outcomes, but we don't know what the outcomes are. Then you can use principles, or some people call them heuristics. I prefer to call them principles. I think of a chess game, where for example, I haven't memorized a bunch of chess moves, let's say, and I'm playing against Grandmaster. I move my queen a certain way. I know the rules of the game, and how the pieces move, and I know when I move my queen a certain way that something's going to happen, but I don't know what all the different configurations are going to be that are going to result.

The best way to make my decision would be to endorse certain principles about how one should move in those situations, even if I can't assess the utilities and compare the outcomes explicitly. That does seem to be one way to start to address these questions. There are interesting things you can do by saying, "Let's look at how people who have come out on the other end, and who have had these transformative experiences." It might be the case that you can eat all the other fruit that you want; it's not going to help you know what a durian tastes like. It turns out that if you've done a lot of sewage work, inhaled a lot of fumes, there's a certain first-personal experience that you have that will give you some insight into the first-personal experience of tasting a durian. Discovering what those things might be, which might be very different from imagining tasting something, would also be another way to get at least a partial value. It's not going to take away all of the problems.

It's also important, not that you were doing this, but sometimes people slip into, "Well, maybe I can introspect a little bit and then use some information and do it that way." Part of my point is that natural way of thinking about things works for familiar situations, but for these radically new contexts, it's just not going to work. We just have to be careful about not slipping into that way of thinking.

LAURIE SANTOS: A lot of the work we know in judgment and decision-making because of choice induced preference changes, because we don't have access to our preferences, because our new situation changes our evaluation of the old, there's a sense in which you could take every choice and every decision as a mini-version of a transformative experience. If that's the case, then this is a fundamental problem, not just for deciding to become a vampire or having a kid. Every time I choose one of those cookies, it's going to affect my future. It's going to be a mini-transformative experience that affects my future preferences.

PAUL: Here's where I make some philosophical distinctions that are relevant. It's a context-dependent situation. I think of it in terms of experiential, natural kinds. If I'm thinking, "Do I want to have a chocolate chip cookie?" I've had chocolate chip cookies before, so in a thick sense, I know what it's like to have a chocolate chip cookie, and I can make a decision based on what I think it's going to be like. Here's what I don't have: I don't have the fine-grained experience of having that *particular*, totally fabulous, amazing chocolate chip cookie.

When you start playing around with the context and the stakes, you do get the problem right back. What I'm trying to push is that we have to be incredibly precise about how we're defining these things, or much more precise to try to avoid unknowingly finding ourselves with these in-principle structural problems.

DAVID PIZARRO: You can't know what a transformative experience will be in that precise way. It's transformative experiences all the way down. The very thing that you're saying should be a warning sign—"when it is transformative, don't make this error"—I don't know yet. Why was your category "chocolate chip cookie" and not "baked good in

Connecticut?" I don't think that you would argue that there are things that are transformative and things that are not. You could say you've had pain, and let's say that severe pain is five percent of being a Mom. You know something. You could say you don't know anything about what it's like to be in space; it's the ultimate transformative experience. Then you could say, somehow you could rank all possible experiences in roughly how much you would know. I don't think that's what you're trying to say. What you're trying to say is that some things are unknowable. I just don't think you even know what's unknowable.

PAUL: First of all, philosophers want to understand the structure as opposed to my particular situation. Second thing—we've been going very meta, and as a philosopher I have to go even more meta—it might be that you discover the category of transformative experiences, but you have to have one to know what it's like to have a transformative experience. There's a little wiggle room there.

SANTOS: You think you know, but you know.

PAUL: You can always raise these questions.

PIZARRO: Some people say, "I've had a dog so I'm totally with you about the mom thing."

CROCKETT: You talked about natural kinds. I wonder if there can be a distinction between experiences that are truly unknowable and other experiences that you've never experienced before, but you can simulate. There's an interesting paper by Helen Barron and Tim Behrens published last year, where they looked at the neural mechanisms of making decisions about novel goods. The goods were food items that were composed of familiar foods—a raspberry avocado milkshake, for example, or tea-flavored JELL-O. Even though you've never had a raspberry avocado milkshake, you can simulate raspberry and avocado and what that would be like. What the brain is doing in these decisions is, you can see the trace of avocado and the trace of raspberry being combined and simulated in that way. I'm just wondering if you make that distinction between non-experienced but potentially simulated.

PAUL: I read that article. I thought it was a cool article. It's important to see that these decision problems are arising at the individual level. It obviously doesn't mean that individuals can't make generalizations or draw on past experience, but the way that each of us faces this problem is going to be highly dependent on the previous experiences that each of us have had. That is crucial.

In some experiences, past experiences of parts can be conjoined to allow us to imagine what the whole experience would be like, but other experiences don't seem to be like that. Again, having a child, speaking from experience, didn't seem to be that way. There are interesting empirical and philosophical questions here about which experiences can be collected or conjoined together so that you can perform an imaginative simulation or model, and which ones aren't, and why not.

