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 '13:
WHAT'S NEW IN SOCIAL SCIENCE?

Sendhil Mullainathan, June Gruber, Fiery Cushman, Rob Kurzban, Nicholas Christakis, Joshua Greene, Laurie Santos, Joshua Knobe, David Pizarro, Daniel C. Dennett. Also participating: Daniel Kahneman, Anne Treisman, Jennifer Jacquet.

***Edge* URL:** http://edge.org/event/headcon-13-whats-new-in-social-science

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.

ABOUT EDGE.ORG

"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*

"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*

"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*

"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*

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

INTRODUCTION

In July, 2013, *Edg*e invited a group of social scientists to participate in an *Edge* event focusing on the state of the art of what the social sciences have to tell us about human nature. The ten speakers were Sendhil Mullainathan, June Gruber, Fiery Cushman, Rob Kurzban, Nicholas Christakis, Joshua Greene, Laurie Santos, Joshua Knobe, David Pizarro, Daniel C. Dennett. Also participating were Daniel Kahneman, Anne Treisman, Jennifer Jacquet.

We asked the speakers to address the following questions:

> "What's new in your field of social science in the last year or two, and why should we care?"
>
> "Why do we want or need to know about it?"
>
> "How does it change our view of human nature?"

We also asked them to focus broadly and address the major developments in their field which included, but was not limited to, their own research agenda. The goal: fresh, up-to-date, and original field reports on different areas of social science.

HeadCon '13: WHAT'S NEW IN SOCIAL SCIENCE was also an experiment in online video designed to capture the dynamic of an *Edge* seminar, focusing on the interaction of ideas, and of people. The documentary film-maker **Jason Wishnow**, the pioneer of "TED Talks" during his tenure as director of film and video at TED (2006-2012), helped us develop this new iteration of *Edge* Video, filming the ten sessions in split-screen with five cameras, presenting each speaker and the surrounding participants from multiple simultaneous camera perspectives.

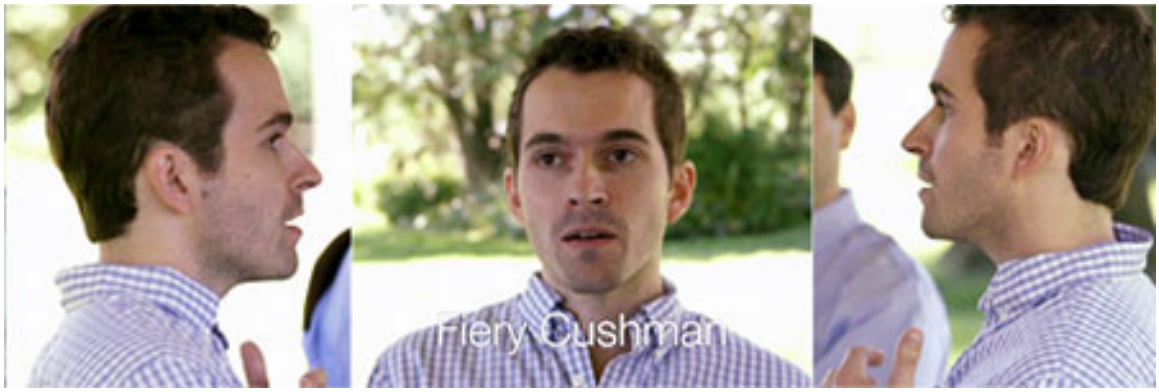
We are now pleased to present the program in its entirety**,** nearly six hours of *Edge* Video and a downloadable PDF of the 58,000-word manuscript.

The great biologist Ernst Mayr (the "Darwin of the 20th Century") once said to me: "*Edge* is a conversation." And like any conversation, it evolves. And what a conversation it is!

John Brockman, Editor
Russell Weinberger, Associate Publisher

PARTICIPANTS

Nicholas Christakis is a Physician and Social Scientist; Director, The Human Nature Lab, Yale University; Coauthor, *Connected: The Surprising Power Of Our Social Networks And How They Shape Our Lives.*



Fiery Cushman is Assistant Professor, Cognitive, Linguistic, Social Science, Brown University.

Daniel C. Dennett Is a Philosopher; Austin B. Fletcher Professor of Philosophy, Co-Director, Center for Cognitive Studies, Tufts University; Author, *Intuition Pumps.*

Joshua Greene is John and Ruth Hazel Associate Professor of the Social Sciences and the director of the Moral Cognition Laboratory in the Department of Psychology, Harvard University. Author, *Moral Tribes: Emotion, Reason, And The Gap Between Us And Them.*

June Gruber is Assistant Professor of Psychology, Director, Positive Emotion & Psychopatology Lab, Yale University.

Joshua Knobe is an Experimental Philosopher; Associate Professor of Philosophy and Cognitive Science, Yale University.

Rob Kurzban is an Associate Professor, University of Pennsylvania specializing in evolutionary psychology: Author, *Why Everyone (Else) Is A Hypocrite.*

Sendhil Mullainathan is Professor of Economics, Harvard; Assistant Director for Research, The Consumer Financial Protection Bureau (CFPB), U.S. Treasury Department (2011-2013); Coauthor, *Scarcity: Why Having Too Little Means So Much.*

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.

ALSO PARTICIPATING

Daniel Kahneman is Recipient, Nobel Prize in Economics, 2002; Presidential Medal of Freedom, 2013; Eugene Higgins Professor of Psychology, Princeton University; Author, *Thinking Fast And Slow.*

Anne Treisman is Professor Emeritus of Psychology, Princeton University; Recipient, National Medal of Science, 2013.

Jennifer Jacquet is Clinical Assistant Professor of Environmental Studies, NYU; Researching cooperation and the tragedy of the commons; Author, *Is Shame Necessary?* (forthcoming).

TABLE OF CONTENTS

I. Sendhil Mullainathan: What Big Data Means For Social Science

II. June Gruber: The Scientific Study of Positive Emotion

III. Fiery Cushman: The Paradox of Automatic Planning

IV. Rob Kurzban: P-Hacking and the Replication Crisis

V. Nicholas Christakis: The Science of Social Connections

VI. Joshua Greene: The Role of Brain Imaging in Social Science

VII. Laurie Santos: What Makes Humans Unique

VIII. Joshua Knobe: Experimental Philosophy and the Notion of the Self

IX. David Pizarro: The Failure of Social and Moral Institutions

X. Daniel C. Dennett: The De-Darwinizing of Cultural Change

We've known big data has had big impacts in business, and in lots of prediction tasks. I want to understand, what does big data mean for what we do for science? Specifically, I want to think about the following context: You have a scientist who has a hypothesis that they would like to test, and I want to think about how the testing of that hypothesis might change as data gets bigger and bigger. So that's going to be the rule of the game. Scientists start with a hypothesis and they want to test it; what's going to happen?

SENDHIL MULLAINATHAN
What Big Data Means For Social Science

I'm going to talk to you today about a project that I've started in the last year or two. This type of thinking, this type of work, is going to be one of the challenges social science faces in the coming three, four, five, ten years. It's work exclusively with Jon Kleinberg. For those of you who don't know him, Jon is a computer scientist, one of the preeminent computer scientists. He's probably the smart one of the two of us, but I'm the pretty one so it's better that I'm being taped.

This is work that starts with the following observation that lots of people have had, so it will be trite to start with but we just have to live with that. The observation is that data sets are getting bigger, and bigger, and bigger in a fundamental way. As the size of data grows, what does this imply for social science? For how we conduct the business of science?

We've known big data has had big impacts in business, and in lots of prediction tasks. I want to understand, what does big data mean for what we do for science? Specifically, I want to think about the following context: You have a scientist who has a hypothesis that they would like to test, and I want to think about how the testing of that hypothesis might change as data gets bigger and bigger. So that's going to be the rule of the game. Scientists start with a hypothesis and they want to test it; what's going to happen?

Now, you might have heard of the scientific method; that is a thing that some of us use. Some people would argue, as an economist I don't really use it, but, in fact, I do. The heart of the scientific method is what you might call "deduction." You start with this hypothesis, it says variable X should matter, you go to your data set, and you see if, in fact, X does matter. If it doesn't, that's a rejection, if it does, that's an acceptance, and we move forward in that way.

What I want to argue today is that that very basic approach changes. A deduction may not be the best thing to do once data sets get really

big. To do that, instead of doing science for a bit, I want to do science fiction, which I'm sure, to make another economics joke, is what many of you think economics is. This science fiction story is going to be set in the nineteenth century. It's going to be set by a budding medical researcher in the nineteenth century who is interested in why people are dying in hospitals—a biological theory of why people are dying.

Somewhat anachronistically, she stumbles first upon a theory—before germ theory, before any of this—a theory of a mind-body connection. This scientist is convinced that what drives people, what drives mortality, is optimism. When people are optimistic, they are more likely to survive. When people are pessimistic, they are more likely to die. That's her theory.

To test her theory, following a good deductive process, she says, "I'm going to turn this optimism theory into a testable hypothesis. What's the hypothesis? Well, if I see my neighbor die, that's got to make me kind of pessimistic." So her hypothesis, or her empirical hypothesis from a theory, is that when your neighbor is sick or when your neighbor dies, you're more likely to die. She would like to test that deduced hypothesis; in fact, she does it well. She finds a hospital where neighbors are randomly assigned. Terrific, now we have causality. And she goes ahead and says, "Okay, now let's see what's in this data." She goes through her data and she finds her theory is accepted. Terrific. In fact, when neighbors die, the patient is more likely to die. Good news for the optimism theory. Of course, hopefully by now, you know that it's good news for some other theory, but to this researcher with her theory, she's kind of accepted it.

What I want to do now is put a pause on that for a little bit and say now this is where the science fiction part comes in. Imagine we live in some steampunk world, where this researcher actually has access to a very large dataset and a computer to use that dataset (even though it's the nineteenth century). She's got many, many patients, and lots of detail about all the characteristics of the patients, all the characteristics of the treatments, everything that's going on. Could she have done something different, and what could she have done? That's the question I want to ask in my little steampunk science fiction story.

To understand what she could have done, I want to take a little pause here and give you a sense of the way in which I think big data has transformed the world of artificial intelligence and how we can use that insight to transform the way of social science. Some of this may be very familiar to you, so be a little patient with me. I want to think about the classic problem of artificial intelligence—natural language processing. I remember when I was a computer science undergrad, this just felt impossible, like how on earth are we going to get a computer to understand anything?

To just understand how impossible it is, I want to take, in this big world of natural language processing, a tiny little problem called, "word sense disambiguation." What is word sense disambiguation? Take the word "bank." In a sentence, does the word "bank" refer to riverbank? The financial institutions, which seem to get in trouble all the time? Or does it refer to bank left, bank right? It can mean many things. How, in an algorithm, parsing a sentence is to determine what it means? People working in artificial intelligence tried what you and I might naturally try in this situation, they said, "Let's figure out what are the rules that might help disambiguate." So people started writing down the rules, introspecting, thinking about what might work. Very smart people went at this problem, and they made - I think this is probably not a precise number- but approximately zero progress. This was a problem. And this is true of all natural language processing. Roughly, when I left computer science, this felt like this impossible task, that maybe in 200 years we will one day have some brilliant insight and crack it. Now I have SIRI right here on my phone. What happened?

When you go back and look, what happened was not some brilliant insight deep and sitting in my phone, what happened was big data. What I mean by that is, you give an algorithm millions, billions of instances of the word "bank," in which you say, "In this case it means 'riverbank,' in this case it means…" you just give it tons of learning data. You don't think about exactly what rules it's going to use. You just code up lots of features, every feature you can imagine. Throw it in there. Just throw it in. Then you have this algorithm learn the associations that predict river bank, and the more data you give, the wider you make the dataset and the better this thing gets towards predictive accuracy.

For scientists, it's about as annoying as things can get. You're like, "How does this thing work? I don't know, but it works. Isn't that great?" So in some sense, this thing, which is the use of big data, is almost quite different than what we tend to do. We tend to form specific hypotheses, but what I want to argue is that this thing can be used for this activity. So now to do that, let's go back to our steampunk world and think of what that medical researcher could have done.

She could have done the following thing: She could have said, "Okay, I've got all of this data, and I've got all of these variables. Well, let me go through all the variables and check off the ones that have anything to do with optimism. Okay. Roommate health may have something to do with optimism." Let's suppose in this data there's nothing else, but there may be other variables in other datasets, like, how much is the patient smiling? Oh, well, that has to do with optimism. She checks out each of the variables that have something to do with optimism.

"The rest," she says, "I don't know what this stuff is. It has nothing to do with me." Okay. Great.

Once she's done that, then what will we do? What we would do is we'd say, okay, here's the algorithm, let it go to work. What it's going to do is come up with the best predictor it can of patient mortality, and we're going to ask the following question: Does it use the variable that you thought was important or not? Now, this is fundamentally distinct in the deductive test. The deductive test is, I take the variable I thought was important and ask whether it predicted. This test, which we're calling the inductive test (and I know half the people here know what the word "induction" means better than I do, so please don't break my heart and tell me this is not induction) is, instead of looking whether this variable matters, you say just figure out all the variables that matter, and is yours one of them.

Think of what would have happened if she had done this with her particular example of roommate health. She would have said, "Oh, look, roommate health matters by itself, but look at this other variable, which seems to be taking up a lot of the action. Did doctor wash hands between patients? I don't know what that is, but look, it's starting to soak up the impact on roommate health. And look, this other variable starts coming in. Shared scalpel." And so on, and so on, and so on. And here's another one that seems to come in: Was the disease tuberculosis, or was the disease buggy accident? Because in this cyberpunk world they don't have cars yet, they just have buggies, it's a very bleak world.

We collected data, and it is expensive to collect. What do you go out and collect? The stuff that you think matters. That's why deduction is so powerful. But once you collect all kinds of things, then you will have the ability to look at all these variables and see what matters, much like in word sense disambiguation. We're no longer defining rules. We're just throwing everything in.

As we go forward what she'd see is her original variable—roommate health—which she thought was proxying for optimism, is being killed off by these other variables that her theories have nothing to say about, which would make her uncomfortable that she actually had discovered evidence for optimism. She discovered an empirical relationship—roommate health mattered, but remember, theory testing is not empirical relationships, it is testing a theory. In this case,

she would have discovered that actually the data is not confirming her theory, as she thought it was.

In fact, if you think of what's happening here, this is almost what the scientific method looks like as it's done by humans, as it unfolds over time and iterations of people by people. If absent—the computer and all this data—she would have run her one test, someone else would have come and said, "Oh, I reran your thing, but I'm noticing this fact." Someone else would have said, "I noticed this fact." And slowly over time, someone else would have come up with another theory, which you might want to call it, say, the germ theory, and then that theory would have overwhelmed it. With these large datasets, we have the ability to supercharge that process at least a little bit, but we're no longer doing the inductive test. So that's the heart of it. So the heart of what we're calling the inductive approach, the inductive scientific method, is, just like with deduction, starting with the hypothesis that you'd like to test, but instead of looking just for the hypothesis, letting an algorithm determine what is the best predictor, and then seeing.

There are a few things to note about induction in this meeting. The first thing to note is induction is only as powerful as your dataset is rich. You have very few other variables besides your theory. Now, why is that important? That's important because this is the reason this type of approach was never really practical until the last ten years. We collected data, and it is expensive to collect. What do you go out and collect? The stuff that you think matters. That's why deduction is so powerful. But once you collect all kinds of things, then you will have the ability to look at all these variables and see what matters, much like in word sense disambiguation. We're no longer defining rules. We're just throwing everything in.

Second, I want you to observe that this is not, absolutely not, about causality. The original researcher's mistake wasn't that she misunderstood admitted variable bias—she randomly assigned roommates. It's about the interpretation of causal relationships we discover. It can also be about causality, but the core issue is not, I'm telling you admitted variable bias, I'm trying to tell you about the testing of hypotheses. So that's the approach, and that's what I think has some room to combine this movement in big data with what we tend to do.

Let me tell you a practical application of this. We decided to try and test this with one of the old facts in behavioral finance. Behavioral finance is the application of the work Danny and others have done to financial markets. I would say that, historically, behavioral finance has been one of the big reasons why behavioral economics really took off. In a way, if you were thinking of attacking some foreign country, this is not quite the capital of economics, but once you take that territory it

becomes easy to take everything else because people are like, "There's no way these psychological biases could matter in markets," and well, they do. And so then you're like, "Oh, we can take the rest to the ground."

One of the early facts in this area was something called the disposition effect, very close to Danny's heart. The disposition effect states that because people dislike realizing losses, what we should see is that when somebody holds the stock that they bought at $10, they're much more likely to sell if that stock is at $9 than if that stock is at 11 because you just don't want to realize that loss. It's quite intuitive, and it's kind of an interesting application of loss aversion combined with one other assumption.

In fact, one of the beautiful papers in this literature was Terry Odean's who went and got a very large dataset of traders—about 100,000 traders from a brokerage house—and what he showed was: Using good deductive science, I took this large data, I looked at people who were in the gain domain, who were in the loss domain, and the proportion of gains realized to losses realized was huge. Gains were realized at about 60 percent higher rate than losses. Very good deductive science.

We thought, well, let's go and just apply inductive science, because this is a large dataset with lots of features, and when you apply machine learning techniques, and you let the algorithm go through this huge set of variables and pick out variables individually that it thinks are important, this is no longer a prediction. If I have to look one by one, in fact, there's good support for the disposition effect. This algorithm is rediscovering loss aversion, because it finds this gain variable and says if I have to use one, this is one of the ones that I would use. It discovers a few others, but it's as if—I'm going to put you out of work soon, Danny—it discovers loss aversion, which is kind of interesting.

But then if you say to the algorithm go ahead and use all the variables you could to come up with the best predictor, it doesn't care about the disposition effect. Absolutely uninterested in that effect. The disposition effect, much like roommate assignment, appears to have been merely a proxy for some deeper avenue of behavior that really has nothing to do with disposition. Disposition has approximately zero predictive value when you add it. So if you said, here's an algorithm, I hide from it the disposition effect variable, or even the gain domain, I even hide from it anything to do with the purchase price so it can't possibly know anything about the disposition effect.

We thought, well, let's go and just apply inductive science, because this is a large dataset with lots of features, and when you apply machine learning techniques, and you let the algorithm go through this huge set of variables and pick out variables individually that it thinks are important, this is no longer a prediction.

Here's another algorithm to which I give privilege advantage of my theory, which is: purchase price matters. It turns out the two algorithms do exactly as well. There is no benefit in knowing the purchase price. What we thought was the disposition effect appears to be a proxy for something else. It's a little unsatisfying, but lots of rejections are unsatisfying. One thing that's interesting about the induction test is, unlike deduction, where we are told it doesn't work, induction actually gives us a little bit more; it gives us at least some sense of north because it tells us these are the variables I used to kill the variable you care about.

In this particular case, what killed disposition? Well, two things appear to kill disposition. The algorithm discovers this variable, which we call quartile. What is quartile? Quartile is the price that you have right now—the stock that you're seeing—where does it fall in the distribution of prices over the last 180 days, or that you've seen, for example? If it's in the top quartile, people are much, much more likely to sell than if it's anywhere else. Now, you can see how that's somewhat correlated with whether you're in the gain domain, but it has nothing to do with your purchase price. It's just where you were sitting.

The other thing that surprisingly seems to matter is if you just look, the last pattern of three prices you saw make a big difference, so, up, up, up—people are very likely to sell. Interestingly, down, down, down—people are very likely to sell. So the pattern that really matters when we put it all together is, you're in the top quartile and the price goes up, up, up, or you're in the top quartile and the price goes down, down, down. That's where a disproportionate amount of the sales happen. Of course, because gain is correlated with being in that space weekly, gain matters. But if you had that variable, which the algorithm discovers, it's not just that gain no longer matters, it's this stuff is much, much more predictive.

Just to close it out, that's the application. The things that we're thinking about now are, in some sense this is, hopefully, a way to combine what I think is the principled elements of science—hypothesis, test—with the data mining inductive—exploratory—just stuff comes up that is hard to interpret aspect of big data that has now

allowed the datasets that we have.

PIZARRO: The big data approach, you're so right about how this is increasingly a method where we're going to reach discovery. I worry a little bit that, well for one, we've had data, maybe not big data in the way that we speak of it, so this approach is weird in that we talk about the scientific method as first formulating a theory and then generating a hypothesis, and then testing it out, because that so often is not how it happens in real life. You can go back to looking at Tycho Brahe's data, right? He was a big data guy. All he did was collect data. Then Keppler comes along and finds them, and says, "Hey, there's some rules to this." But it takes Newton to come and actually offer an explanation. This is what I want to ask you about. Explanation is the heart of the scientific method, and I fear that big data yields better predictions about the future, but we lose sight of getting to the more basic general principals that might actually yield predictions in completely different domains.

MULLAINATHAN: I could not agree more with you, and maybe I wasn't clear enough. I absolutely am in no way proposing that we use big data to induct a hypothesis, and so on, and so on. I was still maintaining, I think, the good features of science, where wherever your hypothesis came from, that's the starting point. The goal is very much to use big data to test that hypothesis. Does that make sense?

PIZARRO: Yes. And I wasn't saying that you were ignoring that feature. I really wanted to hear your thoughts on whether the focus on the success of predictions will possibly make us lose sight of the work that we have to do to try to yield more general explanations—that we might abandon the quest for the more basic because the prediction is so powerful that we just keep collecting more and more data, and saying, "Well, I don't care. This captures the most variance, so that's what matters." I feel like we lose some deeper understanding when we're so focused on that. Do you think that this is actually a danger, or do you think that this is not at all?