FIERY CUSHMAN: I feel like part of what the core question is going to have to be is does the unitary utility comprise the full space of decision-relevant qualia? If it's the case that even parenthood finally grounds out in utility, because I know what utility is, you can tell me how much you have, and then the only relevant simulation I need to perform is one of utility. Not one of the particular experiences that happen to afford utility, if utility is the only kind of qualia that has decision value. But if somehow it were possible that there were qualia which were not utility, which could not be translated into utility, but that would still bear on decisions, then it would be necessary for me to simulate those, and they might be unknowable because I've never had them before, unlike utility, which I've certainly had before.

It feels like there are two interesting ways to go. One is to say utility just is decision-relevant things. If you've ever experienced making a decision at all, then you know what it is to make a good one, and parenthood is a good one. It's going to be one of those ones that you like.

PAUL: I was with you until that last bit. It's absolutely the case that we need to think in a more sophisticated way about utility. One thing that I'm convinced of is that we should not be thinking in terms of simple hedonic pleasure and pain. One thing I'm fascinated by is the intrinsic value of experience. There's work in philosophy about color experience; a lot of that work is focused on defining color terms. Some of the interesting work concerns this notion of revelation. In other words, what do certain experiences reveal? Whatever it is that they reveal, it's very hard to pin down. We also think that they're valuable. This comes from Aristotle, who argued that, in principle, experience has a value to us. It's not clear to me that we can measure that in terms of hedons. I'm not saying the values are not comparable, it's not a straightforward issue about incommensurability either. Rather, this is another place where there are pressing questions that philosophers, in particular, need to think more about. There's been less attention paid to this formal approach towards phenomenology than there needs to be. Again, it seems to me there are obvious empirical ways that we could think about exploring this.

MICHAEL MCCULLOUGH: Someone brought up the usefulness of relying on past experience, and I wanted to get in on that and combine that with Hugo's comments about heuristics. There is another way you can draw on past experience, which is to draw on deep past experience. One of the things that natural selection does is it capitalizes on invisible correlations ancestrally. You could say it's more or less self-evident that ancestrally there was a correlation between having offspring and fitness, on average. I'm just going to put that out there. Dawkins has this idea of child lust. It's as real a quality as sexual lust. That is the product of this invincible correlation in deep time between setting yourself up for having children and getting more copies of your genes out into the world.

JENNIFER JACQUET: But birth rates are going down.

PAUL: I say we grant him this, but there's some work by Gary Brase and some other people about baby fever that might call that into question, but let's pass it on.

MCCULLOUGH: We can use that for anything—food, maybe that's a better one. To what extent can we ask for a free pass on trusting some of our intuitions for some of these big questions? The intuitions, on average, are going to be reliable ones. You can also back this out to after you've had the child and you begin to regret it. What does that say? That's going to happen to some people. The mind is built around these invisible correlations that build up over time. You can imagine, as horrible as it is to contemplate, that compulsion that some people have to get rid of their child, is reflective of some information in the environment. Not 100 percent reliable information coming through, a noisy signal processor that says perhaps this is not the time when taking this child forward is in your best inclusive fitness interest.

PAUL: Let's go back to the philosophical picture. Again, there is an ancient philosophical picture, where the best self is the rational, deliberative self, who thinks carefully, assesses their intrinsic inner nature, and then chooses in a calm, epistemically wonderful way. We've got a picture where that involves a certain reflection on who you are and involves certain imaginative or mental capacities. What I'm hearing you suggest is maybe that's just not the right story for a lot of our big decisions. Maybe there's a different biological story that we should look to. That's worth exploring. But again, it illustrates this tension between this picture that we have of ourselves as introspective agents and what rationality might demand. When you're faced with a big decision, let's say, in cases of informed consent, or you're thinking about writing an advanced directive, you are supposed to think very carefully about what you want. There's something unsatisfying about being told, "I'm going to replace that picture with something else."

MCCULLOUGH: Maybe what I would want to say there is when you ended your talk by saying we find ourselves alienated from ourselves, maybe the thing I'd want to say there is this problem alienates us from our system 2 self.

DAVID RAND: No, I don't think so. From a rational perspective, the problem is how do I predict what my post-child-having utility is going to be? If you believe in natural selection, you can say from a completely rational perspective, "I understand that it must be the case that I will be glad that I had the child afterwards. Otherwise, we would all be dead."

JACQUET: Well, no, we never had the choice before. The whole idea about us having preference on this decision is a new one.

PAUL: This particular case is very modern. The more choices that we get, the more control we have over our futures, the more we face these issues. It's a distinctively modern problem in a certain way, as well as having ancient connections. From an evolutionary perspective, we know that our preferences are going to change, and we know we're going to be happy—supposedly; let's disregard the confusing empirical results—so we should do it. All you're saying then is that we should replace our current self with our future self because we should be utility maximizers. That needs to be questioned. That is one route, but it's not an obviously satisfactory route, especially when you don't want to have kids. It might be like, "You've got to have a kid because you're not rational if you don't choose to have a child. You're not even biologically fit in some way." That's a deeply problematic claim to make.