MULLAINATHAN: It's a great question. I don't think it's a danger, and here's why. I think right now the conversation is there, but having worked a lot with these types of techniques now, very quickly, just pragmatically, you want to cross it—that is, to say there are domains where it's really clear you can just put in your prediction method and plow forward. But in lots of domains, you very quickly have to get interpretations for exactly the reason that you said, which is we're trying to start with big data, but we're trying to move to somewhere where we don't have much data. It's happened so often, but I think it's

just because this is like a new toy that's appealing in this way, but just pragmatic. I don't even think it's going to require anyone to say this. Pragmatically, you start using it, and you're like, "Oh, wait, I can't," and then all of a sudden it breaks. And that's been my experience.

CHRISTAKIS: I had a narrower reflection, not on your broader point, although I have some thoughts on that, too. On the last issue of prediction about when traders might trade, it reminds me a little bit about the challenge physicians face with prognostication. In a way you're talking about position, velocity, and acceleration of a particle. So when you're trying to predict whether a patient will live or die, initially, people put a lot of credence in what is the patient's health status right now. So a patient that, on a ten-point scale was an eight, would be predicted to live longer, have more prospects for survival than a patient whose position was now four.

But, of course, it matters a lot if I told you the patient who's four was yesterday a ten, or vice-versa, the patient who's four was yesterday a three, and the patient who's an eight was yesterday a ten. Now you might make a different prediction about what's likely to happen. And then, of course, you've got the third moment as well, so you can go downstream. So there's an analogy from what you're describing. Eventually, you could even imagine that the velocity is much more important than the position, and dispense with the position altogether.

The other thing that reminded me of what you were describing in that example is the asymmetric loss aversion that physicians have with they make prognoses. So if you imagine that a patient's survival, like the classic "how do you price a piece of real estate"—the classic Zellner—the real estate agent has to pick a price to sell a house, and you can't publish on the front of the house a density function for what price you would sell it at. You have to pick a single price. If you pick too high a price, then maybe you run the risk that the house doesn't sell, and if you pick too low a price, well then you run the risk that you left money on the table. So you calibrate, and you pick a point, and you have to balance these two losses, not selling versus leaving money on the table.

The physician faces a similar problem, which is, "I have to make a prediction to you as to how long you're going to live, you're coming to me for treatment. If I over predict and you die early, then I lose face, and I'm embarrassed. If I under predict about how long you're going to live, well then I maybe look fantastic, or, you know, make treatment decisions otherwise." It turns out that one of the deep reasons physicians consistently overestimate survival and miscalibrate is they feel very differently about selling at the dollar loss than they do about selling at a dollar gain. So a one-month overestimate of survival

means something very different to them in terms of their loss than a one-month underestimate. So just two analogies to your discussion.

BROCKMAN: How do these ideas manifest in your work in government?

MULLAINATHAN: I've done various things, and I've done work with CFPB and worked at Treasury. I think this is quite distinct, and I think this touches back on the big data center question. This really is me trying to struggle with the following fundamental issue, which is, if we maintain the rigid rule that science is a hypothesis test, I'm trying to ask is there something different about science when data gets very large? To me that's interesting because I had always just presumed, until I started down this path, that when data gets very large, the only thing that changes about science is that we have more power. Great! It's almost like the focus gets sharper, and that's all. We continue what we're doing, but because of sharper focus, maybe we can look for smaller effects.

In fact, this stuff has convinced me that it's possible that the qualitative nature of science itself changes. I know that from talking about it in the realm of social science, because we understand that, but I suspect the same ideas can be used in other areas—we're not sure—but in other areas where we also have large datasets and we're testing scientific hypotheses.

DENNETT: I'm trying to put my finger on what I feel is missing from this new approach, and I haven't got it very well figured out, but it reminds me a bit of credit assignment problems in AI and in debugging, and also in connectionism, where you've got a connection to this model, train it up, it works, and I'm thinking of why? How? And there are some techniques, which can tease out pretty well what's doing the work. I think Sinofsky and Hinton have some, for instance. But I think we need something more here.

I guess what worries me is that we'll come to settle for a big data prediction, and just abandon the search for understanding and say, "Well, come on, that's a nineteenth century idea, a twentieth century idea. Who needs formuli? Who needs understanding, when we just push the button, and the algorithm gives us the prediction?" That is, to me, a depressing prospect.

MULLAINATHAN: That's related to your question. So let me tackle it, a bit more. Why I don't think we'll settle there, is that several people have written about this. Donahoe at Stanford has written some very good things about this. There's sort of a misnomer in the word "big" in big data. This may be familiar to all of you, but at least let me just talk

it through. We could break the word "big" into two parts: Long data and wide data. What do I mean by that? Long data is the number of data points you have. So if you picture the data set as sort of like a matrix, or written on a piece of paper, length is the length of that dataset. The width is the number of features that you have.

These two kinds of "big" work in exactly the opposite direction. That is, long is really, really good. Wide, some of it's bad, and it poses a lot of problems. Why does wide pose a lot of problems? Picture the prediction function working as a search process. The search processes find the combinations of features that work well to predict why. You could see, with just a little back of the envelope calculation the mathematicsare such that as the data gets even a little bit wider, this thing is growing exponentially, I mean, just crazy exponentially. As a result, when data gets wider, and wider, and wider, the problem gets harder, and harder, and harder, and algorithms do worse, and worse, and worse. As the data gets longer and longer, algorithms do better and better.

Why I'm saying this is because ultimately as data gets bigger, it's not like it's reducing the need for "curation." What is curation? Curation is the human element of going in and saying, sure, we've got lots of these variables, but let's pick these few that matter. And I think this is a part that's missing in what we're talking about. Something interesting is happening in the curation process. Deduction is one end of curation, which is I pick the one feature I thought mattered and I'm putting that in. With big data we can put in more than the one feature we thought mattered, but we can't in many cases, throw it all into the mix. In a few cases we can. The dataset is sufficiently thin that you can throw it all in, but in most cases, you are left with a pretty massive curation problem. And in most machine learning prediction applications that's all swept under the rug. The truth is, some domain expert had to curate this thing; somebody had to decide because there's just way too much. So there is some interesting and important interaction there.

Notice, my comment is a statement about the inherent mathematics of the problem. This is not something where more computing power is going to solve this. It's not something where more data is going to solve this because this is a double exponential situation where when things get sufficiently wide we're talking about more data than there are atoms in the universe. We're not even close to that.

So this width problem, in blowing things up, is at the heart of why fundamentally we have to have explanations. I'm not saying we have an answer to that, but I think that that's why there is the need for having a hypothesis, testing, and then continuing.

DENNETT: It's like search and chess too. If you had 7500 first move possibilities…

MULLAINATHAN: That's actually a great example. It's worth noting the difference between search and chess and this example. In search and chess, my input is not data, it's just computational power to walk down the tree because the world is given to me. So the only constraint I have is steps down this tree. In this machine learning world, the length is the limit. I can't just say, "Get me more data!" Well there are only 6 billion people in the world, I've given you everybody, I can't give you any more. So the length constraint is a very real constraint on the width that we can search through, which is why I'm optimistic that we're not fundamentally going to somehow lose the need for explanation.

KAHNEMAN: Let's start at the other end. Let's stipulate for a moment that loss aversion is real, and that people really hate to realize a loss. Now, that would predict some disposition effect. Not predict the disposition effect. That is, there is a difference, and I think that's extremely interesting. It might not be the best predictor, but it might still be true. So when you say "kill the variable," you might have killed something that, in fact, is valid, and interesting, and important. It is just not the best predictor of when people sell or don't. We know that that can happen. I'm wondering about that. Because then you might actually be losing something through big data, because there is a consistent story. There is a broad story about loss aversion, and it seems trivial that when somebody has a choice between rewarding themselves by selling something that has gain, and punishing themselves by selling at a loss, they're more likely to reward themselves. If that isn't true, if big data can kill that hypothesis, then we're in real trouble.

So here is a hypothesis that somehow must be true, and what you have shown is that it's not the best predictor of people's choices. And because it is not the best predictor, it's not an independent predictor, you have come up with a conclusion, which is a strong conclusion—we've killed that variable. I am not sure you have killed it. I'm not sure you should kill it. That's the question I'm raising.

MULLAINATHAN: I think there are two ways in which we could be wrong in saying we've killed the variable. One is that we focused narrowly on stock sales, and that's the only thing in the world, and we're trying to figure out did we say something meaningful about the disposition effect there.

The second way is look, we don't just care about stock sales, we care about a variety of things. Even if this variable is not the most important predictive variable here, it's possible that it's the third most

important predictive variable, but across many, many, many domains, and so as a result, it's a foolish thing. So I think these are two separate elements of it, and let me take them in part.

The second one, I think, is the easiest to talk about, which is that I think what I find valuable in the inductive test at some level, is it's worth just comparing it to the deductive test. In that sense, induction, will lead to many more of these instances, where variables look unimportant, or far less important in this context, but I think that's something social science has to come up with, which is that our theories can never be so good that in every context they do very, very well. So we have to be amassing evidence across a variety of contexts, and so I would completely agree. If you said to me, "We are going to go and get data on housing," do I believe that this fact that disposition is less important here, or unimportant here, given the other variables, does that mean disposition wouldn't be important here? I don't know.

I think of this as having supercharged one part of the process, but by no means having in any way supercharged the other part, which is important for social science, which is to look context by context, and start to understand. And I totally would believe a world where we found that disposition was third most important or second or was important if we didn't know, but because the quantile is not something you can look at in other places. I'm agreeing, though in this case, I will say that it gave me a moment of pause because when we look at the variable that matters, in this case the quantile, and the price dynamics, that made me feel that now that I go to the housing world I would also be curious about those variables. So that gives me guidance, but kill is too strong a word for the second one. I completely agree with that. It's more saying we've learned the signal that this thing wasn't independent.

KAHNEMAN: Then I have a small technical question. I assume you did it the way I would have done it, but the disposition, as I understand it, is that you take an individual who has a choice, that there's a portfolio, there's some winners and some losers in the portfolio, and the question is: Which is he more likely to sell? So is that the way it was set up? In general, that is, if you don't set it up as a sales problem, if you just predict what, as a choice problem, between selling a loser and selling a winner, if you just predict selling, you could get something entirely different, and loss aversion would be completely irrelevant.

MULLAINATHAN: No. It was very much set up as you're given this string, here's a person, and you're given this string of everything that's happened to them, as well as their purchase price. And so, therefore, you can say, okay, for this person, in this stock, here's everything you know about them, and now you'd like to make a

prediction about is that person going to be selling or losing. Just to get a sense, that type of string, there are many such strings, and so there will be ones where that person will be in the loss domain on another stock. And so that's the signal that I think the Odean paper correctly got at, because it collapsed all those strings down and said all the times that you're in gain versus all the times in loss. But I don't know if that's what you mean.

KAHNEMAN: No. My question was whether you restricted your analysis and your prediction to cases in which the portfolio included a winner and a loser, because if you didn't, then your results get a completely different interpretation. The disposition effect is about a choice. The story you were telling us is about a prediction, and you could have a portfolio that is all winners or all losers, and if you allowed those portfolios inside your analysis, you could get your result, and that would even bear on the disposition effect.

CHRISTAKIS: And more particularly, does the outcome of other stocks held by the same trader affect the likelihood of selling or buying the index stock. Right? Across stocks, not across individuals. Are you comparing individuals who are at a gain: they are more likely to sell than individuals at a loss, versus comparing within the individual?

MULLAINATHAN: That's right. We've done a little bit of stuff, but maybe not as much as we'd like, but we've done a little bit stuff on the entire portfolio. Is the entire portfolio in a gain? Where is the stock relative to the other stocks in the gain? Those never get off the ground, which I think is sort of a mental accounting thing, that these things are being individually accounted for. But we haven't done, and we can do this, it's a great idea, is just literally compare this person and say here are two stocks that the person themselves held at this time, which one is more likely to be sold? That's a valid experiment.

KAHNEMAN: I mean that's interesting, because of something it tells you about big data and the analysis, that is that when you construct it in the deductive way, when you construct a story about the disposition effect, you really very clearly have a choice in mind, and …

MULLAINATHAN: To be fair, the Odean experiment wasn't the one you had either.

KAHNEMAN: No.

MULLAINATHAN: When we did deduction with Odean, we also did just take all winners and compared all losers. So in some sense, I take your comment, but I think that's almost a slippage, a mental slippage

we all have had around what is the prediction, and maybe this is nice, because it's really forcing us to sharpen exactly what the prediction is.

~~~
~~~

What I'm really interested in is the science of human emotion. In particular, what's captivated my field and my interest the most is trying to understand positive emotions. Not only the ways in which perhaps we think they're beneficial for us or confer some sort of adaptive value, but actually the ways in which they may signal dysfunction and may not actually, in all circumstances and in all intensities, be good for us.

JUNE GRUBER:
The Scientific Study of Positive Emotion

It's really nice to have a conversation with everyone here today. I've met most of you, but for those of you I haven't, I'm June Gruber, and I'm a psychologist at Yale University, and I direct the Positive Emotion and Psych Pathology Lab. What I'm really interested in is the science of human emotion. In particular, what's captivated my field and my interest the most is trying to understand positive emotions. Not only the ways in which perhaps we think they're beneficial for us or confer some sort of adaptive value, but actually the ways in which they may signal dysfunction and may not actually, in all circumstances and in all intensities, be good for us.

I thought I'd first start briefly with a tale of positive emotion. It's a really interesting state because in many ways it's one of the most powerful things that evolution has built for us. If we look at early writings of Charles Darwin, he prominently features these feelings of love, admiration, laughter. So early on we see observations of them, and have some sense that they're really critical for our survival, but when you look at the subsequent scientific study of emotion, it lagged far behind. Indeed, most of the research in human emotion really began with studying negative emotions, trying to build taxonomies, understand cognitive appraisals, physiological signatures, and things like anger, and fear, and disgust. For good reason, we wanted to understand human suffering and hopefully try to ameliorate it.

When we looked at the first scientific study of positive emotion, what we really saw is a rather simplistic treatment of it. We would see people talking about positive emotions as if they were some single uni-dimensional construct that we would call happiness, whatever that was supposed to mean. Even looking at work on what are thought to be some of the most basic universal emotions cross-culturally to man, in some of Paul Ekman's early observations, five out of six of them are negative. Again, we really had just one that was truly positive—this idea of joy or happiness. But everyone here knows that there's more than just one way to feel good, right? So it seems to be that science, though, hadn't gotten there just yet.

Furthermore, when we thought about the fact that emotions are functional—we have them for some reason, they help us serve some sort of adaptive survival purpose—positive emotions, again, were relatively neglected. When we thought of what their function was, a lot of the early treatments suggested, well, really, they're there perhaps just to undo the sort of deleterious effects of negative emotions, right? We saw work by Barbara Fredrickson, Robert Levenson, and others, that showed really profoundly that when you got sympathetically charged by some kind of negative emotion, positive emotions could kick in to kind of help you recover, or come down to some resting baseline.

Although that may be some important finding related to a consequence of positive emotion, it's not the sole function in its own right. Positive emotions are not simply healers, and they're not simply there to counteract the effects of negative emotions. They have their functions of their own, and in their own right. I think only recently have we really seen these empirical tides begin to shift, and to really turn our attention towards trying to understand what exactly positive emotions are—what their functions are for us, and I think in many ways what good are they for us.

This is an exciting time in the field, but I also think, as it's grown and gained momentum, a lot of the research has focused on trying to understand what benefits positive emotions confer for us. We know from this research there's been a robust domain of findings that have really said (not surprisingly), positive emotions seem to have some benefits for us. They help us build vital social bonds. There's even some work in the health psychology discipline suggesting they may enhance physical immunity to stressors, and some work suggesting they place some role in expansive creative thinking.

Positive emotions are not simply healers, and they're not simply there to counteract the effects of negative emotions. They have their functions of their own, and in their own right.

When I entered this field I looked at this and I said, "Well, let's wait a second here. Is this really all there is to the story?" I think here's where I've gotten most interested in this field of positive emotion and where a lot of the most interesting insights about positive emotion have really come into being in the past couple years. So when I thought about where is this field of emotion, and positive emotion, particularly in the past couple of years, this is really where my attention turned. I think it's suggesting that how we traditionally think

about positive emotion and the role it plays in our wellbeing is not at all as simple as we thought; it's far more nuanced and far more complex, especially if you think about the relationship of positive emotion to our general sense of wellbeing and how we survive and flourish—it's not some simple linear relationship.
Some of the critical factors that it really depends on in understanding the role that positive emotion plays in human well-being varies as a function of the, I would say, balance of positive emotions. I'll say a little bit more about what I mean by that, the context in which they unfold, and I would say the specific aims or goals by which we go about trying to experience this thing we call positive emotion in the first place. I think that's what I'd like to talk with you about today—thinking about this really nuanced, almost delicate, interplay between this thing we call positive emotion—this thing that we sort of have some intuitive sense, "It must be good for us, right?" and to actually say, "No, it's not quite that simple. It's a far more rich and deep relationship that I think not only tells us something about emotion, but I think just says a little bit about the role that psychological states perhaps more generally play in better understanding, or human nature."

I wanted to start out with in playing out these three themes for me was a quote that I remember first seeing by Aristotle, who I often go back to and think had many of the most prescient, fascinating observations about human emotion. As psychologists we're just trying to catch up now and sort of build some empirical data to really flesh out his observations. So he has this wonderful quote, and I'll read it to you. He said, "Getting angry or sad is easy and anyone can do it, but doing it at the right amount, at the right time, and in the right way is not easy, nor can anyone do it."

Here he picked up some of these key themes right away. He talked about the amount or intensity of an emotion experience. He talked about the timing, or the particular sensitive context in which this emotion reveals itself, he also talked about the ways that we try to achieve these states in the first place.

I want to start out with this first theme, which is really about the amount of emotions we experience. When you think about positive emotion, probably most of you see things not only in scientific form but also in pop culture that suggests that what we ought to be doing is really striving to maximize these positive states or just the general term "happiness," that what we should be going out and doing is finding ways to increase the frequency, intensity, and amount of positive emotions that we experience. What I would say is that may be going about it in all the wrong way, and that what actually may be most important is to think about positive emotion as a very delicate balance, and an equilibrium that we want to constantly keep check on.

What I'm talking about here in a way is like the magnitude or intensity of any emotion state we experience, and this applies to positive emotion as well. So here, really what we're going back to is this principle of moderation, and the idea that positive emotions are no exception to that. When you look at some of the recent data that's come out in the past few years about positive emotion, it's really suggested that intensely high or great magnitudes or degrees of positive emotion don't necessarily confer these direct benefits in terms of increasing our wellbeing, or our psychological health. Some of these findings at first, when I saw them, I found them really counterintuitive.

For example, we all think that positive emotion is something that should enhance our ability to creatively think about solving problems, that it just opens our repertoire to pick from different possible ideas or strategies. We find, though, that when people actually go beyond a critical threshold—hit a peak and pass that —they actually have a harder time solving problems effectively; they become more rigid or inflexible in their behavioral repertoires. It seems to be the case that too much positive emotion, thinking especially about these high arousal states of excitement and joy, actually leads us to become less creative.

Then the piece that I love the most is thinking about what are the action tendencies associated with some of our most common positive emotions. If we think of some of them, especially excitement or enthusiasm, they motivate us importantly to seek out rewards in the environment, to try to obtain them, and once we obtain them, to go about savoring them. In many ways I think it narrows our focus on rewards, how can we find them, attain them, and keep them for as long as possible.

What we find is that individuals who experience this sort of heightened magnitude of positive emotion (this is measured in a lot of different ways, using self-report scales, and also with children, parent and teacher-rated observations) out of balance, it causes you to neglect important threats and dangers, and pieces of information in the environment around you. And so as a result we see associations with greater risk taking—engaging in reckless driving, substance abuse, unsafe sexual practices. Some people would argue that this may help explain this one finding: looking at children who were rated in terms of their dispositional cheerfulness, and followed them longitudinally over the lifespan, and what you find is that children who were rated as more highly cheerful actually had a greater mortality risk later in life.

There could be many reasons to account for this, but I think one possibility might be, at least tentatively, that there's something about heightened positive emotion beyond a critical threshold that we need to be careful of, and think about keeping in balance. In my lab, we try

to study this in the clinical context of individuals suffering from emotional disorders. One entry point that we've begun to look at is among individuals with mania that show some characteristic signs of heightened positive emotion and this appetitive system that's kind of go, go, go towards rewards, and finding (not surprisingly) that these are individuals who engage in all kinds of reckless behavior. They wipe out their bank accounts, they destroy some of the most important social bonds in their lives with their partners through lots of sexual promiscuity. They will report when you talk to them, and I interviewed a lot of these people clinically, that they just felt so good—that nothing else could enter their mind, that it was a one mind that was really all about feeling good, and finding ways to keep that going.

I think this first theme, and it's a new theme, needed a lot more empirical attention on it. What it's beginning to suggest is something about human nature that suggests that maybe we need to put aside these conventional notions of trying to maximize positive emotions, and that positive emotion may be in line with many psychological states that are subject to this principal of moderation. We really want to be experiencing things in balance— not too little, or not too much— and in many ways it's also consistent with biological theories, postulating optimal functioning, and moving towards a sense of homeostasis, or equilibrium.

This is important because it suggests that a realm of psychology that's getting a lot of pop culture attention really needs to be cautious, and think about, in these interventions that are being discussed, how can we keep it in line with a sense of moderation? So that's one thing that's been getting attention, in the past couple of years that is interesting.

The one thing that comes next that I find even more fascinating is the idea of context. I've been talking about positive emotions very generally, and I haven't been talking about when they occur, what is the timing. If we think about a functional approach to emotions, the idea is that they have functions, but the functions are really tied to a specific context. Emotions are geared to help us find opportunities, solve challenges, respond to immediate threats, and inherent in that definition is that they arise in a particular context in which those goals are activated.

When we think about the function, thinking about positive emotion now, we need to consider the context in which it occurs. We can all probably readily imagine times when we're hanging out with friends, and it's a wonderful appropriate context to experience amusement, experience joy, but there's many contexts in which that would absolutely not be productive, and may lead to rupturing professional relationships—if you're laughing inappropriately, or if you're in a

dangerous life-threatening situation, you don't want to be standing back content with the world around you. For me, this is the piece of positive emotion that I've been most interested in, and I think has some of the most powerful implications, at least for affective scientists in terms of how we think about emotion, emotion regulation, and understanding.

In many ways, when you experience positive emotions in a context that doesn't match that function, here's where we're finding that difficulties arise, and that we shouldn't be trying to promote positive emotions at all times and in all situations, and for all people for that matter.

I start with negative emotions because those are the ones that have received still, to this date, the most attention. We can think of anger, for example, it mobilizes us to overcome obstacles, or fear that alerts us to threat and danger in the environment. These are obvious functions I don't think many of us would disagree with, but when we think about the role of distinct kinds of positive emotions, what role do they serve? In general, if I talk in a general valence level of positive emotions, they're thought to, in many ways, help us pursue personal goals and facilitate cooperative behavior. You can take a more nuanced perspective and look at the discreet types or distinct flavors of positive emotion, and there you see wonderful taxonomies that are not now being developed, saying that certain emotions like gratitude, and very different functions from pride, very different functions from feelings of contentment, and awe or inspiration. So we see that not only does the family of positive emotions have some sort of broad-based function, but that each individual variety or flavor of positive emotion serves a important goal in our lives.

The important point I'm trying to make here is that positive emotions are really suited to perform a function. In many ways, when you experience positive emotions in a context that doesn't match that function, here's where we're finding that difficulties arise, and that we shouldn't be trying to promote positive emotions at all times and in all situations, and for all people for that matter. There's been a couple of interesting findings that have come out in the last couple years that I think really hit this home.

One of them is by a psychologist, Maya Tamir. She did a study looking at what kinds of affective states promote successful outcomes on competitive tasks. And so she did a task that involved a competitive computer game. Participants were experimentally induced into either a

positive mood state, which was a high arousal state that many people would say is something like excitement or joy, compared to individuals in an angry mood state, which is also this sort of appetitive high arousal state. And then there was a neutral comparison condition. After that people played this competitive computer task. She found that those who performed best on this task were not the people who were induced to the highly positive arousing state, but those who were angry. This has a lot of implications and thinking about when you're trying to overcome obstacles, and in some competitive situations, there may be something about anger that helps motivate the kinds of behaviors that could lead to successful outcomes.

Now, that doesn't mean that we should be angry all the time when we're competing, but it suggests that, depending on what your goal is, and if your goal is to win in some competitive situation, we know that highly arousing positive states may not be the best affective state or path that's going to get you there.

We've also looked at the context of experiencing positive emotions and everyday social interactions with romantic partners. What we did is we brought couples into the lab who had been in long-term relationships. These are, perhaps, our most highly valued social relationships in our entire life, and highly ecologically valid in the context of trying to understand emotional dynamics between two people. What we asked people to do was to think of a time in their life where they experienced great suffering or personal loss, and to share that with their partner. And then we had the partner report the kinds of emotions that they experienced after hearing their partner tell of this time of suffering or loss.

Now, many of you might imagine that if you could list an array of emotions that would be appropriate in that circumstance, both to accurately interpret the significance of what your partner is saying, and be connected to them empathically, they might range from things like sadness, frustration, to compassion, which is an interesting emotion that has elements of both positive and negative feelings. What we found, though, is that in this particular sample, and this had both healthy community adults and individuals on a spectrum of clinical symptoms of mania, is that the higher people were on the spectrum of having symptoms of mania, the more they reported feeling positive during this interaction. And by positive, these were feelings of joy, amusement, and even contentment. What we were finding is that not only our signatures of emotional dysfunction we associated with experiencing positive emotions and inappropriate contexts, but that this is not surprisingly predictive of decreased relationship satisfaction. So it just tells us something about the importance of context when experiencing positive emotions.

This may seem really obvious and in some ways trivial, but every time I see some book that tells you how to maximize happiness, to think of three great things every day, and just constantly try to use this facial feedback monitoring, to put a smile on your face, what I don't see in it is under what context is that appropriate. So I always worry that what we need to be stressing more of is that emotions really only serve their function best in the particular context in which they are suited for.

For me, this says something important about positive emotion, but also our emotional states as human beings, in general, and insofar as it suggests that any kind of emotional state is only adapted for us insofar as it has a particular fit with the environmental demands or needs in that situation. In many ways, there are no absolute value judgments we can place on emotions to call them adaptive or maladaptive—good or bad—and this goes for emotion regulation as well as a field where we no longer can call certain kinds of strategies like reappraisal adaptive, or behavioral suppression, where we don't show expressivity in our face maladaptive, that this just isn't the way that these emotional states work. Nothing is inherently adaptive or maladaptive.

So that's a second theme that I think has been getting a lot of attention, what people will call context, or the context sensitivity of our emotional states, behaviors, and associated regulation strategies.

The third theme, and this is the one that I find perhaps the most important when we think about ourselves in our everyday lives, is thinking about how do we set goals that will make us feel more positive. There's a lot of strategies out there about how to maximize feelings of positivity. There's ideas that there are certain ratios we should try to obtain, or certain kinds of frequencies in which we should experience positive emotions.

When I think about this, to one, it's suggesting that we care a lot about, (as human beings) experiencing pleasant feelings, maximizing them, and trying to make them last as long as possible. Especially in the U.S. this seems ingrained in the way we think about what our rights are. We have this notion of life, liberty, and the pursuit of happiness, that this is sort of a culturally embedded value, and positive emotions are at the forefront. They're hypercognized, so you have a lot of words for them, and we put a lot of emphasis on them.

But I think the bottom line in all of this, and I'll tell you some really interesting studies by Iris Mauss at U.C. Berkeley, is that we spend a lot of time trying to find ways to make ourselves feel positive. People call this feeling happy, often colloquially, and I think recent science suggests that we're going about it all in the wrong way.

And, in fact, research is finding that the more people (1) spend time and effort trying to increase how positive they feel, and (2) the more they set as the end goal point feeling more positive, that they actually somewhat paradoxically set themselves to feel less of that very state. There has been work looking at individuals in laboratory studies, where people are told, for example, try to make yourself feel as happy as possible while listening to a piece of music, and those people that are told to do that, not surprisingly, report feeling less positive, right?

If you look at people who report these kind of tendencies in daily life, endorsing items such as: I spent a lot of time trying to feel happy, or I go out of my way to select activities that I think are going to bring me pleasure, that I think are going to make me feel good, If you bring these people into the laboratory and put them in contexts that are ostensibly positive, like you're watching a positive film, or reading a positive sort of vignette, that it's in those circumstances that they feel or self report less positive affect, you see it even more pronounced than compared to a negative film or some neutral film. The idea that researchers have tried to explain these paradoxical findings is going back to basic theories on human goal pursuit—that the goals people value determine what standards they're going to set for achieving those goals.

In many ways, you can think of, for example, someone who highly values academic achievement—they place a lot of value on that goal, it's likely they're going to subsequently set a very high standard for achieving that goal. So in many ways the more we seem to value experiencing positive emotion, whether it is excitement, or pride, or love, or contentment, the more we set that as our emotional value system, inadvertently, probably the higher we're going to set a threshold for achieving it, and subsequently set up ourselves for disappointment.

We've seen research translating this to the clinical science realm recently, finding that people who highly value the experience of positive emotion, and who put behavioral energy towards obtaining it, they're at great risk for depression, and they subsequently report at baseline, cross-sectionally too, a greater incidence of clinical diagnosis of major depressive disorder.

We really see that this is telling us something profound about just the kinds of goals we set for ourselves. For me, this is an important thing to translate to human nature, because it suggests that the amount of our positive emotion is really affected by the effort we put into it. It's almost this ironic effect. I mean we know that the more we try not to think about white bears, we think about white bears, and in many cases the more we try not to be unhappy, the more unhappy we seem to be. So it suggests that in many ways this is this paradoxical

backfiring, and in many ways that if we want to have affective or psychological goals for ourselves, then we ought not to make that the end focal point in itself, but perhaps to be focusing on other things from which those emotions might emerge.

So I mean just in closing, thinking about what positive emotion can tell us, not just about positive emotion, but more general about human nature, I think is that the relationship between our feelings, and perhaps this goes for our thoughts and behaviors as well, is way more complicated than we ever thought, way more complicated. I think we have a lot more work ahead of ourselves, and that it really depends on things like intensity of a given psychological state, the context in which it occurs, and just the way we approach trying to achieve it in the first place.

So in other words, I think balance, intensity, context, and timing are important, and as psychologists, I mean, I always go back and think how much we can learn from philosophy as we try to move forward in understanding human nature, too. So that's what has happened in the past couple years, and I hope that we continue to move forward in realizing just how much more complicated human emotions are than we ever thought possible.

CUSHMAN: I felt as if, in a way, there's a struggle in the remarks that you made between two visions of what a theory of emotion should deliver on. One vision is that what it should deliver on is: tell us whether it's good or bad, right? Then another theory is that it should tell us what its function is, and what role it plays in a psychological system. If you thought of other kinds of psychological systems, like vision, it would never occur to you to ask the question like the visual system, good or bad? The questions would be: what is it's functional role? How does it enable behavior? I wonder whether, in a way, it's kind of been a poison pill for research in positive psychology that it's the kind of thing that it seems like one would want to have a lot of, but that sets up a question which is really a red herring.

In thinking about what the function of positive psychology was, the other thing that struck me about your remarks is this asymmetry—lots of bad emotions seems like just one good emotion. You were saying we could draw some distinctions, but just at a broad level it did feel like there's disgust, there's fear, there's anger, and then there's this happiness. So I guess I was just curious why? I mean can you think of a reason, at a functional level, why we ought to have many different flavors and varieties of negative emotion, but why, if you were going to design a good psychological system, it would actually be best just to have one kind of generic goodness.

GRUBER: The answer is no, I can't. I try to think why did this start in the first place, and perhaps when we think of our early categorization systems of emotion, a lot of it came from facial expressions, right? That they were thought to be these automatic universal signals of emotion. At least when it comes to different kinds of positive emotions, they're not all readily apparent in the face. We have this Duchenne smile that's supposed to signal some kind of joy. But then there's a lot more that goes on when we think about the way that our body—through nonverbal behaviors, touch, vocal intonations—that help differentiate, at the behavioral level, positive emotion.

So I think perhaps one reason it began is because when we were looking at cultural universal displays of emotion, at least in the beginning there seems to be one that …

CUSHMAN: That's one of the tests.

GRUBER: Yes. But I think any functional account of emotion would suggest now that perhaps there may even be a wider variety of positive emotions and negative emotions; it's fascinating. there's these taxonomies that have been developed that show distinct cognitive appraisals that uniquely differentiate … you know, you can look at classes of self versus other oriented positive emotions, and then you can also get emotions that take us outside of ourselves completely and give us a perspective on the broader world—things like awe and inspiration. They're incredibly important for us, and there's recent work suggesting we have different vocalizations, and that even these positive emotions can be distinctly and reliably communicated through physical touch.

For whatever reason, this negative bias we had on emotions because we thought that in many ways they were causes of suffering that led us astray, these made us these irrational beings. Now we're catching up and seeing just how important they are for us, and the more we take a profiled approach to look at distinct varieties of positive emotion, the more we're going to better understand the different psychological functions they have and bear on our everyday lives.

But I think you're exactly right. It's an interesting thing about why there was this value system placed on emotion, but not on vision science, per se, and I think it gets even more complicated when you talk about value systems, and we think of cross-cultural value systems. Much of the positive psychology movement is driven by westernized U.S. notions of positive emotions, so as a result it's focused on high arousal positive states, as opposed to if it had emerged in more collectivistic cultures, it might have focused on more low arousal positive emotion states. So we seem to care a lot about emotions and have value systems, and I think that has clouded our

scientific judgment, and operationalization of where to begin and what these are. So I'm with you.

MULLAINATHAN: I have one question, which is you mentioned on the last thing that you were talking about two things that might be related, I'm struggling to understand. Are emotions the end goal state themselves, or are emotions merely the signal of some goal state? In other words, take a totally different analogy. Say my goal is to be rich, but the signal could be, oh, people drive BMWs who are rich. So I could say, "I'm going go get a BMW," it's not going to make me rich. So I'm just trying to understand is the emotion. I guess I mean just both normatively, so that could be one way you could be giving bad advice, to say, "Think positive, be happy, think positive." You're looking for those things that generate it, but I also am trying to understand this positively, that is, when we understand the decisions people make, how do they think about what they're trying to accomplish? Are they mistaken to be chasing this? Do they understand?

That's related to my second question, in what you talked about, you were taking these reports of emotions, and maybe there's much more literature, so I'd love to just hear more about that. When I think of my own emotional state, it's almost like an illusion in my mind that I can report it. It feels like as I gain maturity, one thing I learn is that actually my emotional state is not as accessible to me as I thought it was. And how do we think about those things? You know what I'm saying?

GRUBER: I know exactly what you're saying. To start with the first point, it's interesting because when we look at some of the most basic accounts of what emotions are in terms of are they the goal themselves, or are they simply a pathway to get to the goal? In many ways what we think of emotions doing is eliciting a certain set of actions, tendencies or behaviors that are going to help get us towards a goal. So an emotion is a signal to us; it's a source of information, and that information is going to guide us to decide do I approach or avoid that particular person. But it's not the emotion itself that's the goal. It's the emotion that gives you information to set you off on behaviors that are going to get you to your goal. So I would say many emotion theorists would not think of emotion as the goal in and of itself.

That being said, if you ask for people's everyday intuitions about emotions, and you ask them about their goals in life, well, it's to be happy. "My goal is to be happy. My goal is to feel less sad and less anxious." And even from a clinical psychology perspective, working with patients who come into therapy, their goals are often very emotionally defined. They want to minimize negative emotion intensity

usually, generally speaking, and maximize positive emotion intensity. I think what we need to do is use this information and leverage it in an educational way to say, "Well, emotions are certainly important facets of our lives, they give us information, and they signal to us something, but they're not the goal itself." And so I think that's the confusion, I think, that many people have this desire to want to move towards emotion as the goal, when it's anything but that. It's what are the behaviors that happen after the emotion is elicited, and do they take you towards or away from where you want to be going.

PIZARRO: Just a quick clarification on this, because it strikes me that we talk, and I think it's an American thing to be honest, this focus on happiness as the goal. But even you, when you're talking about your goals, you frame it as if, well, if we just stop focusing on being happy, it will make us more happy. So it's even an end goal in your talk.

My parents, being the immigrants that they are, always would tell me it's weird to say I'm doing it to make myself happy, and many times I opt to do things that will make me very unhappy, knowing so, because I have a goal. And sure, like at some ultimate level, I want to achieve all my goals, and presumably happiness is the signal that I've achieved the goals, but I don't ever feel directly motivated by an attempt at happiness.

GRUBER: That's a good thing!

PIZARRO: I think it is. I think it is. And I think that it's weird to make happiness the goal, because then let's just pop ecstasy, you know.

MULLAINATHAN: All right. Let's do it.

PIZARRO: Yes. Well, I already did.

GRUBER: No. I think you're right. It's funny because you totally caught me, and that's exactly it. So just don't try to be happy, right?

PIZARRO: And then you'll be happy.

GRUBER: And then you'll be happy, right? I think a lot of what people are focusing on now, especially in the past five or so years, is this idea of mindful acceptance, of not focusing on any one particular emotion that you ought to feel or ought not to feel, but simply being present with whatever emotions you have. And that takes the spotlight away from emotions as a goal, but more focuses on being and experiencing whatever emotions you have, and using them as pieces of information to tell you something important about the environment and about yourself. So that's one approach, and there's many different kinds of

approaches out there. But I think they're all important insofar as they're telling us to get away from looking at emotion as the end goal.

BROCKMAN: To what extent is the whole construct of happiness cultural? I noticed just the word drove Dr. Kahneman from the table. And I've never framed my life in terms of happy.

GRUBER: No. I mean to be …

BROCKMAN: And I wonder like how much of this is driven by Prozac.

GRUBER: It's a scary time right now. The word "happiness," what does it even mean is one thing, and I think people are really, this has been an age-old question of what does happiness mean, but I think the problem right now is that it's used in incredibly vague and interchangeable ways to mean all kinds of things. When this then gets disseminated to the public it becomes really tricky, because people just have this word of "happiness." Maybe it's about a bigger sense of subjective well-being, or maybe it's just sensory pleasure in the moment, but they just know that that's something that they ought to have. So we see a host of prescription rates skyrocketing, and you could hope that maybe it's just more accurate detection and diagnosis of depression in this day and age, but you have to wonder if it's more something else and being driven by this zest to minimize negative and maximize positive.

It worries me in this age of happiness, because what it also does, and I think this is especially an American problem, is pushes us away from just simply experiencing negative emotions, too, that are incredibly rich sources of information for us, and incredibly important components of what gives us rich and meaningful lives.

BROCKMAN: I mean what's the science here? You have a lab, but what does an experiment consist of?

GRUBER: In studying emotion, what does it look like?

BROCKMAN: Yes.

GRUBER: Just on a general level?

BROCKMAN: Well, on a particular level.

GRUBER: Let me think of a good experiment in our lab. One of the studies that we've done has been to try to look at emotional responses—Laurie will remember this task—emotional responses that

are self-referential. So what we do when we study emotion is we take a multi-componential approach. In one study we were trying to look at the experience of self-conscious emotions, so we had people come into the lab, they sit in front of the computer screen, and what we're doing with them are three things we are measuring simultaneously. We're measuring their subjective or self-reported emotion. That's one piece of emotion responding. It may not be the absolute truth, but it's an important component. Second, we are videotaping participants and coding their expressive signatures of emotion in their face, coding them using many different standardized systems, such as FACS or the Facial Action Coding System that look at features of emotion in the face. And then we look at their physiological signatures.What we think is that you can't say anything about emotion from any one single channel. It's going to lead you down the wrong path. In studies like this we'll have people sing along to a karaoke task, and unbeknownst to them, they have to watch the video of themselves.

What we're doing here is we're really trying to quantify when a person has an emotional response, which component is most centrally featured. Is it something about their subjective representation? Is what they're signaling or communicating to others? Is it shifts in their body? Their heart rate, their skin conductance, their temperature, their breathing. And then we also do some studies taking people into the scanner, and trying to understand their own mechanisms at that level. And I think when you study emotion, what you really need to do is at every single moment be looking at it simultaneously across multiple levels of analysis.

BROCKMAN: Your last comment that we need to bring this together with philosophy is coincidental with the recent publication of Leon Wieseltier's attack on science, saying that philosophy tells us about happiness, not science. What do you have to say about her comment?

DENNETT: I don't want to say anything about Leon Wieseltier.

BROCKMAN: I know, but do a final comment. I mean what is that philosophy?

DENNETT: You mentioned in passing a study which showed that cheerful kids had a higher mortality, risk of mortality, and I thought, no that doesn't surprise me in a way, because I remember when my children were young, I remember talking to a wise older colleague and saying, "You know, I really worry about my kids, because they're having this sort of ideal upbringing. It's a very nurturing house, and they've got books, and music, and everything is perfect."

DENNETT: "And I'm just afraid they're going to be as soft as grapes, and be completely vulnerable when they get out in the real world. They won't have been tested at all, emotionally tested." And he had a very wise response. He said, "Don't worry. They'll make their own trouble." And he was right. And recovered from it, of course, and learned a great deal from it. But there is this question of whether we're making a big mistake in trying to cocoon our children in a world of positive emotions, and shield them from ever really experiencing fear, or loneliness, or boredom, and I wonder has there been research on this.

GRUBER: Your intuition is absolutely right, and there's been some work on this. We've been doing some with a colleague of ours, Michael Norton, that many of you know, looking at this concept of emotional diversity. If you think about it just within a broader sense of ecosystems, diversity is really important for health and survival of that particular system. We started taking this looking at the inner psychological system and what is most important for well-being. And when we talk about well-being, we're talking about not only psychological function, but actually physical health functioning, so we have these large medical reports from people. What we're finding is that it's the diversity of emotional experiences that both cross-sectionally and longitudinally are predicting some of our best outcomes.

You want a mixture of things. It's fine to have some joy, but you also want sadness; you want the experience of guilt; you want the experience of loss. All of these things are really important in building a psychological strength to know how to experience these emotions, to know how to cope with them, and to get information from the world around you, too.