PIZARRO: But you're also saying you're not rational if you don't choose to have a child.

PAUL: Well, that's right. I was just responding to Dave's suggestion there.

PIZARRO: Here's one example of perhaps the most epistemically and phenomenologically unavailable state: Death. Surely, one implication of what you're saying is that it is not rational to not want to die. Because after all, not only do I not know what it is, I can't even ask anybody.

SIMONE SCHNALL: But you don't have much of a choice in the matter.

PIZARRO: Sure you have a choice—suicide.

PAUL: Death involves the absence of experience, in a certain way.

Here's a problem. In principle, we can't know until we do it. That's why there are these decision-theoretic problems, because look, do you want status quo bias? Is discovery always good? There's no simple, straightforward answer.

PIZARRO: I'm was trying to use this as a *reductio* where you would say, "Surely it is irrational to prefer death, all things being equal!"

PAUL: I wasn't suggesting that we should prefer death, all things being equal.

PIZARRO: No, but the rationality of it. That you would say that I am able to make a rational decision based on something that I know nothing phenomenologically or epistemologically.

PAUL: You have to be very careful about how you're framing that decision. My claim isn't that there's no way to make rational decisions about these different things. You can, for example, rationally choose to have children. You can rationally choose to try durian and other things. But there are certain bases that you can't use to make your decision. This way of thinking is a mistake: I want to commit suicide because I know it's going to be better when I'm dead. That doesn't make any sense, for obvious reasons. It's also the problematic form of the decision that you see in a number of other cases: I want to be a parent because I know it's going to be better, because I know it's going to form my preferences. That reasoning is problematic. We have to be quite careful here because although I do want to say that this is a fundamental problem and it's fairly far-reaching, it doesn't destroy decision theory, and it doesn't destroy the way that we can rationally plan. I'm a big fan of using principles, and I'm a big fan of building more sophisticated decision models, where we distinguish between different types of utility. My point is that we need to see the structure here and see the complexity so that we can successfully attack the problem.

Michael McCullough: "Two Cheers For Falsification"

What I want to do today is raise one cheer for falsification, maybe two cheers for falsification. Maybe it's not philosophical falsificationism I'm calling for, but maybe something more like methodological falsificationism. It has an important role to play in theory development that maybe we have turned our backs on in some areas of this racket we're in, particularly the part of it that I do—Ev Psych—more than we should have.

MICHAEL MCCULLOUGH is Director, Evolution and Human Behavior Laboratory, Professor of Psychology, Cooper Fellow, University of Miami; Author, *Beyond Revenge.*

TWO CHEERS FOR FALSIFICATION

I'm Mike McCullough. I'm a psychologist at the University of Miami. I want to talk a little bit about some thoughts I've been entertaining about falsification, and particularly its place in my bailiwick, which is Ev Psych.

Most of you, when you think about falsification, think about Karl Popper, who had this idea that is pretty compelling, which is that we can never have positive evidence for a hypothesis. Hypotheses give us predictions about how the world should be ordered. What we like to do is take the data from the world and then make inferences about whether the hypothesis is true. This is a problematic kind of reasoning. It's not valid reasoning. As I'm sure many of you know, he suggested that there is a valid reasoning we can use for making inferences about the truth value of hypotheses from observations, but the way to do that is to look for predictions that are falsified in the world.

Your hypothesis says the world should be structured this way, and when you find evidence that it's not structured that way you're then in a position to make a valid conclusion about how the world is structured. More specifically, you're in a position to say this is a valid conclusion about how the world is *not* structured. That's valid reasoning. Modus Tollens works.

The thing is, science has done pretty good with basic induction. Most scientists feel this way anyway. Taking observations about the world that are true and then making inferences about hypotheses has been a pretty decent way to do science. For most practicing scientists, affirming the consequent looks like a reasonable way to approach our jobs.

What I want to do today is raise one cheer for falsification, maybe two cheers for falsification. Maybe it's not philosophical falsificationism I'm calling for, but maybe something more like methodological falsificationism. It has an important role to play in theory development that maybe we have turned our backs on in some areas of this racket we're in, particularly the part of it that I do—Ev Psych—more than we should have.

In general, we don't do much falsifying. As an empirical matter this is old news. Everyone knows that most of what gets published in journals is supportive of hypotheses: "The prediction supported the hypotheses." There are lots of methodological, sociological, and cultural reasons for that. I'm not that interested in that other than to say we are comfortable with doing confirmatory research, and there may be some non-methodological reasons why we like that. One is that falsification is just hard. We don't do it very well. Think about something like the basic Wason selection task. We know from that work by Wason that when people have cards with even or odd numbers on them and the other sides of the cards have either the color red or the color brown, and they're asked to determine whether a rule is being followed—if the number is even, it must have a red color on the other side—people don't choose well. They don't choose optimally. They're good at saying, "We should turn over the even cards and see if there's red to see if that condition is being fulfilled," but then we want to turn over red cards, and see if there are even numbers. We don't do Modus Tollens very well. It just doesn't come naturally in these basic problems that are more like scientific problems.