In terms of how does this relate to raising children, I think as much as you can expose them to different kinds of emotions, and not let any one kind predominate. I think that's what's going to be most critical, the diversity of experiences at the affective level.

CHRISTAKIS: I was just going to ask what your thoughts are on the functional kind of emotions that includes not only their internal function, but also their interpersonal function. And the thing that's interesting to me about emotions is not what you feel inside, but the fact that I display the emotion, and then not only do you read it, but you copy it, that there's kind of an emotional contagion, which is a very fundamental feature, to my eye, of emotion, so people who are depressed, you become depressed, or anxious, or happy. I'm very curious about your thoughts on this interpersonal account of emotions, not just the intrapersonal account.

GRUBER: I teach this course on emotion, and when I ask students to provide an example of a time they remember experiencing some memorable emotion, they always talk about it in a social context. Usually, it's about people, but often it's with people. And many people would say that our emotions are inherently relational and interdependent, and that the function of our emotions is not to keep us as individuals, navigating the world, but it's to connect us to other people, and to interrelate to them.

CHRISTAKIS: I would say in a way, your account, the use of information, I would say emotions might not be about the acquisition of information from the environment, but the delivery of information to the environment.

SANTOS: A conspicuous signal of mental health.

CHRISTAKIS: Of a variety of things.

GRUBER: Yes. There's been some fascinating studies looking at exactly what you're talking about, which is this mimic or contagion of emotion, and finding, for example, in married couples, that those who had the best marital quality, in terms of self-reporting satisfaction, were those who played this dance. They had this mimicry, not only at the subjective level, looking at continuous rating dials of emotion as they were interacting with each other, but even looking at physiological signatures that have been thought to co-vary with the experience of positive emotion, they were in sync with one another, at one person's level, looking at cardiac vagal tone as it shifted, and so did the others.

So it seems to be what's most important in this case is not what emotions you're experiencing with a partner, but that you're in sync with one another, and there's a sense of almost coherence between partners, not just within an individual.

~~~
~~~

I want to tell you about a problem that I have because it highlights a deep problem for the field of psychology. The problem is that every time I sit down to try to write a manuscript I end up eating Ben and Jerry's instead. I sit down and then a voice comes into my head and it says, "How about Ben and Jerry's? You deserve it. You've been working hard for almost ten minutes now." Before I know it, I'm on the way out the door.

FIERY CUSHMAN:
The Paradox of Automatic Planning

I want to tell you about a problem that I have because it highlights a deep problem for the field of psychology. The problem is that every time I sit down to try to write a manuscript I end up eating Ben and Jerry's instead. I sit down and then a voice comes into my head and it says, "How about Ben and Jerry's? You deserve it. You've been working hard for almost ten minutes now." Before I know it, I'm on the way out the door.

This is a problem for psychology not, regrettably, because I was writing anything terribly important, but rather because it highlights a deep tension in a dual process theory of the mind. From one perspective my desire for Ben and Jerry's is the product of automatic or intuitive responses—literally gut feelings in this case—and then it's a controlled, effortful, deliberative process that tries to focus on the paper and put thoughts of Ben and Jerry's out of mind. On the other hand, it would truly be bizarre to say that when I went to Ben and Jerry's it was an automatic response. I mean, I have to go through a process of goal-oriented planning. I've got to get my shoes on, I've got to get out the door. There's a mismatch between the willpower perspective and the goal orientation perspective.

No, the exciting part for me is that I feel like we're finally able to at least frame the questions in the right way such that an answer's in the offing. The questions are being framed these days, I think, by some really foundational work that went on in computer science when people tried to design artificial intelligence systems that could learn and decide and they drew a division between two broad classes of solution. One of them which has been worked out best and is very familiar to psychologists is a kind of stimulus response learning based on reinforcement history.

I want you to imagine a rat that's in a Skinner box. It stumbles on a lever once, it notices that a food pellet comes out, and so it forms an association between the environmental context that it's in—being in the Skinner box and pushing that lever—and the association between

that action and value. It says: whenever I'm in this box, the valuable thing for me to do—the stimulus is "I'm in the box", the motor response is "I'm going to push the lever", that's a valuable thing.

Remarkably, you might think that the association that the rat would form would be between pushing the lever and getting food, the outcome of its action, but it turns out not to be the case at least some of the time.

You run a procedure called the Devaluation Procedure. You put the rat in the box, it forms the association between pushing the lever and value, and then you take it out of the box and you give it unrelated access to food pellets until it wouldn't touch a food pellet with a ten-foot pole, it's completely stuffed. Then you put it back in the box and under the right conditions, it waddles over to the lever and pushes it and the food comes out and it just lets the food sit there. You know that the rat wasn't pushing the lever because it had a goal in its mind and it associated pushing the lever with a particular outcome, rather, it's a stimulus response association. I'm in the box, pushing the lever is good.

What the computer scientists were able to do was to formalize mathematically the kinds of representations that support that kind of learning, and then in a way that I won't have time to describe right now they showed how you could string together series of stimulus response associations to make local decisions that have long-run consequences that are good. Okay?

It turns out that once the computer scientists formalize this stuff and they had the equations that specified, you know, there'd be a parameter here, there's an alpha, there's a gamma, and then you go look in the brain while people are making decisions in these types of tasks, you find that if you ask, "Are there voxels in the brain? Are there regions of the brain whose response profile tracks those precise mathematical parameters?" They do, again and again and again; it's happened hundreds of time now, mostly in the basal ganglia, which is a brain region we know, for instance, is impaired in Parkinson's. If you think about what goes wrong in an individual with Parkinson's, they're not able to produce motor actions. So this is the part of the brain that's responding to stimuli and it's producing motor actions, and when it's disrupted you can become, literally, put in a kind of frozen state.

We understand that system pretty well, but it's obvious that tons of human behavior is exactly the opposite of that, it is goal-oriented. And, in fact, the rats do this, too. You can have other conditions in which you put the fat rat back in the box and it doesn't touch the lever, if you run the experiment appropriately, because now it's operating in a kind of goal-oriented planning mode. It has a particular

outcome in mind that it wants to achieve, in this case the outcome is not food, maybe it's just sitting and digesting, and then it selects the actions that are appropriate to get towards that goal. And computer scientists have been able to formalize algorithms that do this type of goal-oriented planning as well.

We know that humans use goals but how do you get goals off the ground? How do you get this planning process off the ground and make it computationally tractable?

The problem with these algorithms is that if you make a task even moderately complex they totally fail because of the computational intractability of planning. We brought up the case of chess earlier, right? Chess is a game where there's many, many opening moves and then there's many, many next moves and then there's many moves after that. It's perfectly obvious what the goal is—you want to get to checkmate, but there are so many possible paths that you could take that you could have all the time in the world and that wouldn't be enough time to evaluate each one of those paths independently.

This is a really deep problem for computer science and also for psychology. We know that humans use goals but how do you get goals off the ground? How do you get this planning process off the ground and make it computationally tractable?

There's a couple of solutions that people have focused on to try to do this. One of the solutions is to arrange your goals hierarchically. I'm going to use a metaphor here, I hope it's a helptful one. Imagine that we took a picture of this group and we turned it into a jigsaw puzzle. You can suppose one way to try to solve the jigsaw puzzle would be to randomly arrange the pieces one by one, as if we were taking random searches down that chess path, right? You'd have to go through billions of random arrangements before you ever alighted on the appropriate one. But if you organize the puzzle hierarchically, you could reduce that search space a bit.

You'd say, "I'm going to just focus on Josh. I'm going to just take the pieces that look like they plausibly belong to the Josh area, fit them together, and then the June area, and the Rob area. Then once I've gotten those units organized, then I can shuffle high level units." You don't have to try every combination anymore because you're working on little local problems and then moving big chunks of space around altogether.

That's a good start but it's not going to get you the whole way. Even just a simple task like making a sandwich; I can say in order to make a sandwich I'm going to have to have some kind of sub-goal, there's going to have to be a first step, but there's kind of an infinitude of next steps that I could take. One of the next steps that I could take would be to get the bread out of the refrigerator, but another next step that I could take would be to start the manuscript that I have to start, or to pick up my wife at the train station, or any number of an infinite number of sub-goals that a person could possibly entertain.

We're going to need some kind of a cognitive system that, in the state of having a goal, selects the appropriate cognitive action of selecting a sub-goal. An insight that occurred about a decade ago in psychology was that you could actually maybe use the basal ganglia—that old rat-like stimulus response learner—to solve that problem, too.

Here's how it works. When we talk about the rat, we talk about the rat being in the perceptual state of a Skinner box and having learned the motor action of pressing the lever. But by analogy you could say that the rat would be in the internal perceptual state, the kind of conceptual state of having the goal of, say, making a sandwich (it's a very smart rat) and then what the basal ganglia has to learn is a particular cognitive action rather than a motor action, which is the appropriate cognitive action in that conceptual state.

You could learn that when you're in the conceptual state of wanting to make a sandwich, the next cognitive action to take is to select the sub-goal of getting bread and load it into the goal unit. Then if you're in the cognitive state of having getting bread in working memory, in your "goal slot," then the next cognitive action that you're going to take is walking over to the refrigerator. This is a way of using the kind of simple machinery that we understand of stimulus-response learning and getting it to perform the individual computations necessary to do a much more sophisticated type of goal-oriented planning and action selection.

One of the remarkable things is when you go back to the Parkinson's patients who are not able to produce motor actions and you ask them what it's like to be in that state of being sort of frozen in this motor sense, they'll talk about feeling cognitively frozen, too, as if their thoughts are moving with incredible slowness or they can't bring the thought to mind that they want to bring to mind. It's quite different than other types of motor impairments. ALS deprives someone of the capacity to produce a motor response but their thoughts are moving just as fast as they ever were. And so, a bunch of research has now shown that that region of basal ganglia is interacting with working memory in order to facilitate the movement of information in and out of working memory.

There are a few things that I find exciting about this. One of the markers of progress in psychology is when you can exorcize just one of the ghosts from the machine—when you can take one of those points in science where you had to say, "And then a miracle happens." We knew we had to make a sandwich so the obvious sub-goal was get bread, but that's a little bit of and then a miracle happens. How did you actually get from Point A to Point B? And this starts to show us the way that you can do that.

The second thing that I like about it is that it teaches me something about why it is that I keep ending up at Ben and Jerry's? The idea is that my basal ganglia has learned that when it loads the sub-goal "get ice cream" into working memory, good things happen. When I'm trying to work on my paper, what's happening is that this basal ganglia that really loves ice cream keeps saying, "Oh, you know what would make me happy? What if you had the goal of then going and getting some ice cream?" What that suggests (this is the third thing that I love about this area of research) is a new way of thinking about what automatic and controlled processes are and how they relate to each other.

For a long time we've talked about automatic and controlled processes as if they were systems, as if we were going to go into the brain, perhaps, and actually find completely dissociable mechanisms. One of them does the dumb stimulus response learning thing, another one of them is going to do the smart goal-oriented planning thing. What this suggests is that the controlled processes are really just a kind of adjunct—they're an add-on, an optional feature, like an app, that you can run on the lower level stimulus response system. Specifically what they do is they take a system that probably evolved to mediate between perceptual states and motor actions, and then turned it inwards and allowed it to operate on conceptual states and cognitive actions.

This feels deeply true to my intuitions about how controlled cognition works. Let's take a classic example of a controlled cognitive process, say, doing a calculus problem. Well, maybe let's make it simpler, let's just make it long division. If I think through long division, the individual cognitive operations that I perform each seem to bottom out in a kind of a habit or an intuition. Like, eventually I'm just going to have to say that five minus three is two. It's not as if when I get to five minus three is two I'm still doing something controlled. When I get to five minus three is two I'm in a conceptual state—five minus three—and then almost habitually two just pops out as the answer to that question.

If you think back to when you were in kindergarten, you actually had to learn that cognitive habit the same way that the rat has to learn

about pressing the lever and getting a reward. That is, the teacher ran laborious drills on your times tables and on the shortcuts of division and the specific operations that you would have to perform at each step in the process of doing long division, right? These are the things that excite me the most about this area of research and reinforcement learning and the ways that it connects with neurobiology.

I said at the beginning that I regard this research as a kind of a promissory note—that where we are right now is that we're starting to be able to frame the questions in the right way but that a lot of the answers are still elusive. It might seem as if maybe I'm overselling the research because I've been presenting some of these ideas as if we had all the solutions worked out, and I want to start by saying, well, we don't. People have a hunch that it's going to work something like this but getting all the nuts and bolts to fit in place is still to be done. But there's also a much, much deeper question which I think has largely been ignored and is one of the areas that I'm excited to move into over the next couple of years.

...humans as a species stand apart from all other species in our ability to engage in controlled cognition, to organize our lives around very distant goals, to inhibit more automatic responses and hold and use flexibly information in working memory. We should ask ourselves what is it that enables humans to do that? What are the other things that are unique and special about humans that might explain why we're able to engage in controlled cognition?

What's been ignored is the learning problem of getting the right cognitive stimulus response habits. What do I mean by that? When we plan very complex things, like John planned this lovely weekend for us, and there are a lot of pieces that had to come together, from the film crew to the specific set of speakers, to getting these tablecloths and the tables. When we're putting together something as complex as that, is it really plausible to say that John himself learned each of the appropriate cognitive stimulus response habits such that all of these things that have never interacted in this way before would fall into place perfectly? John in his lifetime just hasn't had enough experience…well, I shouldn't say this about you in particular. I suppose that even some of the younger of us around the table could have put together something like this but we wouldn't have benefited from the experience that John has had through his lifetime putting together many such events.

What's going to be the source? Where would the knowledge come from that we can then flexibly assemble into novel plans and novel

procedures? A hint towards the solution to this problem comes from observing that humans as a species stand apart from all other species in our ability to engage in controlled cognition, to organize our lives around very distant goals, to inhibit more automatic responses and hold and use flexibly information in working memory. We should ask ourselves what is it that enables humans to do that? What are the other things that are unique and special about humans that might explain why we're able to engage in controlled cognition?

The one other thing that stands out to my mind as being incredibly unique about humans is culture and is social interaction. One obvious source comes back to the point that I made about how we learn to do long division in school. One place that we could learn the appropriate sets of cognitive stimulus response actions—the ways to manipulate information internal held in working memory—would be through the scaffolding of other people teaching us and other people showing us that if you want to get a Ph.D. you're going to have to publish a certain number of papers, and in order to publish the papers you're going to have to run the experiments and you're going to have to be able to run a t-test, so you'd better go to statistics class, right? We can't learn all of that stuff through trial and error. The first time I tried to get a Ph.D. I didn't go to statistics class, so let's try something new this time. Right? That knowledge comes to us through cultural channels.

In the literature right now there's a debate between two rival theories for what makes humans unique. One theory calls itself the "cognitive niche" and it basically says what makes us unique is that we can think very, very carefully and hard about things in a controlled way. Another hypothesis calls itself the "cultural niche", and it says, no, what makes us unique is that we get for free the answers to problems culturally. Other people have worked it out through trial and error and they tell us.

What I find really exciting is the idea that it's not just that both of those things are true but that they're codependent. That in principle you could not make the mathematics of controlled cognition work, you couldn't solve the computational intractability without the support of cultural input, and that cultural knowledge wouldn't be much good if you couldn't flexibly reassemble it in the way that hierarchical representations allow you to.

That's a promissory note, we haven't done the research yet. I haven't, certainly, and I certainly don't think the field has either, but that's what I see as being really exciting right now. And in a way I think if we were able to use some of these ideas to address the problem of what it is that makes humans so different than other organisms it would be

rather more exciting than just having figured out why it is that I keep eating ice cream instead of writing papers.

KAHNEMAN: How does your treatment relate to the Newell and Simon idea that basically you solve problems by working backwards from where you want to go? If I want to go there, I must get here first and ...

CUSHMAN: There's a lot of virtue to thinking about working backwards. Introspectively, all of us do this a lot. But you still have the same computational intractability problem. If where I want to be is having a Ph.D. and there isn't some trick, there isn't some ghost in the machine that's going to help direct my attention to of all the things that could directly precede my having a Ph.D., the one which is relevant, which is having written at least three papers, then I've got an enormous search problem on my hands. You see my point is you could say, "Oh, I want a Ph.D. What's going to be important? Let me start with the A's. Having aardvarks; alpanaering…" You've got an enormous search problem, you've got to constrain that space somehow. There's got to be some part of the brain that is able to direct your attention to the correct response, given the sort of cognitive state that you're in, whether you're moving forwards or backwards. Another way of thinking about it is: could we go to the programmers of Deep Blue who are trying to design a computer that plays chess very well and say, "Oh, you guys, a couple of decades of work, just go backwards. Start with checkmate, work backwards." Right? But it's not that easy.

KAHNEMAN: You would know that people who get a Ph.D. have something in common. That is something in your experience and you get that culturally. So that would very considerably narrow your search space.

You know, people who get their Ph.Ds., if you look backwards they all took prelims, they all wrote a thesis. There's actually a lot that you know from the fact that somebody got a Ph.D. that allows you to go backwards.

CUSHMAN: That's right. And there are two important conclusions to draw from that observation. One is that it's certainly not the case that having learned it culturally or in school would be the only answer to how we could narrow that search space and find an appropriate

solution, and I wouldn't suggest that it is. But the second point is that, suppose that by hook or by crook one did know that the appropriate step before getting the Ph.D. was having three papers, what brain mechanisms would one then use to codify that knowledge and allow it to be used when next planning to get a Ph.D.? The Ph.D. case is tough, one only does that once, hopefully, you know, but making a sandwich is something you're going to do everyday; you want to cache that knowledge somewhere.

Gary Marcus has this wonderful title for a book of his: *Kluge*. The basic idea is the brain has a bunch of pieces sitting around and you're just going to kluge one of the pieces you've got, and it looks like at least one of the kluged solutions to where you would cache that knowledge is in the basal ganglia through this kind of analogy between the perceptual motor linkages and conceptual cognitive linkages. Almost certainly not the only way that people cache out that type of information but one of the ways that we've begun to understand and that I find exciting.

KAHNEMAN: There is direct evidence that the basal ganglia are involved in that?

CUSHMAN: Absolutely. For instance, my colleague, Michael Frank has this very beautiful work where he takes Parkinson's patients and manipulates whether they're on or off L-dopa and then looks at the impact that that has on their ability to use working memory representations, or looks at people with different genetic variants that impact dopamine function in the basal ganglia and, again, shows systematic effects on working memory. Where I'm pushing a little bit beyond—I say this as a warning, not as a self-congratulation—the sort of state of the art in the field is in suggesting that one of the most critical functions of those working memory representations may be to solve the problem of hierarchical goal planning. He's used N-back tasks, you know, very standard measures of working memory but ones that don't necessarily involve hierarchical representation. But folks like Matt Botvinick and David Badre have been starting to take those models and say, "Ah, these look like just what we need in order to understand hierarchically embedded goals." Who knows, it may go flop, but I think it's got a lot of promise.

GRUBER: I love this stuff and I particularly love the cultural angle but it strikes me that there's some empirical stuff that we might not know yet, or maybe you know if we know yet, which is the extent to which your social inputs and other people's social rewards can feed into these reinforcement learning models, right?

CUSHMAN: Yes.

GRUBER: There's one question about whether I can watch other people getting rewards and whether I can structure from that my own set of plans—my own kind of perceptual motor kinds of schemes. Because ultimately if that social input's going to get in there, it strikes me that from your argument it has to enter in the kind of dumb rat level as opposed to the kind of goal level. Can you get reward prediction error from other people's rewards? Are people looking at this stuff yet?

CUSHMAN: Gosh, I'm embarrassed to say that I don't know right off the top of my head for reward prediction errors. The best work that I know on this topic is by Liz Phelps, and a lot of her work focuses on aversive or fear conditioning rather than (she has some stuff on reward, too), but fear has been the mainstay of her research. She has some work showing that the amygdala, which seems to play a somewhat analogous function in these sort of conditioning processes in the fear domain, responds equivalently. In one experiment you're hooked up to a shock machine and a computer shows you different shapes, whenever you see a blue square you get a shock, your amygdala starts to activate, whenever it sees the blue square even absent the shock because it's formed this predictive association.

She finds that you can show somebody a video of somebody else participating in the experiment and then if you show them a blue square on the scanner, you get the same amygdala response, even though they've never experienced it themselves, they've only seen it through the video. And she additionally finds that you can get a similar response just by telling them "Hey, you know what about blue squares? They predict shock." But the interesting thing is that that, unlike when you observe it directly, there you get a bilateral amygdala response. When you learn it verbally, you only get a left amygdala response. And of course a tantalizing possibility is that that language is left localized.

KNOBE: I really liked your example about the rat, that even when the rat doesn't want food at all it will still... According to the theory that you've been developing it should be that that same thing with the rat is exhibiting its behavior, we should be exhibiting cognitively. That even though we don't exhibit the behavior actually, when we're just trying to figure out what should be my sub-goal, we're going to show that exact error?

CUSHMAN: Yes.

KNOBE: Do people actually do that?

CUSHMAN: You've drawn me into inside baseball. But yes, we do and the best example of it is actually in the moral domain. We consider it worse to harm somebody as a means to an end than as a side effect of our behavior. For instance, the Catholic Church has a very peculiar version of this doctrine in which if you are pregnant and your fetus presents a threat to the pregnancy, then it's impermissible to terminate the pregnancy in order to save yourself because you're killing the child as a means to saving yourself. The death of the child is the sub-goal, right? It's like, what's threatening me? Child. Then the sub-goal is: kill the child in order to avoid the threat.