There are two other reasons why we are not as good as we should be at falsification. One is that we tend to have weak methodological beliefs. This has been beaten to death over the past three years. We tend to imagine lots of ways in which the methods might not have been good enough to provide a decent falsification of the hypothesis. I'm more interested in the obverse side of that, particularly as it concerns EP, which is we tend to have very strong theoretical beliefs. This hits home in Ev Psych, where I live, because evolutionary psychology is working on the theory of natural selection, which is one of the strongest theories in all of science because it has just a couple of basic ideas that you get very powerful axioms from.

If you assume you've got these replicating units in the world—let's just create a planet and put these things on it that have found a way to take material from around them and hack off copies of themselves, that’s all you need to get evolution by natural selection. If you want to slip in there that the replication has to be imperfect, you can put that in as a second condition, but that’s life. Everything is imperfect.

Once you’ve got this world with replicators in it, we know exactly what’s going to happen. They’re going to go through time cloaking themselves in these nifty design features that increase their reproductive rates. It’s axiomatic; we don’t need to prove this in any sense. It’s an algorithmic process that’s going to happen.

As a function of this building of design features, critters are going to come to look like they were built to optimize something. That’s one thing we can take as canon. We don’t know quite what they’re optimizing until we get to Bill Hamilton, and then finally we see what it is these design features are, in a sense, for—to maximize inclusive fitness.

This is a strong theory. It allows you to predict that there are features in the world that have purpose to them. The purpose is increasing inclusive fitness. Here's the problem: Once you get to that point there are no more axioms to be had. You don’t know what design features to expect natural selection to provide you with. Thomas Nagel had us wondering why don’t pigs have wings. It seems like from a certain way you could think about what’s good for pigs, "Gosh—they should have had wings by now! Wouldn’t wings be great? They could fly, they could skip over empty troughs, and they could go find the muddiest mud holes. It would be fantastic."

Why don’t they have them? The reason they don’t have them is many-fold. There were powerful constraints against it; every gene that would have helped to build wingedness was downward from an optimal peak. All of those reasons they don’t have wings are highly contingent on natural history, highly contingent on phylogenetic constraints. As soon as we start trying to predict the design features that are going to be out there, we’re stuck. Natural selection only takes us so far as a theory. It cannot tell you, for any organism, what you're going to find. You know what the things you’re going to find are for with reasonable certainty, but you can't postdictively predict them.

Ev Psych is a big idea, and big ideas get criticized for good reason—they’re big ideas. But the criticism of some of it has crystallized around a few substantive areas of research. It's been interesting to think about what’s going on in these areas that make them such lightning rods for criticism. Some of it has to do with falsification or maybe even a lack of appetite for falsification.

We’re going to have to discover adaptations ultimately, not deduce them from first principles. One of the areas where this has become a

PR problem for Ev Psych is in an area of research on how women’s behavior varies as a function of their ovulatory state. There’s a lot of work coming out these days based on the idea that when women are not fertile they have a set of preferences, a set of emotions, a set of motivations, and a set of behaviors that are quite different from when they’re ovulating. They prefer to wear red when they’re ovulating. They find certain features in men more sexy, attractive, and worthy of pursuit when they’re ovulating than when they’re not ovulating.

One of these hypotheses behind this idea—and it’s in the news as we speak, the Internet is burning up with commentary on this—is a hypothesis called the ovulatory shift hypothesis. The hypothesis is that women who are not fertile have a different reproductive agenda than women who are currently in the fertile part of their ovulatory cycle. The idea goes something like this: When you’re not fertile, what interests you in men—among other things, but particularly these things—are traits that might make them valuable, long-term mates. When you’re ovulating, what you shift your focus toward are indicators that that man has got good genes that are free of a heady mutation load, and that the man is in good condition.

The idea is that you want to get those good genes, so you’re looking for features in men that, ancestrally, would have been correlated with the possession of a low mutation load, got enough to eat while developing, were free of a lot of blunt force trauma as an adolescent, any features that would indicate that this is somebody who’s carrying around a decent set of modules for me to merge with my modules.

This could be right, this hypothesis. There is something I like about it. I find it a provocative idea. What I realized about a week ago is it’s not inevitably right. It’s not axiomatically right. There’s lots of ways we can think about natural selection having worked on women’s psychology that would have been inclusive fitness maximizing, but that would not have involved an ovulatory shift in preferences, behaviors, emotions, and motivations.

One thing you can imagine is that the same genes that make you not so attractive as a man—when there are mutations that come out in reducing your body symmetry, your attractiveness, or masculinity; What if it’s the case that those same defects, those same genetic problems also make you a worse dad? Why aren’t the dad modules equally run down by these genetic defects that are piling up in a visible way in reducing your condition, or your masculinity, or your symmetry? That's one possibility that would say maybe there is no ovulatory shift. Maybe women want and find attractive in men the same things throughout their entire ovulatory cycle.

I don’t think these are outlandish possibilities. They are hypotheses that you should put on the table. Another one is maybe men were sufficient, but not necessary co-provisioners for children, ancestrally. There’s some decent evidence that if you’ve got another co-provisioner around, men might be facultative rather than obligatory co-parents. If that’s the case maybe what women want all the time is just the good genes. That’s what they’re interested in 24/7 because male parenting effort is not something that’s making a marginal, unique contribution to women’s inclusive fitness.

All to say, there are data to suggest that those hypotheses are not laughably, absurdly wrong. They're reasonable alternatives. Here's what is happening in some of these areas that have become so fractious, like this research on ovulatory cycle shifts. For those of you who don’t know what’s going on, there are a set of reviews and meta-analyses by some researchers that have looked at all of the evidence for the ovulatory shift hypothesis. They’ve meta-analyzed it very carefully, and on some measures of this shift in women’s preferences for certain traits in men in certain perceptual contexts, the data looked pretty supportive. They're not overwhelmingly supportive, but they’re supportive enough that you’ve got to give the hypothesis some serious thought.

There is another group of researchers who think that the literature is so beset by methodological problems that we ought to ignore it and start over with other studies. Then there’s a sub camp of that camp who says the hypothesis itself is fundamentally flawed. What has happened, in part, to make this so inflammatory is that some of the EP researchers who are working on this hypothesis—in a lot of different areas, not just in mate choice, but also in advertising one’s own qualities to prospective mates—maybe have come to feel that the hypothesis they’re supporting is *the Darwin hypothesis*. It’s the hypothesis that’s available that has to be defended if Darwin is right. If there’s not something profoundly wrong with the theory of natural selection, this inevitably has to be right. It has to be right in some way.

For people in that other camp who may have motivations to find problems with this literature, maybe they feel that the hypothesis is so offensive to what we understand about cultural variability and the problems that women have faced traditionally that we need to undo, that it’s become for them a cause that is a righteous one because they view it as the propagation of the mistreatment of women using the tools of science.

What I want to suggest is this thing can be unwound. I’m confident that the empirical matters of fact are going to get settled one day. What I want to suggest is that methodological falsificationism could have a huge role to play here for people who care about the theory of natural selection and applying it to understand human behavior. I’m going to go one step forward axiomatically and say there are some forced moves. Dennett talks about forced moves—things that are inevitable consequences of natural selection. You’ve got to have a skin. There’s got to be a place where you stop and the other organisms start. That’s a forced move. It’s going to happen to all replicating things. You have to know where your interests end and the next guy's begins. That’s a forced move.

Another forced move is you’ve got to get something to eat. Every organism has got to reverse entropy locally. For sexually reproducing species I think it is the case, there is a forced move, which is to get the best modules you can from your mate. We should expect the female ovulatory cycle to be well tuned to operating the way it should have ancestrally to maximize women’s inclusive fitness.

I’m going to go a little bit axiomatic on you in that respect. That’s the way we should expect that system to be designed, but I don’t know that we know enough about the initial conditions. I’m not so worried about available genetic variance, but I do worry about phylogenetic constraints, social ecology. There are a lot of possibilities on the table—I just made up those fun alternative hypotheses—that could be, if stated more seriously, reasonable alternatives.

What I would like to see is the next wave of research on that involve some evolutionarily interested, engaged researchers who take other hypotheses like those seriously as alternatives, so we cannot evaluate the fit of a certain model in absolute terms, but we can say which of the viable models fit these data better and which ox gets gored by the data. We should be in a position at the end of the day to take some hypotheses off the table through a greater psychological comfort with disconfirmation, with seeking out falsifying evidence. To do that comfortably for somebody who cares about using the tools of natural selection to understand human behavior, you’ve got to have multiple hypotheses on the table.

There are a couple of areas where this is happening in a healthy way. I’m not a mating guy. I mean I’m a mating guy, but I don’t study mating 24/7. Probably, I do, but not for pay. That's an area of what the putative functions of life after menopause is about. There’s not just a single hypothesis that’s being taken seriously by evolutionary demographers, psychologists, and modelers about why it is women live past their reproductive years. We’ve got at least two plausible hypotheses. One, which is the very famous one, is the grandmother effect. Women improve their inclusive fitness by withdrawing their reproductive effort from their own direct fitness maximization and then toward providing care to their daughters’ offspring. It’s a good hypothesis. It’s a winsome hypothesis. It’s not the only hypothesis.

There’s also a hypothesis that menopause came along to minimize parent-offspring conflict, which is a distinction with a difference. They lead in some ways to some interesting differences and predictions. If you ask me about them after the camera’s off I’ll tell you I’m not able to describe them to you in a whole lot of detail. That’s one area.