However, if you're pregnant and you develop uterine cancer, the only way that you can save yourself is to have a hysterectomy, but as a side effect of the hysterectomy of course the pregnancy will terminate—that's permissible. And notice the one critical difference between the two cases is that in the hysterectomy case there's no sub-goal. You didn't say, "Oh, my sub-goal, in order to achieve the goal of saving myself, has to be to kill the child."

People draw a distinction between these two things. I hope you can see the way that this connects with some of the ideas that I was describing before. If you had a system that assigned values to sub-goals, then that system when it looked at the sub-goal—kill a child—I mean, of all the things to assign a negative value to, that would be very high on the list, right? You'd get a big response out of not wanting to do that. But when it occupies the role of side effect, a system that assigns values to sub-goals would miss it.

What's interesting about that case, and, again, it comes back to the challenge of working out a kind of dual process view of the mind. Usually when we think about goal/sub-goal hierarchies we think we're in the part of the mind that's fully controlled, that has promiscuous access to all knowledge in the brain, but a system that was fully controlled that had promiscuous access to all knowledge in the brain would focus on the fact that the baby is equally dead in both cases. Right? You have to understand: why is a sub-goal representation really critically important and yet important in a blind way—important in a way that can't put together all the consequences in a kind of forward planning sense of an action that we take?

If you think that there's this sort of peculiar marriage between a relatively more dumb system that just does the stimulus response stuff and places values on actions, including sub-goals, and then the process of goal planning itself ... I feel like you might be able to have your cake and eat it, too.

MULLAINATHAN: Danny's question got me worried because I found your chess metaphor very appealing. Then the question about working

backwards made me realize it's actually ... I'm worried we're over-applying the math implicit in that metaphor. So, picture the chess tree is actually a tree that's exploding out. And when you say we want to reach checkmate ...

CUSHMAN: There are many instantiations ...

MULLAINATHAN: On the other hand, there are many mathematical problems like the puzzle example you gave, that exactly does not fit that. There is exactly one instantiation by which the puzzle was assembled. The idea that there's this many is kind of an illusion. There is this thing and of course, then working backwards makes total sense, which I think is what you were getting at with your PhD example.

CUSHMAN: Yes.

MULLAINATHAN: And is that just a superficial problem or is that a more basic distinction between these? I know in the actual programming literature these are two very different kinds of things and everything in between happens. But is that notion that there are tasks where there is only one thing you're trying to get to, so you can work backwards versus a task where so many things can lead there and the end goal is… is that distinction important? Is that something people thought about? Because the math would be quite different.

CUSHMAN: The wonderful thing about having started this discussion by saying, "I'm going to present questions that I find exciting rather than answers" is that I can now say I just don't know. I think it's a perfect question and I don't know the answer.

MULLAINATHAN: I love the long division thing and the idea that you had to understand the rote learning thing ... but yet, and maybe this is just an illusion, but it feels to me that some amounts of my almost automatic cognitive responses, it's not just that they were never learned by me as you were getting at. It's also they were never rehearsed in this way. Even if they were learned by Danny [Kahneman] who told me, it's not that he didn't tell me and tell me again and again. I didn't in my mind rehearse the thing again, and again, and again until I got it. It was almost like I got that module, sometimes from someone else, sometimes just by figuring it out and saying, "Oh, this is the right thing to do." But once I had it, it was almost like I could pop it into a system.

CUSHMAN: Yes.

MULLAINATHAN: And it felt like it pops into a system and maybe what we'll discover is it doesn't pop in in the same way as something that was learned but at least by my own intuition it feels as automatic as five minus three is two. Do you see what I'm getting at?

CUSHMAN: I do. One of the areas of research where people have really investigated that pop-in effect is what gets called fictive rewards. Read Montague is one of the leaders in this area of research. The idea is in the simplest version of making a person into a rat you just give them a bunch of levers they can pull and they get rewards. If they pull Lever A they get the reward from Lever A, and over time they learn which levers are good.

But a somewhat more complicated version that reflects the way that humans sometimes learn about things is that when you pull Lever A you get Reward B, but then it's revealed to you what reward you would have obtained had you pulled Lever B or Lever C.

Behaviorally you observe that people use that information. Quite sensibly, they should. It turns out that when you look in the brain and you try to look at neurosystems that seem to be responding to those moments of revelation, it's the very same mechanisms that learn from direct experience.

What the prefrontal cortex seems to supply is a fictive reward that the basal ganglia then treats as if it had been a veridical reward, so that the next time you have to choose one of these three levers the basal ganglia itself can evaluate their relative values. I really don't want to give the impression that that is the complete answer as to how humans behave in gambling tasks, it's certainly not. But the beautiful thing about the basal ganglia is we've learned an awful lot about how it works, so at least with respect to that system we know that there's a way that you can take a verbal representation and somehow create the type of input to the system that ordinarily a reward occupies, even though no reward was experienced.

MULLAINATHAN: It's like you had said: exorcising the ghost from the system. It feels like, to me, that particular ghost would be very interesting. Because when you start talking about conditioning, I thought, "Wow, this is really interesting, we're going back to this ... " Now we have this ghost creeping back, it that feels totally different and interesting. And I'm not saying that we haven't made progress. It feels like that understanding that ghost would be a particularly high return for understanding your findings.

CUSHMAN: Some of the most exciting stuff that's happening now working on that problem—I'm so embarrassed, I've forgotten the

researchers who are responsible for this work—but a team has used optogenetic techniques to be able to selectively active neurons within basal ganglia that respond to reward. Now they can be the ghost. That is, they can direct the response of these neurons and then observe the subsequent impact on behavior where the rats prefer Lever C because just at the right moment in time they'd used the Lever C neuron. But that's a long way from answering how one's own prefrontal cortex does it.

~~~
~~~

The first three talks this morning I think have been optimistic. We've heard about the promise of big data, we've heard about advances in emotions, and we've just heard from Fiery, who very cleverly managed to find a way to leave before I gave my remarks about how we're understanding something deep about human nature. I think there's a risk that my remarks are going to be understood as pessimistic but they're really not. My optimism is embodied in the notion that what we're doing here is important and we can do it better.

ROB KURZBAN:
P-Hacking and the Replication Crisis

The first three talks this morning have been optimistic. We've heard about the promise of big data, we've heard about advances in emotions, and we've just heard from Fiery, who very cleverly managed to find a way to leave before I gave my remarks about how we're understanding something deep about human nature. There's a risk that my remarks are going to be understood as pessimistic but they're really not. My optimism is embodied in the notion that what we're doing here is important and we can do it better.

I really wanted to take this opportunity to have a chance to speak to the people here about what's been going on in some corners of psychology, mostly in areas like social psychology and decision-making. In fact, Danny Kahneman has chimed in on this discussion, which is really what some people thought about as a crisis in certain parts of psychology, which is that insofar as replication is a hallmark of what science is about, there's not a lot of it and what there is shows that things we thought were true maybe aren't; that's really bad. This is a great setting in which to talk about these things, and I want to talk about it in part from my experience in this because I started to come into contact with this in a way that I'll describe right now.

Let me just give you a quotation from Barack Obama, President of the United States; he says, "I'm trying to pare down decisions. I don't want to make decisions about what I'm eating or wearing." He was discussing the fact that all of his suits are the same, so he doesn't have to actually pick suits each day. He says, "You need to focus your decision making energy." He's relying here on an idea from psychology which is that there's this stuff called willpower, and this connects to Fiery's previous remarks, and you just sort of use it up and if you use up your reservoir of willpower you've got less of it to make a decision. This is some work that was developed by Roy Baumeister and colleagues more than a decade ago. The paper on which it's based has been cited over 1,500 times, which, for the uninitiated in psychology,

that's a lot. The modal number of citations on papers tends to be zero, so 1,500 is bigger than that.

I remember coming across this and thinking this is very contrary to the kind of model that I would have predicted. It feels like this kind of hydraulic model, which felt to be very 19th Century, as opposed to a computational model which would've had the properties that Fiery was talking about, which is, okay, what's the demon in there that says: "Okay, how do I figure out if I should be doing my paper or eating ice cream?" and then some kind of process where these two different sorts of systems fight against each other and then one wins and then some kind of behavior pops out. I thought: if this is right, then I've got to step back about what I thought I knew about the way the mind works. It was pretty important to me in that respect. I also think it's an important issue: how do we make these decisions about exerting willpower? Lots of decisions that are really bad in the long run can be understood in this context.

I looked at the literature and one of the first things I did, (which was what many of us around this table would do) I planned an empirical agenda, the first part of which was to replicate the finding. I had a first year graduate student who was just starting and so, of course, I did again what all of us what do, I made him do all the work. He put together a replication of one of the big kinds of bedrocks of this literature. In this literature there are two different tasks. I'm looking this way and I'm not supposed to look at some words on the screen over here, and this is supposed to drain my willpower but it's hard not to look. Then I do a task that takes willpower—solving some unsolvable anagrams or doing a Stroop Test, which requires saying the color that some word is in as opposed to what the word says, which is just hard.

We did that and we ran about 100 subjects and we got nothing; we couldn't get the effect. We contacted the people who developed these stimuli. We also read the literature more carefully, because, of course we didn't read it that carefully to begin with, and we specifically went through and we tried to find where the big effect sizes were. Maybe we just picked a weak one. We ran that with their guidance, including their original stimuli—couldn't get it. We ran a third round, and by this time it was the end of the student's first year because, of course, these things take time, he had to write a paper up from his first year project to punch his ticket to get to the second year; we couldn't get that one either.

Then, as all of us do, again, I started just talking informally with colleagues about this. I would go to give talks in places and, lo and behold, it turns out there's this kind of background radiation—there's the dark matter of psychology, which is a few people who fail to

replicate and don't publish their work and also don't talk about it because the fact that you've failed to replicate has a reputational effect, right? The person who's in charge of this literature says, "Oh, these guys were going after me," and so maybe you don't talk about it in polite company. Right? It's sort of like sex, it's the thing that we're all doing, we're all replicating, we just don't want to talk about it too much, right?

Once I did that I started getting the sense that I was fishing into literature where there's no there there. There are other things that I won't discuss that lead me to that conclusion, too. I will say one piece of it, which is that more and more work is coming out that's very difficult to interpret under the willpower model. Carol Dweck has this stuff that says it only works if you have this belief that willpower is a limited resource. And that seems like it's pretty bad for the model.

Just a couple of days ago, so before I came out here, there was a piece that came out in Psych Science, obviously a prominent place where we'd all like to see our stuff. I'm just going to read the title to you: "Heightened Sensitivity to Temperature Cues in Individuals of High Anxious Attachment. " That's not important, what's important is what comes after the colon: "Real or Elusive Phenomenon?" It was a failure to replicate, so basically there was this effect, which does not matter for this conversation, they tried to replicate it, they quadrupled the sample size, they can't get it, and then they published this. What they published is that it's either real or elusive, but you can't even say out loud real or it wasn't there to begin with, right? These are not things that we say in polite company.

We have this responsibility to be better scholars, particularly as scholars whose work is consumed by the world.

What I want to talk about today or what I've already talked about and what I want your ideas on is how do we get rid of bad ideas? Because my experience is a little bit like the pessimism that you see in Max Planck, which is that it happens a funeral at a time. I'm confident we can do better than that; and people discuss this. As an addendum on the willpower stuff, which is really in terms of content where my interest is, one of the things that was striking to me is, I went back and what these authors are saying is that the substrate of willpower is brain glucose. If you run out of sugar in your head, then you can't exert your self-control. And I did two things. The first thing I did after I read that is I just did a back of the envelope computation about how much sugar we're talking about. The theory's off by two orders of

magnitude. They're giving people 100 calories of lemonade in these manipulations and, in fact, your whole brain is using like a quarter of a calorie per minute so you just use this massive sledgehammer to talk about this little bitty effect.

The other thing I did is I reanalyzed some of the data in the paper on which many of these claims are based. First of all, I should also say the key dataset I asked for, I was told that the data were corrupted (this is a paper that was two-years old). Already there's something funny going on in our discipline when we can't get the raw data from scholars. But then I looked at the data they did give me and rather than supporting their claims, it undermined it. If you just run the math, which wasn't too hard, it was five minus three (that joke is lost without Fiery), it turns out that their own data undermined this idea. I want to emphasize again, this is work that's continuing to be exciting. I just did a Google search on it before I got here. I just searched "organize my time." Within the last month another dozen papers have come out on this. And it's going to turn out that this is not true, it's not true in an important way, in the sense that, okay, Obama's not making policy today, he's not deciding on whether or not he's going to bomb Libya based on how much glucose he's got in his head or something like that. But he's aware of it. This is penetrating into policy sectors.

My point is not specifically narrowed on the willpower stuff, although it's important we clean this up, this also has public health implications. There's this great advertisement we found in 1971, this is work in collaboration with my good friend and colleague, Angela Duckworth and Joe Cable, a behavioral economist, a neuro economist. We found an advertisement which basically says: If you want to lose weight, what you do is eat an ice cream cone because that's got the sugar, it'll give you the willpower not to keep eating. This is something that people in my casual conversations, if you are a consumer of psychology as opposed to a producer of psychology, this is the kind of stuff which is penetrated into the popular consciousness. If I just eat a couple of simple carbohydrates, then I'll lose weight. Not going to turn out to be right.

We have this responsibility to be better scholars, particularly as scholars whose work is consumed by the world. Uri Simonsohn, my colleague at the University of Pennsylvania in the Wharton School, has written extremely well on this. He's discussed all the things, again, like sex that we all do but don't talk about. He calls it P-hacking, where we run enough subjects and the P value sneaks below .05, and then we've decided we're done. That's just one of a large number of things that people do, including selective reporting of dependent measures. I'm not saying people have not addressed this. I understand that people have. I'm saying that in many ways the replication crisis in psychology

is a little bit like the weather, right? We all talk about it but no one really does anything about it. We do a little about it here and there. I'm going to kind of wrap up. I'm in the speaker slot between the audience and their food, so I'll keep this brief. But I really do feel like this is a good opportunity for us to talk about it. It's really important.

I want to talk about, again, getting back to me, a couple of things that I've done. I have a journal,*Evolution and Human Behavior*. As of January of this year, mostly influenced by things like this and Uri Simonsohn, we implemented a policy that says you have to post your raw data, all your raw data. Now, there are a couple of exceptions. If there's identifying stuff in there that violates some kind of HIPAA requirement, privacy requirements, also if your data are drawn from a publicly available source, it's already available. And we've had good compliance. At my journal I don't want people to run into the problem where they publish this result and then people who are trying to recreate the statistics can't do it. That's the first thing we've done.

The second thing we've done is we've got a professional statistician on board. All of our action editors know that if there's anything about the stats that feels dicey to them, they can ask this person to take a look at just the stats. This is a person who doesn't have to worry about the content, the theory, what have you, and is more or less indifferent; he's not even in my discipline; he doesn't really care. He's actually a kind of critic of the discipline, which is one of the reasons I feel lucky to have him. And so just the stats can be looked at by a professional. This is a nice innovation which I would love to see elsewhere.

Another thing we're talking about is, (this seems like a separate issue but I don't think it is) I bet all of you have looked at people that cite your work and you look at the sentence, like, "I didn't say that. How did the person take me to be saying that?" Even the reverse in some cases. And, having published in law reviews, one of the things I thought was so interesting about their model, and I see why it's difficult to implement but, on the other hand, I don't think it's impossible, they're law students, they go through every single footnote and make sure the source says what you say it says. Now, they err in the other direction. I got this funny comment from a law student who says, "What's your citation for the claim that correlation coefficients go between positive one and negative one?" Fine, they've gone a little bit too far, right? But, nonetheless, if we think that science is accretive to the extent that people are mis-citing these sources and they don't go back and the thing doesn't say what it says, first of all, that means the author was probably confused or lazy and we should clean that up, but the second thing is we're not really building an edifice anymore, we're doing kind of this weird smorgasbord kind of thing.

I should just say, *Evolution and Human Behavior* is published by Elsevier. The downside of this is they're big and evil, right? Everyone hates Elsevier because they're making a ton of money off of our labor. I take that point. One of the good points is we get the royalties, right? We get a fraction of that. We have the resources to have graduate students go in and check these citations. As far as I know, in psychology and economics, This is true in philosophy, not only doesn't this happen, it wouldn't really occur to people, right? The onus is on the author. I don't see any reason why these institutions need to stay in place.

My point here is not yet another handwringing, how horrible is this? It's that we need concrete proposals to move the field forward in a way that's going to be productive. I'm optimistic about psychology. I think that in ten years it could be that in addition to a council of economic advisors there's a council of psychological advisors. You're already seeing this a little bit. As people who talk about nudges are getting into policy positions. But what that means is our powder better be dry, like, we better be giving ideas to policymakers that are right. And the way to do that is to do better science.

To take it back to where I started, I'll be honest, the idea that there is a reservoir of willpower, that's just obviously wrong. If you look at it in the context of what we already know about the way the mind works, it wouldn't have passed the 10-second test, but it's a very appealing idea; it has a certain resonance. And I'm willing to say that as an empirical matter, when people start going into this to try to replicate it and they bring together all the replications with the existing data, they're going to find there was no there there. Now, maybe I'm wrong. Maybe that turns out not to be true. We can talk about that if you want. But I do think that it's going to turn out that there's lots of ideas that people like us—not us, obviously we're too careful for that—and, in fact, Uri Simonsohn uses Danny Kahneman as an example when he talks about: what should your curve look like in terms of P values? All of them shouldn't be between .045 and 05. It should actually look like a lot of them are really small and then a few of them are a little bit bigger and Danny's curve looks just like that, which means that it's being drawn from the right distribution, that is, he's looking at actual effects and doing appropriate statistics on them. But that many of our colleagues are not, and I that this is the kind of community that has a responsibility to be more assertive about making positive changes so that we do a better job of doing our jobs.

GRUBER: This is a wonderful initiative and I'm happy to see someone not, like you're saying, wallowing in the despair that we see ourselves in, but practically thinking of steps we can take. A couple of questions I had related to this is: how do you think we can expand this kind of

approach towards datasets that are much more complicated to grapple with? For example, there's issues that often arise in behavioral coding where people worry there are biases within the lab to be coding a behavior in a way that sort of supports their hypothesis. Do you have people submit the videotapes? This is just an issue to think about how we expand it. Or datasets that are far more complex and require a lot of preprocessing—I'm thinking neuroimaging, even psychophysiology to this extent. What do we do to really insure the integrity of the data we get in those raw data files? SPSS, Excel—how can we really trace it back to some of the biases that really could be problematic earlier on upstream?

KURZBAN: Let me say two things about that. The first thing I would say is that the solutions for cleaning things up are going to turn out to be specific to the area. And that's something that's a challenge. I take that point. And so the short answer to your question is, I don't know. But let me say something that is really important, which is journal incentives. If I did the kind of thing for my authors where I said to them, "If you're going to give me a behavioral coding experiment I need you to send the data to another lab and have them verify it before we're going to look at it," I would stop getting submissions, right? I'm not saying this is an easy problem. As an editor, to the extent I make my authors' lives harder, which is going to improve the science, I'm going to look worse because down the road I'm going to get fewer papers and they're not going to be cited as often. There's no doubt that my incentives have something to do with keeping things easy for my authors, which is going to push down the quality of the work. That's part of what we've got to start talking about. I mean, already there are some suggestions along these lines.

BROCKMAN: What about issues of code? You talked about the data you get from your experiments, but a lot of researchers have proprietary code—some of it corporately financed where there are patent issues. Unless you have the code or the software, you don't have anything. And yet often that never comes up. Victoria Stodden has done a lot of research on this. And how do you handle that?

KURZBAN: Again, my answer to that is ... I'm going to go ahead and punt and say that's a great question for the smart people at the table to figure out. I will say this. In my word, code is usually agent-based simulations. And Rob Boyd has introduced what I think ... it's actually crazy that we didn't do this before, which is that he has two different people simultaneously do the agent based simulation code to make sure that when they look at the output they're not looking at something idiosyncratic because someone made a mistake. They want to make sure they have two replicants of the building of the code. This, is a no-brainer. Yes, it adds work and, again, there's an incentive problem here, which is now he's going to get half as much stuff done

because half the time he's doing his buddies' coding. That's the kind of innovation that has been very good. And in terms of what you're talking about, I have to throw up my hands. I don't think these things are insoluble. There might be legal mechanisms that allow people to verify these things. People don't talk about this, they talk about how horrible it is that Elsevier takes our labor, but it makes pools of money available. One of the ways that we can deploy resources might be to ... there's various ways you can incentivize people to behave in a better way or at least to check and verify.

KAHNEMAN: A couple of remarks. One, I don't want to defend the reservoir model of willpower but there is a difference between the way that people operate when they're tired and when they are not tired. My understanding of Baumeister's research is that he's studying fatigue and that looks more plausible than the glucose work.

On the other point of practices, I have a suggestion—an observation and a suggestion. There is a line of research which includes Baumeister's and the priming stuff, where it takes an hour to collect one data point.