Whatever hypothesis gets gored at the end of the day, there’s going to be an evolutionary hypothesis still standing, and it might be a functional one. Gould told us to be pluralistic about all of the possible mechanisms that could generate design features—phenotypes. What I want to say is we need to be adaptive pluralists as well. We’ll do a lot better as a science with adaptive pluralism in the same way that we are multi-evolutionary force pluralists.

The other area where there’s some nice work going on that’s falsificationist is in the area of cooperation. Fifteen years ago there was an idea that Ernst Fehr championed and Herb Gintis and a few others, which suggested that humans are as nice as they are, as cooperative, as disposed to punish bad guys as they are because, through some group selection—either cultural group selection, or gene-culture evolution, or genetic group selection—we developed this propensity they called strong reciprocity. Because it relied on certain flavors of group selection that some evolutionary psychologists found objectionable, there was a lot of grumbling in the field for a lot of years. It took a while, but 7 to 10 years later people figured out how to put together some empirical horse races for some of the claims that would leave one ox gored. Somebody was going to get bloodied, some hypothesis was going to be damaged, and another one hopefully would be less damaged.

We are far from solving that problem, far from resolving those debates because it’s still true that theories die funeral by funeral in science. We at least know how to frame debates as empirical matters that ultimately could enable us to adjudicate hypotheses that give proper props to Darwin, give proper props to the only force we know in the universe that can build complex functional design, and also help us build a better science.

THE REALITY CLUB

SARAH-JAYNE BLAKEMORE: Do these kinds of hypotheses—Ev Psych hypotheses and particularly the ovulation shift hypothesis—only apply to humans? Is there research on other species that form pair bonds, like prairie voles? Do the women prairie voles prefer ripped men when they're ovulating?

MCCULLOUGH: I don't know. There is a lot of sexual choice.

BLAKEMORE: Isn't that a key question? Whether it applies across species?

MCCULLOUGH: It's an interesting question, but it's not decisive. It wouldn't be decisive because they're not humans. They don't have our same history.

BLAKEMORE: In a way Ev Psych does just apply to humans. Well, humans are what you're interested in—human behavior and human psychology.

MCCULLOUGH: Yes. It is an attempt to unite evolutionary biology and behavioral ecology more closely with psychology.

LAURIE SANTOS: The problem that you described, about sticky hypotheses, is that folks don't want to propose other ones (like this ovulatory shift hypothesis) in the face of a lot of folks who said that the data are not there. It stuck around, and folks haven't proposed alternatives. But I don't think the same has been true in cases of behavioral ecology, where you'd imagine the same constraints apply. Behavioral ecologists believe in inclusive fitness theory, they believe in sexual selection, yet when some life history variable does work for one individual species, ecologists propose a ton of other alternative hypotheses. So why in EP did we get stuck? That's an interesting sociological question because you don't see it with folks who have exactly the same commitments about the meta-theories.

MCCULLOUGH: You took the words right out of my mouth. One of the people who works with me as a primatologist, and when we were talking about this issue at a lab meeting she said, "This is just astonishing to me." She's working with humans in the lab for the first time. She said exactly what you said. We see something unusual, we come up with three, four, or five hypotheses, potentially, and you see which of those hypotheses gets a better fit from the data, and you throw the worst ones away. It is interesting to me that these are sticky. I would love to know why that is. I wonder if there's an interesting set of attractors that have to do with how we intuit about humans.

SANTOS: Another question I always have for EP is basically Sarah-Jayne's question. There are other species that have parts of our life history, which, if we feel strongly that some set of hypotheses should be true of humans, we should expect to those traits in them. But it depends on the specific hypothesis about some aspects of a species' life history. If we have a hypothesis about mating and pair bonding species, then we should expect to find it in other pair bonding primates. But if our hypothesis is about being a species where men have access to resources and females don't, we should look to a species like pair-bonding primates, and expect to see that trait there.

As an animal person it's always felt like there's a reluctance in EP to even speculate about what's going on in animals for the same answer you just gave, which is: "We're just interested in humans," but if you're interested in the mechanisms that shape humans, then those same life history variables should be of interest broadly. It feels like they're not.

MCCULLOUGH: One example that is of interest to me in this particular debate is some researchers who have taken up the hypothesis that women advertise their ovulation through dress, which I find interesting. I don't see a lot of evidence that they have evolved other, more direct ways to advertise their ovulation. In fact, I don't know if they've evolved to conceal it, but they certainly don't seem to have evolved much more straightforward ways to advertise it. Relative to things that they could be doing, if you take into account all of the interesting things primates have managed to do to advertise their ovulation.

I don't mean to say don't tell me about those other animals. Personally, I love that work.

SANTOS: It seems like that could be a useful way to do falsification in EP. In part because sometimes there are experiments that are hard to do with humans, but you can do in other animals. You can do stuff with the pair bonded prairie voles that IRB-wise is a little bit sketchy to get approved with undergrads.

DAVID PIZARRO: Facebook would do it.