Those are between subject experiments and they're expensive. They're very costly, and the samples that people run are too small, they're too small —I think a reasonable idea is by about a factor of four—and they haven't increased in the last 50 years. There were complaints about that. Jacob Cohen in the 1960s, and it's been replicated by Gigerenzer 15 to 20 years ago, and nothing has changed. I think that's a matter for editors.

Bobby Spellman and I have been thinking of writing a piece where editors of psychology journals would treat a problem as car makers do with respect to fleet gas mileage, that is, that within ten years we have the goal now of achieving that level of power. And that is agreed across all journals so that you don't get competition with the wrong incentives. And it's slow. But if you have that as an objective, that you can measure the average power of your studies against small effects in a reliable and consistent way across time and show improvement. That would reduce the problem very significantly.

KURZBAN: Let me address the first one very quickly. I'm not denying the empirical phenomenon. There's definitely something going on. I'm arguing with the explanation.

Bobby's done great things with perspective and exactly what you're talking about is the kind of thing that I want to point to, which is we've got to get groups of editors in fields together either physically or in terms of correspondence, or in terms of policy. That kind of initiative is

going to wind up being successful. And this speaks to the question of incentives. Because if I'm no longer competing with editors who are willing to go low in terms of power, now I'm able to insist on well-powered studies without putting myself at a disadvantage. This is like hockey helmets. I don't mind if there's a rule that says everyone's got to wear a hockey helmet, but if there's no rule, I don't want to wear one because I'm going to lose that advantage, right?

That, is where we've got to go and in order to do that you've got to have leaders who step forward and say, I'm going to gather these editors and I'm going to say, "You guys, you can do better as long as you're not mutually affecting others in a negative way." That's exactly the kind of thing I'm pointing to.

MULLAINATHAN: It's not helpful in the sense that we're not proposing any magical solution or even a non-magical solution. I'm talking a little bit what the problem is—that is, it's easy to sit and say the problem is replication. That's part of it, but I actually think there are two problems that are actually unrelated to replication or at least weakly related to replications that contribute to the problem, and this is coming from looking—economics has this issue as well—So let me start with the first one. You mentioned that the sub title of it was "Real or Elusive." There's something to that and here's what I mean. It seems to me a very crude way of describing psychology is that you have quite a bit of control in the lab over the nature of the context you're used to to show the effect.

What I mean by that is even taking the original willpower studies. You had said you tried to replicate, tried to follow their exact methods, and supposedly it turned out that it wasn't the hand press, that you had just used some other thing and you didn't find it. You went and did hand press and you found it. Oh, good, they were right. But were they right? And, after all, if I replace hand press with this other thing, but in some narrow sense, and that's what I mean by replication. There's a question of if you literally replicated the original studies and found they worked perfectly—you replaced hand press with some other proxy for self-control and found it no longer worked—that's a different kind of failure to replicate. It's almost like a context sensitivity. And the reason I'm emphasizing that is, I feel like there's quite a bit of that that goes on, which has nothing to do with p-hacking generally, it has to do with "Oh, I would like to show this effect. Let me try it with these types of things. Let me try it with this." And to the extent that that's in the DNA of the field to say, "I'm looking for context when certain things happen," then I feel like this type of problem is going to be immense.

Let's contrast it with areas where we're trying to find psychological effects but where everything is already pre-specified—line up versus

show up, or how do we show data—so then it's really clear. Are you picking the right guy? Are you doing this? We're looking for effects. There's still a lot less freedom and replication is much more robust than sensible. Does that make sense? And seems to me this is as much a problem of what the goal is as it is about the method. I don't know if I'm being clear. But it's that you can really see that.

Take biology, where we're all trying to study this cell and this mechanism, then it's very clear when there's a failure. The fields where I've seen this happen, the biggest is the subfields within psychology is a terrific example. It's a very abstract big concept, and if I happen to show the effect using hand press, I've shown it. I'm only pointing this out because you talked about the council of psychological advisors. I would be exactly as concerned if you told me all of these willpower studies perfectly replicated but they only hold for hand press. Well, what on earth does that have to say about policy? There's a different kind of replication that fields like psychology need to take far more seriously, which has to do with robustness across settings. And there are areas where it's taken very seriously, which has to do with robustness across the thing. And there are areas where it's taken very seriously and there are areas where it feels like it's just left to something else.

KAHNEMAN: Sendhil, that's one of the major problems. I feel conceptual replication is what social psychologists do because they go across contexts. And that is one of the major forms of p-hacking. That is, that there are no constraints on how many failures can replicate. Hal Pashler asks, "when has anybody reported a failure of a conceptual replication?" So, p-hacking is mediated in large part by conceptual replication.

KURZBAN: Yes, I completely agree with that. It's a different form of p-hacking. But another way to think about this, again, having one foot in economics and one foot in psychology, it's just a different. The entities in our explanations in psychology tend to be way wigglier than one finds in at least some areas of economics, and it leads to these kinds of problems. I's the nature of the explanations and then some of these methodological practices, which are permanent. So I don't think I disagree.

MULLAINATHAN: The reason I'm pushing is, I understand you use this p-hacking, but it's actually statistically pretty significant.

KAHNEMAN: You're right. It's being misused in a big way. This is right at the core of the debate between social psychology and its critics. Social psychologists say, "We replicate conceptually all the time and, therefore, a failure of literal replication doesn't concern us." That's a big part of the debate.

KURZBAN: You said you had two points.

MULLAINATHAN: The other point is this ... look, to be honest, economics is as prone to this as psychology, and it's terrible. I looked at a lot of medicine and how the ideas in medicine have evolved and here's a striking fact. There are a lot of weird theories that just look crazy. The germ theory is crazy. You're telling me that there are these invisible things called germs that are deadly to us ... oooh. And yet at the same time, here is the anomaly: there are 100 crazy sounding theories and one of them turns out to be right.

The reason I'm emphasizing that is that in contrast, when you work in these fields that have these problems, we're so driven by our intuition about what's right. Like, "That sounds wrong. That just feels like it can't be true." I don't know if you've noticed this, but there's a fundamental dominance given to the theory and the intuition as opposed to: it's a mess, it's too bad that the data suggests this, and who knows what it's going to be saying? Even the structure of a psych paper or an economics paper is of the variety, it's like you're forcing your hand into saying, "Write down the theory. Here's a test." Yet, you look at biology—a vibrant field, it's a mess. And it's that mess from which stuff arises. And I feel like there's something else, forgetting replication, forgetting everything, why is there no space for empirical exercises where you say, "I don't know what this is. I hope in 50 years we figure it out." I mean, that would be the worst. You would say, "What an unintellectual researcher. They just aren't thinking." That's the other fundamental problem, is that we're not allowing empirics to have its own vibrancy and strength and structure.

PIZARRO: We were just talking about this, and that you're confusing the way that we write up our ... psychology is a mess and it is driven a lot by this sort of empiricism that you're describing. We do lots of studies, and we discover an effect that we never would've predicted, and then we write up the paper by saying, "We sought three tests of the hypothesis..." Maybe the error here is in the way that we communicate what we do, but that's never how we do it.

MULLAINATHAN: That is what I mean. Because that that is not just narrow communication. All that stuff that you did to arrive at the thing, that's what should be on the other side of the thing.

PIZARRO: Absolutely. I just wanted to point out that it is a mess.

KNOBE: A lot of our discussion has been about ways that we can avoid making errors in the first place. But it seems like a really central question is if we have made an error, so we said something and it's wrong, how can we make it the case that we will recover from that

error? And it seems like if we work backwards in the way that Sendhil was suggesting earlier, what we would want to have is the person who initially made the error would change his or her mind—would say, "Oh, now that there's new data on the table, I guess I was wrong." But then if you work backwards from that and we think: what can we do to make people actually change their minds? It seems like a really central thing is something cultural. I feel like right now there's a kind of spirit in our discipline where if someone says something and then other researchers investigate it further and it turns out that they're wrong, that they've been somehow de-bested, they've failed, they're being crushed by this other person who was smarter.

If we as a culture could eliminate that kind of feeling about what happens when you turn out to be wrong, then people would be much more willing to say that they're wrong. Maybe if we all had a different feeling about what happens when people just say, "Oh, I guess it wasn't true," people would be much more likely to do that because we recover from errors. I was thinking maybe—like we have the Nobel Prize—we could also have the "Noble Prize." It could go to that person each year who most publicly admitted that someone else had done an experiment which just … …

KURZBAN: That's a great idea.

KNOBE: Then the field can just move on. That idea—it seemed like a good idea at the time—it's over.

BROCKMAN: What's the name of that prize?

KNOBE: The "Noble Prize."

KURZBAN: Yes, that's right. Scientists are supposed to be the sorts of people who just admit they're wrong. I always remember that nice Arrow quote where someone asked him why he seems to change his mind in print so often. I don't know if this is true or not but he says, "Yes, that's what I do when I realize I'm wrong. What do you do?" And that's good. The rhetorical question is there. I don't want to name names, but in this literature I actually asked someone in this literature, I said, "Well, there's this stuff out there. What pattern of data would cause you to change your mind?" And the reply I got was, "You know what? I should really think about that sometime."

~~~
~~~

If you think about it, humans are extremely unusual as a species in that we form long-term, non-reproductive unions to other members of our species, namely, we have friends. Why do we do this? Why do we have friends? It's not hard to construct an argument as to why we would have sex with other people but it's rather more difficult to construct an argument as to why we would befriend other people. Yet we and very few other species do this thing. So I'd like to problematize that, I'd like to problematize friendship first.

NICHOLAS CHRISTAKIS:
The Science of Social Connections

The part of human nature that I'd like to talk about today is that part of our human nature that is relevant to our interactions with others. There's been a phenomenal amount of work taking place in the last ten years, certainly, and even in the last year or two that seeks to understand how we interact with each other and how we assemble ourselves into social networks.

If you think about it, humans are extremely unusual as a species in that we form long-term, non-reproductive unions to other members of our species; namely, we have friends. Why do we do this? Why do we have friends? It's not hard to construct an argument as to why we would have sex with other people, but it's rather more difficult to construct an argument as to why we would befriend other people. Yet we, and very few other species, do this thing. So I'd like to problematize that; I'd like to problematize friendship first.

Second, not only do we have friends but we prefer the company of people we resemble. There's an enormous literature on in-group bias and on why this might be the case. A lot of this literature, to my eye, takes the form of what I would regard to be a tautological explanation. Why do we prefer the company of people we resemble? Because we're more comfortable when we are with people we resemble. Why are we more comfortable when we're hanging out with people we resemble? Because they resemble us. And I'd actually like to try to find a deeper explanation for why we befriend other individuals; why we assemble ourselves into networks with what turn out to be very fundamental, reproducible topologies (structures); and why we prefer the company of people we resemble.

And, in fact, the ubiquity and necessity of social interactions carries with it a suite of other phenomena, like cooperation, which is very deeply and fundamentally important; sensing (the ability to see what's happening in others); communication; social learning; epidemics; violence – all of these phenomena arise not so much within individuals,

but rather at the interstices between individuals. They're not so much nodal phenomena—having to do with the nodes on the networks—but edge phenomena—phenomena that have to do with the connections between the individuals.

In fact, I'd like to think that the focus on networks calls into question some very old ideas about human nature, and about what the state of nature really is for human beings. Joe Henrich, in an interview he did for Edge a couple of years ago, had a very nice, pithy summary of this. He asks why do we see market economies as all about competition for advantage? Actually, you can just rethink the existence of market economies as all about cooperation. Why do we have to see them as being competitive rather than as cooperative enterprises?

We can shift our perspective on lots of things when we think about people as being nodes on a graph, as being connected to other people. And this shift in focus might, in fact, prompt us to begin to think about —not the individuals themselves-but the ties between them. This calls to mind an analogy, which I don't know if some of you may already know, of streets in the United States and in European countries. So, streets have names in our country, and the houses on the streets are numbered numerically and linearly as you move along the street. And the blocks between the streets don't have names or numbers and are seen as the things that are between the streets, and we don't pay much attention to them. But if you go to Japan, it's the blocks that are numbered. The blocks have names and the houses on the blocks are numbered in the order in which they were built, not numerically or linearly in any kind of systematic way. If you ask the Japanese, "What's going on with the streets?" they say, "The streets are the spaces between the blocks." They don't pay attention to those.

We can even begin to think about human beings in this fashion. We're so interested in understanding human beings that we lose sight of the connections between them. And just like we can efface the individual, to some extent—and I don't have a strong argument that we should do this, but I have what I would regard to be a weak argument why it's beneficial or useful as a heuristic to do this—just like we can begin to efface individuals by thinking about the selfish genes within them, we can also begin to efface individuals by thinking about the connections outside them.

So, the question I'm asking myself lately is: What would a social science of connections, rather than a social science of individuals, look like? What would it mean to take connections as the focus of inquiry and to think about the individuals as the spaces between the connections who are not so important? And then we begin to think about all the dyadic interactions between individuals, which are themselves natural phenomena, just like we are. I'm an object of the

natural world, but so are my connections between me and all the other people, so are those connections objects of the natural world which warrant an explanation and a kind of deep and profound—in my judgment—study.

In fact, this would have a variety of conceptual and methodological problems. And some people would say that this is a really horrible perspective, that it obliterates our individuality, that it's dehumanizing, and so forth. But I would retort to that by saying, What makes us think that the ties between us are any less important or worthy of attention than the individuals themselves?

One of the things that we've been doing is asking ourselves what is the reason that we form these ties? What's the function of these edges and these connections between us? One of the things that's very interesting to us is that these edges between individuals, these networks that we form, have properties that are not reducible to the individuals. They offer us a kind of an understanding of emergence, a new kind of emergent phenomena. And these properties, while they are properties of groups, actually, as it turns out, have implications for individuals.

We're so interested in understanding human beings that we lose sight of the connections between them.

Let me give you an example. This is very visual and, given this format, I'm not supposed to use visuals, but I'm going to cheat and use one slide in a moment. Let's say you had 1,000 people, and, on average, they each have five connections, so you have 5,000 ties between them. Mathematically, you could construct a number of ways in which you could organize these networks. You could have a random network where people are jumbled together; you could have a big ring network; you could have a kind of “scale-free” network; you could have the kind of network that we humans actually make (which has a variety of properties). It turns out that if you were designing the network from mathematical principles so that the network would be the most resistant to pathogens taking root within it; so, you say, "I want to organize these people in such a fashion that this group, when so organized, resists epidemics;" whereas, if they'd been organized some other way, these same people who otherwise were identical—had the same immune systems, the same biology—this group no longer resisted epidemics so well. If you wanted to give the group the epidemic resistance property, the way you would organize the people is to give them a property in network science known as degree

assortativity. You would make popular people befriend popular people and unpopular people befriend unpopular people. You could give them this property, it would make the network as a whole resistant to germs being able to make inroads.

And I can cultivate this intuition by asking you to think about the airport network in this country. The airport network is degree disassortative. Chicago is connected to lots of small airports but, in the small airports, you can't fly from one to the other; they are disconnected from each other. Whereas people don't have that property. Popular people befriend popular people, and unpopular befriend unpopular. Now, think about which of those two networks, if you were a bioterrorist and you wanted to seed a germ in, which network would the germ spread more rapidly? In the airport network, right? If you start any random node, like an isolated small town, it will go to Chicago, and, in the next hop, it will reach the whole nation. But if you had the hubs and the spokes or the peripheral airports connected to each other, it would be relatively more impervious to a pathogen spreading.

I don't think it's a coincidence that of all the kinds of ways human beings could organize themselves into networks, that's what we do. We evince degree assortativity, and I don't think it's a coincidence that we do that. We assemble ourselves into groups, the group now has this property, this germ- resistance property, which is a property of the group, but which, as it turns out, also benefits and affects us. Now, being a member of that group, we are less likely to acquire pathogens.

And this sets the stage for a set of ideas that we and others have been exploring that shed light on multi-level selection and other kinds of contentious ideas in the biological and the social sciences. And we have a number of fellow travelers on this road—László Barabási, Dirk Helbing, Tooby and Cosmides, Frans de Waal, Nowak, Rand, Santos—people working on these related areas of interactions among animals and people, and what this means. In fact, David Rand and Josh Green and Martin Nowak just had a nice paper this past year — I was asked to highlight some papers—looking at whether you can use time to response as a kind of heuristic for understanding are people intuitive cooperators and rationally selfish, or do they exercise rational self-control over a kind of instinctive greed? The data they presented in that paper, to my eyes, was quite compelling—that we are intuitively wired to cooperate.

James and I published a paper last year as well, also in *Nature*, where we had the following idea: We said, well, what we would love to do is, if the claim is that there's something deep and fundamental about human social networks and the structure of networks, we would love to be able to go back 10,000 years to the Pleistocene and look at what

kind of networks did humans assemble themselves into, before we invented agriculture, and cities, and communication, and so forth?

We did the next best thing to that, which is to map the social networks of the Hadza hunter-gatherers. There's only about 1,000 of them left; only about 500 of them still live in the traditional way. They are a natural fertility population; they have no material possessions to speak of; they sleep under the stars. And when we map their social networks, their networks look just like ours. So, despite all of modern technology, telecommunication, the Internet and everything else, the structural features of their networks are indistinguishable from the structural features of our networks, suggesting to my eye, again, that there is something very fundamental, not just about the structure of our bodies and our minds, but also about the structure of our societies.

This is some of the work that's been going on in a number of fronts the last few years, trying to understand the social interactions, social networks, and the kind of constituent elements of that – cooperation and the like. But then that leads to what I like to call the so-what question. So what if we can understand the structure and function of networks? What can we do with this knowledge, not necessarily to make the world better, but actually to intervene in the world in some way? And if you think about it, that's also one of the tests of science. I mean, as a scientist, can you actually understand the natural world well enough that you can actually seize control of the natural world in some way and make it obey certain fundamental rules?

I'm going to close with some summaries of a few experiments that have taken place over the last couple of years, and then a bigger idea as the final point. Let me just summarize a few pieces of work that are going on in my field that are very cool at the moment. There are two broad categories of work: One category of work is, can we manipulate the structure, the topology, of the network? Can we take control of the nature of the ties between people and drive the network to desired states? The second is, can we manipulate, not the connection, but the contagion within the network? Given the structure of the network, how can we seed the network? How can we introduce information strategically within locations that make the group behave in desirable ways that we specify? Can we show that we've mastered and understood this world well enough that we can actually intervene in it?

One experiment that was done by a former postdoc of mine [Damon Centola], that was published a couple of years ago now, is this. And I have to show you this image. So, this is an image of experimentally constructed networks. There are two networks in this image. There's just no way you could describe these two networks. Both of these networks have 128 people in them, and in both of these networks each

person is connected to exactly six other people? So, if you talk to the human beings in these networks, and ask them, "How many friends do you have? " and they say, "I have six friends. " And every one of them in both of these worlds would say, "I have six friends." They cannot tell the difference between the two worlds which they inhabit?

Now, suppose I'm going to infect the person—the yellow dot that's up here—with a germ. In which of these two worlds do you think the germ would spread more rapidly and more completely throughout the network? From the point of view of the individual, there's no way of telling what world they're in, but from the point of view of us, with this God's eye view, we should have an intuition in which of these two worlds is the germ more likely to spread? And the answer is the network on the left. This random assembly means that, ping, ping, ping, in the next step, the germ will spread from the yellow dot to the six red dots, and then from there to the others, and you'll flush through the system, you'll get a blooming of the information spreading or the germ spreading or whatever. And these are things that spread by so-called simple contagion.

Now, I'm going to ask you something different and more difficult. Imagine now what is to spread within the network is not germs or information but, a behavior, for example, smoking cessation or cooperation. Something more complex. It turns out that the world on the right is the world that is more conducive to the spread of such phenomena. So the topology of the network, which can be seen from above, is what's relevant to whether or not these group-level properties can emerge and be sustained. So this was an experiment that was done to show that.

We did an experiment in our lab where we recruited over 2,000 people online, and we brought them into these virtual worlds, and the subjects played a public goods game with people near them, a kind of cooperative game with those around them who they were randomly assigned. Then, we controlled in that world whether or not people could rewire their networks and the amount that they could rewire them, by which we meant not only can you, if you defect from me, can I reciprocate by defecting, or, if you cooperate, I can reciprocate by cooperating, but we gave me another tool, which is that I could cut the ties or form ties to people. So I could form ties to cooperators and cut ties to defectors. And then we manipulated the viscosity with which that could be done.

What we found was that actually we could control the amount of cooperation that emerged in this group of people by specifying the rules of interaction. If we allowed people to rewire their ties just the right amount, then cooperation in the group would appear above and beyond and independent of the individuals themselves and their own

tendencies. So we can elicit from the group a property, namely, cooperation, by controlling the nature of interactions. Second experiment.

A number of other experiments have been done with contagion phenomena. So, given a structure of human interactions in an African village, in a trading floor on Wall Street, in schools in the United States, whatever the setting is, can you strategically introduce information in such a fashion that you can get people to behave in particular ways? There was just a paper published by Matt Jackson and Esther Duflo a couple of weeks ago in *Science* looking at microfinance. So if you want to get the adoption of microfinance in a setting in Indian villages, who do you target so that if you get them to use the microfinance you get the most spillover and the most rapid diffusion of innovation?

If we could find ways of identifying central people using big data or other techniques, and monitor them passively or actively, when we observe a spike in central people it means that an epidemic's about to strike the population.