SANTOS: Not Facebook, but OkCupid. It seems that other animals provide a useful possibility for falsification. One of the things I teach in my EP course is this idea that children should look more like their fathers at first. There is another evolutionary psych hypothesis that's out claiming that males won't murder children who look like them because they think they are the fathers of the kids. This claim makes a strong prediction in primates that when the females are mating with a bunch of dudes and it doesn't matter who your kid looks like, you should see that those primate kids look less like their dads than human kids look like their dads. We should be able to do a comparative study. It seems like sometimes other species could form this sort of beneficial outgroup, and I've always been curious why EP folks seem frustrated to use that as a potential falsification.

MCCULLOUGH: Maybe it's just a matter of getting the groups together who have the resources to make that work happen. That's a reasonable first approximation. Psychologists go to school and learn to do studies in the lab with Homo sapiens, and to the extent that they know researchers who have access to non-human animals it might be fantastic, but perhaps the network isn't there.

PIZARRO: Just listening to who's been doing this work, it strikes me that as you set out the task of evolutionary psychology coming up with maybe some empirical horse races to pit hypotheses against each other, there are two ways in which this seems to be done. One is by people who are committed to the view that, say, in this case sexual selection would have placed specific pressures on female mate choice or ovulation. How did that manifest itself? As you were laying out two potential hypotheses, but both of those share the commitment that the way the natural selection worked was by changing women's judgment about sexual selection during ovulation.

MCCULLOUGH: No. Not necessarily during ovulation.

PIZARRO: Yeah. But as a change in whether or not they are fertile.

MCCULLOUGH: Well, no. It may be that they just like a certain set of traits all the time, 24/7. My alternatives were to explicitly not shift hypotheses.

PIZARRO: It's still framed as accepting that there would be direct natural selection pressure on female mate choice in a way that ought to be obvious with the right test, so you can have three hypotheses. It strikes me that the people who disagree with this, and the majority of people who find this controversial or object to it, it's not because they think you have the wrong view of how sexual selection drove female mate choice and it's also not that they don't believe that evolution shaped the human mind, it's simply that they disagree about whether or not natural selection or sexual selection, in this case, happened to shape something like sexual selection in the way that evolutionary psychologists say.

You have people, for instance, in cognition in general, who believe that the way natural selection works was by giving us a lot of general-purpose learning mechanisms. They’re not disagreeing with the view that evolution shapes the mind, but they are disagreeing with a very specific view about how evolution shapes the mind—very specific, functional, modular—and it’s rare to see somebody in this camp who has the commitment to saying that natural selection shaped the mind so specifically across human cognition, or who accept as evidence the studies that the people in the other camp do. It's almost what you take to be axiomatic is different than what they take to be axiomatic about natural selection, therefore, it generates very different predictions. Then there’s this tension because they think you guys aren’t paying any attention to specific hypotheses about how culture works, or how learning might cause diversity, so they accuse evolutionary psychologists of that. Then evolutionary psychologists accuse them of just being misguided about how sexual selection had to have shaped the mind with the specific modular decision-making mechanisms.

How do you get those two to agree on the terms by which you can say it turns out sexual selection has nothing to do with the way that women choose their mate or choose their clothing? (Maybe not just with their mate, but a large variety of other decisions that women might make.)

MCCULLOUGH: The claim that I want to make is that mate choice is pretty close to the engine of natural selection. Organisms are going to be under selection pressure to identify reliable cues to mate quality. That’s about as far as I want to go without some more information, please. I just don’t see how natural selection can build general-purpose things. It can build a lot of special-purpose things that combine together to make you look awfully general purpose. But that, to me, is something that I have found useful for helping me.

PIZARRO: Right. And that’s the heart of the controversy.

MCCULLOUGH: It is absolutely the heart.

DAVID RAND: It seems to me that part of the point that I’m taking from what David was saying is that if there’s going to be a concept of horseracing, all the different sides have to agree on who the horses are.

MCCULLOUGH: Yeah. Absolutely. That’s right.

SANTOS: And if it's domain specificity versus domain generality that’s going to look like a different horserace than: When in the cycle do they move?

JOSHUA KNOBE: This point people are making with the horses seems helpful, in terms of thinking generally about what falsification is. There’s this picture that you started out with, from Popper, where you take one individual hypothesis, and there’s just some fact about whether they did an experiment that falsifies it. Even if you had no other hypothesis, you just have to abandon it.

Then there’s this other metaphor that looks much less like a trial and much more like a race. The idea is that whether one horse wins depends on its status relative to the other horses; it’s not just some individual fact about that one horse. In the latter way of thinking about falsification, the only way something can be falsified is dependent on how it did relative to these other hypotheses. It’s not that you can know from a given experiment whether it falsified this theory about the ovulatory cycle. It depends on whether you’re comparing it to other hypotheses that also involve adaptation, or to other hypotheses about the general mechanisms.

MCCULLOUGH: There’s another problem here, which is methodological, which is what if this stuff is so deeply buried by culture, by cultural variability, by time lags, modern environment being so different, that it’s so deeply buried that we can’t find it with the current methods, in any case. I share what tends to be a healthy skepticism about many layers of the feasibility of this enterprise. It's all quite valid.