Another nice paper that was done by my colleague, James Fowler – and all of the work that I'm describing to you, virtually all of it, has been done jointly with James—is the following: James did a beautiful paper as well last year in *Nature* where they randomly assigned 61 million people online to a voting intervention and were able to show that actually showing people a very seemingly trivial stimulus drove, not only the individuals themselves to be more likely to vote, but their friends to be more likely to vote, and their friends' friends to be more likely to vote. So he showed a spread of civic-mindedness to two degrees of separation within this massive experiment done with 61 million people. In fact, it's estimated that an extra 300,000 people turned out to vote on that election because of James's experiment. Actually our democracy was improved because of the scientists actually doing their work in that particular occasion.

There's been some other nice work on product adoption using experiments online: how can we get people to adopt products? And we're in the field right now doing some experiments where we've mapped the networks of 32 highland villages in Honduras, and we're trying to see, if we can only reach 5 percent of the people, which 5 percent should we reach so that we get the whole village to change its mind about clean water and nutrition outcomes? And we're randomly assigning the villages to different targeting algorithms. In some

villages, we pick 5 percent of the people at random; in other villages we pick them according to one targeting algorithm; and still another according to another targeting algorithm, and we have very promising results from this study.

There's also a sense in which you can now use networks—and there's been some nice work done in the last year or so, summarizing my field—wherein now, instead of introducing information into the system, you can think about networks as kinds of sensors—extracting information from the system. So, for example, if you think about it, just a moment ago, we cultivated the intuition that if you target information to particular individuals, they're going to be more able to spread whatever it is that's happening in the network.

Let me ask you to think about this, since I can't use slides. Imagine a network. There are ties and there are little nodes between them. Most of you probably have an image that, in the middle, there's a kind of jumble, like Christmas tree lights. When you open them up after a year, there's a thick knot in the middle, and there are these little tendrils that spread out to edges, that's what a network kind of looks like. Imagine that I can ask you, "You can be a person in the middle of that, and have four friends; or you could be a person on the edge of that and have four friends. Now a deadly germ is spreading through the network. Who would you rather be? The person in the middle or the person on the edge?" The person on the edge. You have the intuition that the person in the middle is going to be on more paths through the system – and you can formalize this mathematically – and is going to be more likely to get whatever's spreading through the system. This very simple idea was an idea that we exploited by recognizing that if we could identify who were the central people in networks, and passively monitor them, we would have an early warning system for epidemics. So, the epidemic curve is a classic S-shaped curve that goes up like this. That S-shaped curve should be shifted to the left in central individuals compared to random individuals within the system.

So, if we could find ways of identifying central people—using big data or other techniques—and monitor them passively or actively, when we observe a spike in central people, it means that an epidemic is about to strike the population. This can also be done with economic information or any kind of information that spreads through the system. We were able to show that this works with an outbreak of H1N1 flu a couple of years ago now, and in the last year we also showed that it works on Twitter. James and I know nine days before anyone else what's going to be popular on Twitter, because we see it spiking in the individuals that are at particular topological locations within the network.

To sum up, this is new work that has been taking place over the last year or two in my field, which is network studies and the study of social psychology relevant to interactions and in sociology (not all of sociology or all of psychology, just my little niche where I sit), and the biology of these types of things, has a number of features. First, this work is increasingly experimental in nature; so, more and more people are doing experiments. This move to experimentation is a kind of rediscovery of a tradition of experimentation in the social sciences. We always did experiments, but beginning in the 1950s, we became besotted with regression models. Psychology is a bit of an exception because they consistently have done experiments. But we're moving back to field experiments in broader swaths of the social sciences, and this is being abetted in part by the development of the Internet and online experimentation. So the big data revolution intersects with the experimental revolution by making it easier for us to do experiments.

This new work reflects four things: First, it's experimental. Second, it's exploiting online and Internet technology. Third, there is (to my eye at least) an increasing desire to try to find things that are deep and fundamental about our humanity. The best social science now that is being done seeks to go to a deeper, more fundamental level to try to explain human behavior, at least when it comes to human interactions. And, fourth, this work is involving interventions. If you want to construct an almost Popperian sort of theory of science, the ability to actually...: we observe the system; we have a hypothesis about the system; we do experiments about the system and conclude things; and now we actually manipulate the system (we introduce genes, we excise the genes, we do experiments in particular ways): this shows a level of control or understanding that's very commendable.

Collective behavior has always captivated people's interest, but, in the last couple of years, we've been making phenomenal progress in understanding what I would regard to be – for me at least – the key aspect of our human nature, which is our interactions with others.

BROCKMAN: You mentioned point three. Could you repeat that?

CHRISTAKIS: Something that transcends individuals. Something that's very deep and fundamental, but that transcends individuals.

BROCKMAN: I think you mentioned humanism?

CHRISTAKIS: I didn't mention humanism. No, but you're asking ...

BROCKMAN: You were imputing some kind of goodness?

CHRISTAKIS: No, I'm avoiding that because those were my marching orders from you, John.

BROCKMAN: There is a sense in all the discourse about networks and big data that it means good. But as Steven Pinker pointed out, the Internet hasn't changed much in terms of human nature.

CHRISTAKIS: Yes, I think I know what you're talking about. Any technology—atomic power, guns—can be deployed for good or for evil. So, I've been highlighting or imagining some ways in which a better understanding of social interactions can be exploited for good. But it can also clearly be exploited for bad. Now, this bad could be getting people to buy products they don't need; it could be whipping up political fanaticism. Actually, if you understand networks, you can be much more effective at fostering Nazism. Actually, there's a way in which you can think of extreme political ideology and how it takes root in populations, and how you would go about structuring populations precisely to reinforce these kind of extreme ideologies. So there are all kinds of bad things that you can use the same technology for, and I am not unmindful of that. But I mean, the things that we're trying to do I would think, we're trying to increase cooperation, and make people healthier, and increase economic development in the developing world, and everything else that Sendhil and everyone else here is trying to do.

SANTOS: I can't help but ask the psychologist question, which is a chicken and egg question, which I'll illustrate with chickens. So, imagine you ran your network analysis on chickens. I don't know what chicken networks look like ...

CHRISTAKIS: Someone's done that, by the way, but go on.

SANTOS: They don't look like humans, right?

CHRISTAKIS: No, they don't. But elephants do.

SANTOS: The primate stuff we're getting out of Cayo Santiago suggests but other animals form networks too, but the question is that why is that, right? And so, you started by talking about this fact that humans might have networks that are unique or unique to more closely related primates or whatever, but then why at the psychological level could that be? Is there something about human cognition or human cognitive mechanisms that allows us to form those networks, and not other species? And, if so, then it seems to me that the individual, at least what's going on in the individual's head, shapes this

CHRISTAKIS: I think it's fascinating. Leaving aside the eusocial insects and clonal species, where the interactions between the individuals are necessarily amongst kin, we're talking about non-kin relations, so we've got primates, including us, elephants, cetaceans; what's amazing to me is that what's known about the network mapping of these individuals, of these species, is that those networks look incredibly similar. Elephant networks and primate networks and dolphin networks look very much like ours.

To me this begs what is a really interesting question, which is, maybe there's only one way to be social. I mean, why would it be the case in the natural world that whenever we go looking at social species, leaving aside the eusocial insects, would they evince these network properties? Because the last common ancestor between us and whales was 60 million years ago. So whales clearly have evolved independently, and with elephants, it's about the same.

So, they're converging, by convergent evolution on a similar solution, not on a bodily phenotype, but what James and I are calling an exophenotype. So think about this (to borrow an example from Richard Dawkins): Why is it that if a spider evolves bigger mouth parts to capture more prey, we think of that as a kind of evolutionary adaptation; but if a spider evolves the construction of a more elaborate web that basically achieves the same thing, we don't necessarily think of that as a phenotype? Well, actually we should. Let's start thinking of it as a phenotype—spiderweb morphology is a phenotype. If that's true, by a few short leaps I could get you believing that social network construction is a phenotype. My manipulation of the social world to construct the network around me, I would argue, is no different than the spider's manipulation of the physical world to construct a spider web around it.

Second, picking up on your point, what's amazing to me is that dragging with it, not necessarily dragging with it, but walking along with the network structure are all these other things. For example, mirror self-recognition. Dolphins have mirror self-recognition, primates clearly do, elephants have mirror self-recognition, cooperation, self-identity, and other-identity. So if you're going to cooperate and form networks with non-kin, you have to be able to know "Oh, this is June, and that's Sendhil, and that's Danny..." you have to know who they are from moment to moment. And these other animals also do that. So there's this suite of features that seems to be necessary and go together for the construction of social worlds.

SANTOS: Just to follow up, do you have to be the kind of cognitive creature who could do X, Y and Z and then you're like, "Oh, I'll talk to June and then, ooh, the network forms?" Or does the network form

and that creates this crazy selection pressure to have these mechanisms ...

CHRISTAKIS: Yes, it's both, I think. That our social life and our biological heritage are in a conversation across eons.

Think about this: Imagine a beaver, for whatever reason, has a chance mutation that makes its behavior different so it constructs a bigger dam. And now when the beaver constructs a bigger dam you get a bigger flood behind the dam. Now across time those beavers, ideally to exploit the greater linear perimeter of the pond that they've created, which gives them more forging opportunities, need bigger lungs. So the beaver now, because of the behavioral change, has to start evolving bigger lungs to be able to be underwater more to explore this perimeter, or bigger flippers or whatever beavers need to be effective. Okay?

Well, I think humans are like that, actually. We have little things where we begin reworking the social world around us. That creates selection pressures on our brains and our cognition; it makes us social. The more social, cooperative, mirror self-recognition, all that other stuff we do, the more able we are to create these webs around us, and it feeds back on itself. But what's so interesting to me and James about the social world is that, unlike the physical or the biological world, which is 'God-given' [or exogenous] and all around us, we create the social world. We create the selection pressure that then feeds back and contorts our minds and contorts our bodies. That's what we think is happening.

GRUBER: You're talking about social contagion and I know earlier we talked a little bit about emotional contagion. I'm just wondering to what extent do you think the spreading of this phenomena is going to vary depending on the type of network we're talking about, whether it's specifically, I think of offline or in vivo interaction and now with the social media networks growing and ever-increasing, and the degree to which you're expressing emotions in these two domains is radically different.

CHRISTAKIS: One of our arguments has been that, with respect to emotional contagion (in which we're very interested), there has to be some relationship at stake. My emotional response to my child in pain or my colleague in pain even, depending on the colleague, is very different than my emotional response to a stranger in pain. I still have empathy, I'd like to believe, and sympathy for the stranger in pain, but there's clearly something different about it. Plus, it's also different to see the person in pain than to read about the person in pain. So, (a) the nature of the social tie, and (b) the visibility, are crucially important. However, I think that you can transmit emotional states to

a lesser extent, but still, through online interactions. Like if you get a sad letter from your sister, you're going to feel sad about it even though it's a printed word and not quite as powerful as seeing your sister.

We have an unpublished paper, which I can talk about very briefly, in which we exploited weather variation as an instrument. We looked at all the residents of New York City, and if it rains in New York City, with Facebook mapping of the whole country, their Facebook friends in cities outside of New York are affected by the weather in New York City, to two degrees removed. I won't go into all the details, but we did this in the econometrically way, and so we can discern, in a kind of quasi-natural experiment, to the extent that you believe the literature that weather affects people's moods (which there is a nice cottage literature on this), you can use that as a kind of, what's known as an instrument, to identify these effects between online friendships.

KURZBAN: Like all models, there's a certain degree to which you're abstracting, and that's a necessary feature of modeling, right? You've got to take some stuff out. And you highlighted something that's right about psychology, which is that we don't spend that much time thinking about what friendship is for. There's been some assumptions about it being for exchange and so on, and you have a different proposal.

Just to connect you back to some stuff that we've been thinking about, one thing that seems to be important in our data about friendship is that the nodes aren't equally weighted. The amount of time and the degree to which I'm close to my best friend is really different from my fourth and fifth friend. What I'm really curious about is, first of all, as a technical matter, how easy it is to build things like that into the model. Just for the record, my suspicion is it's going to be really important. It might even change your packaging data.

CHRISTAKIS: Yes.

KURZBAN: Because if I spend a lot of time on doing that … Right? So, as a psychological matter, that seems like a reality, which would be very cool to build into these sorts of things. And, again, as an empirical matter, we're finding that there's a relatively nice function that one can use to map these things. So, is the future of this kind of weighted edges ...

CHRISTAKIS: Yes.

KURZBAN: And what's going to happen?

CHRISTAKIS: Yes, there's a big move to weighted graphs exactly for the reason you're describing. Every tie can get a weight now, so you can describe ties and not just the nodes. In fact, ties can become just as complicated as people. How long has the tie been lasting? How intimate is the tie? How frequently do you see the person? What's the vector? Do I say, "You're my friend" or you say "I'm your friend?" And so you can begin to have all kinds of details, which are highly relevant, and you can weigh the ties and use a variety of methods which allow you to take advantage… And it falls mostly as you would predict, right? So, just as you suggested, people with whom I spend a lot of time are more important paths through the network when it comes to germs, for example.

BROCKMAN: Unfortunately, Wallace Stevens couldn't be here today, but he asked me to read the following excerpt from his poem, "United Dames of America":

> "The mass is nothing. The number of men in a mass
> Of men is nothing. The mass is no greater than
>
> The singular man in a mass. Masses produce
> Each one its paradigm.

~~~
~~~

We're here in early September 2013 and the topic that's on everybody's minds, (not just here but everywhere) is Syria. Will the U.S. bomb Syria? Should the U.S. bomb Syria? Why do some people think that the U.S. should? Why do other people think that the U.S. shouldn't? These are the kinds of questions that occupy us every day. This is a big national and global issue, sometimes it's personal issues, and these are the kinds of questions that social science tries to answer.

JOSHUA GREENE:
The Role of Brain Imaging in Social Science

We're here in early September 2013 and the topic that's on everybody's minds, (not just here but everywhere) is Syria. Will the U.S. bomb Syria? Should the U.S. bomb Syria? Why do some people think that the U.S. should? Why do other people think that the U.S. shouldn't? These are the kinds of questions that occupy us every day. This is a big national and global issue, sometimes it's personal issues, and these are the kinds of questions that social science tries to answer.

What I want to talk about today is the role of neuroscience, specifically brain imaging, in social science. So far, neuroimaging has told us something between very little and nothing about these kinds of big questions. What I want to talk about is, why is that? What has neuroimaging accomplished in the last 15 years or so? What has it failed to accomplish and why? And what's the hope for the future?

I should say a little bit about my background. I'm a neuroscientist. And at the moment I think I'm the only neuroscientist here, but I'm not a neuro-evangelist. I didn't really begin my academic career as a neuroscientist. I started out as a philosopher, and I think of myself as much as a philosopher and a psychologist than as a neuroscientist. I use neuroscientific tools but I use other tools just as much. So I'm giving you my perspective as someone who's a user but not an evangelist.

The key psychological distinction behind what I want to say is the distinction between process and content. What functional imaging has done very well is connect parts of the brain to different processes and has taught us some specific things about how certain systems process information in a general way. What neuroscience has not done very well, but is starting to—not so much in a way that connects very directly with big social scientific questions, but is starting to—is to actually get at the content of thinking.

I'm going to talk a little bit about where we are now, how we got here, and how I think important developments in functional neuroimaging may finally deliver on the promise that a lot of us felt for neuroimaging when it was really getting rolling in the late nineties.

When I started doing this in the late nineties, I was very excited. I thought that this is opening the "black box" with everything that the metaphor entails. The brain scanner would be something like the microscope of cognition. Just as biologists ground some lenses and were able to see these little things swimming around in exquisite detail, we'd finally see the little critters swimming around in the brain—the little cognitive critters doing their thing. We would understand things on a fundamentally different level than the way we had understood them. I think there've been a lot of triumphs, but we haven't gotten there yet. We haven't gotten our mental microscope, and the question is, why haven't we? What do we have? And what's it going to take to get there? Will we ever get there?

I'm going to start off telling you about a brain imaging study that I did recently, and I'm focusing on this experiment more as a matter of convenience. You can call it narcissism, whatever you like, but it has certain features that I think illustrate where things are, but not necessarily where I think things are going. This is an experiment done with Amitai Shenhav, and the experiment was looking at how people make moral decisions when there are outcomes with varying magnitude and probability involved. So, a specific situation would be like this:

You're a Coast Guard sailor. You're on a rescue boat going to save one person from drowning, and you know that you can definitely save that person. Then you get a radio signal that says that in the opposite direction there's this boat that's capsized. There are five people there and you can save them, but there's only a 50 percent chance of success. Or, say, there are not five people, there are ten people; there are 20 people; there are 40 people. It varies in this experiment the number of lives at stake for the chancy option, as opposed to the one sure thing of saving one life. And we also varied the probability of saving them (with a little twist that I won't get into).

Perhaps we can explain something like why we just don't care if it's 100 lives that we can save or 1,000 lives we can save, why past a certain point it all just sounds like "a lot." That makes sense from the perspective of a brain that's designed to seek out valuable things until it's had as much as it can use. Once I've saved a dozen lives, I'm "morally full." My ventral striatum only goes up to here.

The question is, what in the brain is keeping track of that probability parameter? What in the brain is keeping track of the magnitude parameter? The number of lives you can possibly save if you change course. And what's putting those two things together to give a sensible answer? If you want to give a sensible answer you have to think about your odds and you also have to think about how big the moral reward is.

What we found is that when people were making these kinds of hypothetical moral decisions, the structures that were doing this kind of work—keeping track and integrating the relevant parameters—were the same kinds of structures that you see in humans when they're making self-interested decisions about real rewards like food and money. You see homologues of this in other mammals, like rats.

In our case we found that people's behavioral sensitivity to the probability parameter was associated with the sensitivity of their anterior insula to the probability parameter. Do they care about probability, and how much? You can look at their insula and make a better-than-chance guess about that. How much do they care about the size? If you look in the ventral striatum, which is one of the brain regions that we heard about earlier in Fiery's talk, that's a brain region that's sensitive to the magnitude. Then the ventral medial prefrontal cortex seems to be sensitive to the interaction—that is, putting these two things together. This parallels very closely, as I said, what people found when they looked at self-interested economic decision making.

What does this tell us? Well, it says that there's this quite general process of assigning values to outcomes, and there are these general parameters that apply to a lot of different things: Saving lives versus getting more money for yourself or versus getting more food for yourself. We have these domain-general systems that we use, and when we think about something like saving hypothetical lives, we rely on the same kind of circuitry that a person or even a rat might use when they're deciding, "Should I go this way and try to get the really good food or should I go this way and get the less good food that's more certain?"

From a neuroscientist's perspective, this is not terribly surprising. What else would it be? We've been studying this stuff in rats, and human brains look more or less like big rat brains. It's an important caveat. But from a moral perspective, as a philosopher and as a psychologist, you might've expected something different. Not long ago, not a lot, but at least some people thought there was a dedicated system for making moral evaluations, a kind of "moral organ" or

"moral grammar." These kinds of results and others suggest that that's not tenable.

We've identified a kind of process that's involved and it seems to be a quite general process. Does this tell us anything interesting about moral decision- making? Well, maybe a little.

Here's an interesting thing about moral decision-making that many people have documented: People seem to value human lives and other moral goods with diminishing returns. Saving one person's life, that's really good. Two… three… That's a little bit better. By the time you get to the hundredth life it's leveling off. Now, why would that be the case? In a sense it's very strange. Why is the hundredth life worth any less than the first or second or third life that you can save?

If you know that the system that we're using to attach values to these things is a system that evolved in mammals to attach values to things like food, then having this kind of diminishing returns built into the system actually makes a lot of sense. In that sense, from an evolutionary perspective it's not surprising that you would see diminishing returns built into the system. Not that it makes normative sense. Perhaps we can explain something like why we just don't care if it's 100 lives that we can save or 1,000 lives we can save, why past a certain point it all just sounds like "a lot." That makes sense from the perspective of a brain that's designed to seek out valuable things until it's had as much as it can use. Once I've saved a dozen lives, I'm "morally full." My ventral striatum only goes up to here.

That's a theory about why we intuitively have diminishing returns for saving people's lives, and the theory comes from the neuroscience. But the more general point is this: Understanding the kind of process can give you some insight into some of the quirks of the decision making process.

That's a hypothesis. There are other explanations for what might be going on there, but at least if that hypothesis is right, it tells us something interesting about the ways we make judgments, including ones that may have life-and-death consequences. That's an experiment. What it didn't tell us is how this actual thinking works. What we're doing is implicating in both the moral and non-moral case the same kind of reward system—same system for representing the value of outcomes. But, of course, somewhere in the brain you understand that now you're talking about saving hypothetical lives rather than foraging for food or making gambles. In the moral version I'm imagining that I'm working for the Coast Guard as opposed to making a gamble in which this button will give you $1 with 50 percent probability. Somewhere your brain obviously knows the difference

between those two things despite the fact that the same general valuation mechanisms are involved.