RAND: Another issue with the horse racing is that not only does everyone have to agree on which horses are in the race, but they have to agree on the details of the horses. In a lot of situations, people make evolutionary hypotheses using words that mean one thing to the person that said the hypothesis, but different things to other people. Then there’s a lot of disagreement. That’s where formal models, like from evolutionary game theory, can be useful. You say, "Look, this is exactly what I mean, and we can say exactly from this set of assumptions that this is what the result should be."

What that reveals is, from however many decades of evolutionary game theory models, one of the basic lessons is that what happens depends on the details of the assumptions that you make. You can falsify one specific model, but then people can think that framework just wasn’t the *right* group selection model, or it just wasn’t the *right* individual selection model. You can’t be like, "Look, this is an experiment that distinguishes between these two broad *classes* of theories." I guess you could if people that wanted one side to be right could try every single model they could possibly think of and never come up with a model that worked. Based on my understanding of models, that’s probably not going to happen. For any given observation, you’re always going to be able to come up with some individual selection and some group selection model that can make that prediction. So, it's hard.

FIERY CUSHMAN: For at least the first specific case that you brought up—the ovulatory cycle effects—we’re being our own conceptual sophistication in thinking about Popper and about a model comparison approach. It seems like what the meta-analyses disagree on is the question: If you assess women’s mate preferences over the course of the ovulatory cycle, do they change? The meta-analyses are not even addressing the issues of whether given the existence of such a change natural selection provides the right answer, or which form of natural selection, whether it’s to remain general or remain specific, that’s not what’s at issue. What’s at issue is: Is this a whole pile of false positives or not? It seems like we don’t have to engage in thinking about Popper or model comparison. What we need to do are some replications.

Everyone else has framed it as the question where there’s much bigger issues at stake. Am I missing something?

JOSHUA KNOBE: It’s just because people are thinking about examples other than ovulatory cycles.

RAND: Say you take something where there is evidence, then the question is what can you conclude from that evidence.

CUSHMAN: Is that a problem that you think characterizes evolutionary psychology? That is, there’re lots of instances where the evidence is clear, there’s no dispute over the effect, unlike the ovulatory cycle issues. The fundamental problem that we face is an adductive one—what is the best explanation for that evidence?

MCCULLOUGH: The latter. I don’t wish to characterize all of evolutionary psychology as having this particular issue. Some areas of Ev Psych do, some areas of lots of psych do. In fact, it’s possible that most areas of psych do suffer from this problem. There’s a nice opportunity here for Ev Psych to take a cue from behavioral ecology in that way and remember that it’s good to train students to pick up another hypothesis and play with it, and hold them all equally dear and equally hostile. Then you can do a horse race that may end up with a more interesting finish.

RAND: One thing that occurs to me is that something that happens a lot in Ev Psych, and also is criticized about economics a lot, is that in both cases there is an overarching theory, a very clear overarching theory, and then there’s an attempt to see how much of the data fits with the theory, or in what ways can you contort the theory to try and make it fit with the data. My sense is that in a lot of the rest of psychology, there just is no overarching theory. It's like we did some experiments, we got an explanation for these experiments. These other guys did an experiment, they have some explanation for their experiments. An overarching theory causes all kinds of issues, and maybe people bend it a little extremely sometimes, but that is maybe one of the things that comes along with trying to have an overarching theory of things.

MCCULLOUGH: That's why I made the point I made about the very strong theoretical beliefs that evolutionary psychologists have. There're two or three beliefs that are very "Father, Son, Holy Spirit," they're very strong.

HUGO MERCIER: To some extent when we are talking about having experiments that can distinguish between the very broad classes of models, such as group selection versus individual selection**,** or even when we're referring to Popper's falsificationism, which isn't meant for this kind of thing, we have to give up this physics envy that we have. We have to put much more weight on the preponderance of the evidence. You have this whole framework, and you can look at things, not only experiments, but real life data. This is completely undervalued in psychology. You have to have that one experiment that shows what you want to demonstrate, and everything else just gets thrown out of the door. This is done partly, for good reasons, because we want to be more scientific, we put a lot of weight on these things, but we might be slightly off sometimes.

MCCULLOUGH: The people who work in this area do take what they can get from what we know about the life history and demography of ancestral humans and they try to interact with it. To the extent that I portray them as being insensitive to other kinds of data in their theory development, I probably did them wrong. They do take that other work seriously and try to interact with it, but there are still holes in the record. We don't have a complete, perfect evolutionary reconstruction of the human ancestral social ecology any more than we do for any other critter. There are going to be multiple ways of getting from there to here.

MERCIER: Then it is not true that science progresses funeral by funeral. In sciences that work, you don't need people to die. Revolutions happen all the time. People change their minds all the time to adopt the best theories. We have less good evidence, so it makes sense that we change our minds more slowly. In physics, clearly people change their mind all the time very quickly, in biology as well, so it's our problem rather than the problem of science.