This comes back to the difference between process and content. What brain imaging has been good at, and what it's essentially figured out, is that there are a relatively small number of major cognitive networks in the brain. You have brain networks that are involved in different kinds of perception, different kinds of motor actuation, in storing and retrieving long-term memories, in, as I described, attaching values to different outcomes under conditions of uncertainty, and so on and so forth. What we've found when you compare mental activity A to mental activity B, is that you see a little blob on your brain image indicating increased activity here for mental activity A. And when you compare B to C, you see another little blob indicating a local difference in overall level of brain activity between these two mental tasks. If you lower the statistical threshold on all these blobby brain images, all these little blobs end up expanding into familiar brain-wide structures—like the peaks of different mountains expanding to reveal the same mountain ranges. The blobs that you see on your brain images are the peaks of these mountain ranges, which are essentially large- scale networks in the brain. And what we've essentially learned is that pretty much any kind of juicy, large-scale, interesting cognition involving perception, involving imagery, involving memory, involving choices, involving social perception… It's going to involve all this stuff--all of the same major networks in the brain. And you get to a point where you say, "Okay, it's all involved."

We can go a step further than that. In my study with Shenhav on saving lives, we didn't just say, "Well, our mammalian reward system is 'involved'." We had a bit more of a complicated story. We said, "This brain region is tracking the probability, and this one's tracking the magnitude, and this system seems to be playing this integrative role, putting those two pieces of information together." People who do this more seriously than I do have more detailed computational models, and you heard from Fiery Cushman some discussion of some of those. We can look at these general systems, we can say, "Okay, these systems are involved," and we can say something more about the general operating characteristics of this system. And sometimes knowing something about the general operating characteristics of the system will give you some sort of insight into something that you might care about even if you're not a neuroscientist, like: Why do I not care so much about the hundredth life I can save?

But what's the next step? What does it take to actually understand our thinking? And this is where I think new advances in functional neuroimaging are—or at least could be—very important.

To flesh out the distinction between process and content there's a nice analogy—it's not mine, I wish I knew where it came from—where you can imagine hanging a giant microphone over a city like Beijing. What would you learn if you did that? Well, you wouldn't learn Chinese. What you'd learn is certain things about patterns over the course of the day or week or month: When this part of the city's very busy, this part of the city's not so busy. And when there's a disaster all of a sudden these things fan out from the central area and go to all of these different areas around the periphery. You would learn things about the function of the city, but you wouldn't learn the language of the city, right? The question is, what would it take to learn the language? To bring neuroimaging closer to psychology?

Consider a classic neuroimaging study: If you want to understand working memory, you can give somebody a word to remember and ask them to hold onto it. And while your subject is remembering that word you can see structures in the dorsolateral prefrontal cortex and corresponding regions in the parietal lobe that are keeping that information online. Looking at the a set of brain images, you might know that your subject is remembering one word rather than five words. You can see the effect of the higher load—more activity, more function at the same time. But doing this kind of neuroimaging, the kind that dominated for the first ten years or so, you wouldn't know what word your subject is remembering. The content is neutral, not reflected in the brain images. What you're learning about is the process.

Starting with a breakthrough paper in 2001 by Jim Haxby, people have started to use brain imaging to look at content. They started with categories, then individual concepts, individual intentions, and so on. Haxby's insight was that there's a lot of information that we're losing doing neuroimang the standard way. The standard way of doing neuroimaging analyses, you have mental task A over here and mental task B over there—for example, remembering a word string, two-long versus five-long. Then you'd look at the activity for those two tasks and you'd subtract, and you'd say, "Okay, the extra work for remembering five words versus two words seems to be here in this circuit, and we think that's the working memory buffer." Haxby's insight was: "Well, look, there's all this information in these patterns of neural activity. And it may not be about what's overall up or overall down at the level of brain regions the size of the tip of your pinky or bigger. The micro-details make a difference."

You can imagine, for example, training a computer to tell the difference between paintings. If you have, let's say, Starry Night over here and an equivalent size Van Gogh painting of sunflowers, that's a pretty easy distinction. You just average the overall brightness, and you can say, "Okay, that's an A, a copy of Starry Night, and that's a B,

a painting of sunflowers. That's a bright one, and that's a dark one." If instead you had two classic paintings by Mondrian, the kind where you have the lines and the color patches, you couldn't necessarily distinguish between them by averaging, by looking at the overall brightness, or even by examining the kinds of compositional elements that are used. You have to look at the pattern. That's what multi-voxel pattern analysis and other multivariate methods are about. With any good thing, it's always possible to oversell it, but I think that there's really a lot to this stuff.

What's been done? Well, the original experiments examined brains performing perceptual tasks. Things usually start with perception. Can you tell whether someone is looking at a chair versus a wrench, versus a face, versus a place? Sure enough, you can look at these patterns—without paying attention to what's generally going up or going down, focusing instead on the microstructure of the pattern—and you can tell what someone's looking at! Then people did it with acts of imagination. You could tell whether someone's thinking of a face or a place or an object by looking at these patterns—and in a way that you couldn't using overall subtractions of what's generally up or what's generally down.

More recently, people have done things with concepts. There's a really fascinating paper by Tom Mitchell and colleagues where they had a huge corpus of words and they created a map of the semantic relationships among all of these words. So, obviously, dog and cat are going to be closely related, dog and nuclear reactor not so much. Then they mapped brain patterns onto a set of words. And then they took those brain-mapped words, and they took a new word that was not ever looked at with the brain imaging before. And they asked, "What should the brain pattern for this word look like?" They used "celery" and "airplane". They showed that you can make a pretty good guess about what the brain pattern will be when someone's thinking about celery or airplane based on what the patterns look like for other words, other concepts, that are more similar versus less similar to "airplane" and "celery".

Another recent classic: You give people two numbers and you say, "Don't tell me, but just in your head plan on either adding the two numbers or subtracting the two numbers." This is John-Dylan Haynes' work. And you can tell ... that is, not "read" the intention, but make a better than chance prediction, seconds in advance, about whether or not someone is going to add the two numbers or subtract the two numbers.

...if neuroscience is going to matter for the social sciences, if neuroscience is going to teach us things about whether or not the U.S. is likely to bomb Syria and why some people think it's a good idea and some people think it's a bad idea, and how our understanding of what happened in Libya, or what happened in Iraq is informing our understanding of whether or not we will or should bomb Syria... To understand those kinds of complex social things, we're going to have to really understand the language of the brain.

Recently people have started—this is Jack Gallant's work at Berkeley—to reconstruct still images and even videos from brain patterns. So, finally getting at the content. Not just what kind of processor is engaged—how does that process generally work? what kind of variables does it use? what are its temporal dynamics?—but the actual stuff that's being processed.

What's the significance of this? Well, the first thing that everybody thinks is, "Oh, so, 'brain reading,' people are going to be able to read your mind, read your brain." I think, at this point, and for a long time, if you want to know people's secrets go through their garbage, read their Facebook page. It's going to be a long time before the best way to get at somebody's secrets is by looking at their brain scans.

The long-term promise of this is really about understanding the "language of thought." That phrase was made famous by Jerry Fodor. He had a specific theory that comes with a lot of baggage, but the idea that there has to be some kind of language of thought in the brain actually makes a lot of sense. If I tell you something in English and then someone later asks you a related question in French, you can take the information that you learned in English and give the answer in French. There has to at last be something that's translating that. Likewise, you might've seen something visually but you can respond with a description of it in words. Or you can point to a picture. Later, you can access the same information from memory. You can use that same information to guide actions with any part of your body. In other words, there seems to be this global informational access within the brain where the same pieces of information can be used by almost any kind of system—a sensory system, a motor system, memory system, systems for evaluating outcomes. There's got to be some general-purpose format for representing the meanings of things.

What we've started with now with neuroimaging experiments are relatively small things, object-like things, such as an intention, a concept, a perception, a visual image. What we don't yet have is, first of all, a detailed understanding of the structure of these

representations for these specific things. Moreover, we don't yet understand how these things get put together—how thoughts get put together in the same way that sentences get put together from words. I don't know if an understanding of that constructive processes is a few years off, if that's decades off. This is something that I along with Steven Frankland have just started thinking about and working on.

But to bring this discussion full circle, if neuroscience is going to matter for the social sciences, if neuroscience is going to teach us things about whether or not the U.S. is likely to bomb Syria and why some people think it's a good idea and some people think it's a bad idea, and how our understanding of what happened in Libya, or what happened in Iraq is informing our understanding of whether or not we will or should bomb Syria… To understand those kinds of complex social things, we're going to have to really understand the language of the brain. We're going to have to really understand the content that is being processed and not just the kind of processing and the general operating characteristics of that processing.

Is multi-voxel pattern analysis—and multivariate methods for neuroimaging more generally—going to answer this question? I don't know. What I do think is that this is our current best hope. This is our current best way forward for actually understanding and speaking the language of the brain, and finally getting the mental microscope that I was hoping for back in the late nineties when I first started doing this.

DENNETT: Josh, the last point you made about multi-voxel pattern analysis reminds me of one of the first points that was made today about "big data." It looks to me as if you are anticipating the future, even with rose-colored glasses on, where now, thanks to big data and really good multi-voxel pattern analysis, we can to some degree read people's minds. But we're not going to know why. We're not going to know what the system is. We can do brain writing, we can do brain reading but doing brain writing will be, if we can do it at all, we won't know how or why it works and we'll be completely dependent on the massive technology, not on any theory we've got about how representation is organized in the brain.

GREENE: I agree with you that that's where we are now, and you may be right, we may be stuck here for a long time. I may be as disappointed 15 years from now about the hopes that I'm expressing now as I am now about at least certain things that I hoped for when I first started doing this. But I think there are some real reasons to think that it's not necessarily a false hope.

A really fascinating recent experiment, also by Jim Haxby, uses a method called hyperalignment. It begins with a technical problem, but it really gets into much more interesting conceptual territory.

First, the technical problem: When a group of people are engaged in the same task, everybody's brains are representing all these things, all of the pieces of information that one needs to perform the task. But surely what's representing these kinds of fine, micro-cognitive details is not going to be in exactly the same place for me as it is for you. Neuroanatomy differs from person to person, just like gross anatomy So, how do you normalize people's brains? How do you put different people's brains onto the same map?

Haxby said, "Well, I'm going to have people watch a very complicated stimulus—I'll have people watch movies—and I'm going to start with one person's set of brain data, and then I'm going to do a factor analysis. Factor analysis is used to find the natural structure in a data set. For example, if you're Netflix, and you want to understand the general structure of people's movie preferences, you can take a massive data set consisting of different people's movie ratings and submit it to a factor analysis. And the analysis will tell you that there are some general factors corresponding to familiar categories such as "romantic comedy" and "action adventure."

Haxby wants to know what the major factors are in the patterns of neural activity that arise when people watch a complex stimulus such as a movie. What is the underlying structure within those patterns? How do the components of these patterns hang together to form more coherent representational categories?

In the beginning, it was just a technical matter. How do you align these brains? But what he found was that if you take one person's brain and you find their major components, and you take another person's brain and find their major components, the components seem to line up extremely well. If you can translate the specific topography of one person's brain into the topography of another person's brain, or into a common topography using this higher order space, that's one important step forward.

If you're a pessimist you say, "Well, factors-schmactors. Are those going to correspond to some kind of interesting cognitive units?" Maybe they will; maybe they won't. But these kind of experiments at least give me hope that we can do more than just, in a brute force kind of way, separate the A's from the B's, predict the A's from the B's. It gives me hope that we can actually find deeper structure, that we can then start saying, "Okay, well, if you were to manipulate this component this way, then you would have an overall representation that looks more like this instead of a representation that looks more

like that." In other words, when you look at the level of components within these neural patterns, it might start looking more like a language and less like a schmeer of big data.

DENNETT: A schmactor-schmeer.

BROCKMAN: We're calling your talk "Factor-Schmactor."

PIZARRO: Every time I say what I'm about to say it sounds like I'm a pessimist of some sort about neuroscience, but I'm a real optimist about neuroscience. I just want to check to see whether we're talking about the same thing. You offer out this hope that we will learn deep things about psychology, and then it turns out that what we're learning is really, really interesting things about the brain. So, for a neuroscientist this is amazing. In fact, it is a level up from just maybe regular neuroscience because what you're essentially saying is we're learning the assembly here, the machine language of the brain. But I'm still not sure that we know more about the psychology. It's one thing to say we know how neurons encode concepts in this sense, but it still seems as if it's dangled out as a hope for a psychological theory that it's not just elbow grease that we need to get to the psychological theory. It's that even when we solve that problem, as Dan was saying, there's still this huge open question about the psychology.

GREENE: It's the mind-body problem. It's transparency.

PIZARRO: I don't know. Don't call it that. It's not the mind-body problem.

GREENE: Sorry. It's not the problem of consciousness. Let's be more specific. It is the problem of understanding in a transparent way the relationship between mental phenomena and the physical phenomena of the brain.

I freely admit that that gap has not been closed and so, therefore, I can't tell you when it's going to be closed, even a little bit. But let me try to do what I did with Dan, to give you some sense for how this could go. In the Mitchell experiment that I described where they're predicting brain patterns for things like "airplane" and "celery," the algorithm hasn't seen the brain data for these words, and yet it can predict reasonably well what those brain patterns will look like. For that prediction to work, there must be certain semantic features that, say, different vegetables have in common, and those semantic features seem to be represented commonly among different kinds of vegetables, or different kinds of artifacts or things. When it comes to concrete objects like celery and airplanes, we can make a pretty good guess about what the kinds of features are likely to be: Does it have

leaves? Is it green? Does it move on its own? Does it make a roaring noise? And so on and so forth.

What happens when we start getting into abstract concepts? For now, to try to get a result, researchers are looking at concrete nouns. But what if you start looking at more abstract things? You can start all the way with things like "justice" or something like that. But you might also consider things that are somewhat abstract and somewhat concrete, like the idea of a market. A market can be a literal place where people exchange goods, but a market can be a more abstract kind of thing. Just as you can identify the features that are relevant for classifying different kinds of more basic concepts such as "celery" and "airplane", where you're probably not going to get any big surprises about the relevant features, you may be able to do the same kind of thing for more abstract things. For most basic concepts, I can say, "Oh, well, plants kind of look similar and vehicles kind of do the same sort of thing." But when it comes to understanding the relationships among abstract concepts, it's much harder to know where to start, and doing something like this kind of analysis can help.

Another thing that neuroimaging has revealed is that there's a huge, important distinction between animate and inanimate things. This seems to be the biggest split in terms of where object concepts are housed in the brain. And while that's not super surprising—it seemed like a good candidate—you wouldn't necessarily know that that's true just from observing behavior, at least the literature that I've seen.

CHRISTAKIS: Could you do sneaky experiments like show a celery-shaped car, for example? I'm serious, like trying to screw with the brain?

GREENE: One foot in front of the other. As far as I know, there haven't been…

CHRISTAKIS: Like a dog-shaped nuclear reactor?

GREENE: You do see things like that in other domains. There's this distinction between objects and places. And some clever person says, You've got a place like a house, you've got an object like this toy. Well, what if it's a toy house? Can you see the kind of shifting back and forth from a more object-like representation to a more place-like representation, depending on how it's framed? In the more perceptual end of things you see stuff like that. That's not too far off from your celery car.

~~~
~~~

The findings in comparative cognition I'm going to talk about are often different than the ones you hear comparative cognitive researchers typically talking about. Usually when somebody up here is talking about how animals are redefining human nature, it's cases where we're seeing animals being really similar to humans—elephants who do mirror self-recognition; rodents who have empathy; capuchin monkeys who obey prospect theory—all these cases where we see animals doing something really similar.

LAURIE SANTOS:
What Makes Humans Unique

I'm going to talk about some new findings in my field, comparative cognition. I'm interested in what makes humans unique. There are findings that I think are fantastically cool, in that they might be redefining how we think about human nature, but first they're going to pose for us some really interesting new problems.

I'm doing this, in part, because I think already having redefined human nature in the last couple of years is sort of a tall order, and that scared me, but also because I think that open questions about human nature can actually be more fun and I couldn't help but use this audience to kind of get some feedback on this stuff.

The findings in comparative cognition I'm going to talk about are often different than the ones you hear comparative cognitive researchers typically talking about. Usually when somebody up here is talking about how animals are redefining human nature, it's cases where we're seeing animals being really similar to humans—elephants who do mirror self-recognition; rodents who have empathy; capuchin monkeys who obey prospect theory—all these cases where we see animals doing something really similar.

Today, I'm going to talk about two sets of findings where we're seeing, at least in the case of nonhuman primates, young nonhuman primates doing something really different than humans. In one case they're doing something different than humans, which you might think of as cognitively less rich. That makes the human looking like, "Wow, they're super smart." But in a second case they're doing something that looks like animals in this case, so the nonhuman primates are doing something that's cognitively a bit more rational, but I think it's also going to lead to some deep insight into human nature. So those were what I took my marching orders to be, and now I'll sort of jump into two separate findings.

As I do that I'm going to violate another principal immediately that John told me to do, which is to stick to questions and findings that are very, very recent. The first set of findings bear on a question that's, in fact, very, very old, and it's a question that Premack and Woodruff posed way back in 1978, asking the question of whether or not the chimpanzee or any other animal has a theory of mind. What Premack meant by this question was, does the chimpanzee look out into its world and see all these agents just behaving—doing different behaviors? Or do they do what we do, which is to intuit inside everybody's heads all these things going on—things like intentions, and theories, and beliefs, and desires, and so on?

This is a very old question, as I said this was 1978. Some of us around the table weren't born yet, but some of us around the table thankfully were born and were writing important critiques of Premack and Woodruff studies, (Dan) which were really important to the field, because it got off the ground this question of what could actually count as evidence for this question. We can verbally talk to each other and come up with the idea that we think of each other as having beliefs and desires, but how could you ask this of a non-linguistic creature? What would really count as evidence that they're thinking not just about behaviors, but about these mental states that are different from behaviors?

...if you really want to know whether an organism is thinking in terms of other individuals' behavior, or thinking in terms of what's going on inside another individual's head, you have to use these creative kinds of cases where what they would do if they're monitoring behavior is different than what they would do if they were thinking in terms of what's going on inside somebody's head.

Dan, and Zenon Pylyshyn and others who commented on this really important paper came up with a set of marching orders for the researchers at that time about how you could design studies to potentially tell the difference. That's what launched, in the eighties, this long field of what's been known as false belief studies. Many of you know about this, but for those that don't, please just be patient with me.

These are studies which are trying to look at whether or not people are actually representing the beliefs inside someone's head as distinct from their behaviors by using this special case of false beliefs—this special case where people are doing behaviors that don't necessarily

match what you might see in reality. So if I had a false belief that this event was over, I might do something crazy in my behavior like, get up, take my microphone off, go inside, have a couple beers, and so on. That would be different than what I should really be doing in reality—what reality should be telling me—but there's this sort of false content in my head, this sort of false thing that's going on. And, cleverly, folks pointed out that if you really want to know whether an organism is thinking in terms of other individuals' behavior, or thinking in terms of what's going on inside another individual's head, you have to use these creative kinds of cases where what they would do if they're monitoring behavior is different than what they would do if they were thinking in terms of what's going on inside somebody's head.

This launched this whole set of inquiry in the field of developmental psychology, where I think developmental psychologists had a bit of a leg up on those of us who are comparative cognition researchers, because they had the tool of language to ask children about different scenarios. This led to a long history of research showing that it seemed like there was some important developmental changes over the first couple years of life, and children's ability to think about what's going on inside the heads of others. The comparative researchers, though, were a bit stymied, and they came up with a lot of experiments that, even though they didn't have the fantastic commentaries that came after Premack and Woodruff's paper, I think if they had, people like Dan, and Pylyshyn and others, would say the same thing, "These aren't good tests to really get at what's going on in terms of what other animals know about other's mental states."

This was the state of the field well into the nineties, until researchers started coming up with what I think are somewhat better tests that use these nice nonverbal measures to come up with good tests of whether or not other animals have false beliefs. Here's where I have to kind of give a nod to a conversation we were having earlier about the "Noble" Prize, which for those that don't know watching this, would be a potential prize they were hoping someone out there would donate lots of money for so we can give prizes to researchers who, upon having evidence that their idea was wrong, admitted that their idea was wrong.

Here I have to give a shout out to one potential winner of this, who is Mike Tomasello, who in 1997 wrote a book that said, "I don't think any other animal has any representation of other individual's mental states," and in 2003 he wrote a paper that updated that and said, "Because of new evidence I have to say that I was completely wrong in that book. I published that book, and I was wrong. Now there's good evidence that they do."

What's that evidence? Well, the evidence comes from a variety of different tasks showing that other animals seem to process information about other individual's perception or visual access. One version of this type of test asks: do other nonhuman animals actually pay attention to what other individuals can see? So if you give them the option of trying to deceive somebody who is looking away versus somebody who is looking at a piece of food, what you find is on the first trial with no training, nonhuman primates know who to steal from; they steal from the guy who can't see.

They also seem to know something about the fact that visual access passes into the future. So if somebody saw something at one point, they might recognize that that past visual access predicts that that individual might know something about what's happening in the future, and, therefore, won't steal from that individual, and so on. As Mike reported, we're starting to get more evidence that primates are doing better than we thought, but so far there hadn't been a really good test that would qualify for the kinds of critiques that Dan and others brought up.

Until such time as a group of clever developmental psychologists came up with a very good nonverbal false belief test—a nonverbal test that allowed us to show that maybe these might be representing something that's going on inside somebody's head, and we, as comparative cognition researchers, like it very much when developmental psychologists are clever like this, because when they come up with a good nonverbal test, we can then take it and do it ourselves and get the same answer.

And this is what happened a couple years ago when Onishi and Baillargeon came up with a good nonverbal test of false beliefs that they used in 15-month-old infants, that we and others were later able to import to nonhuman primates.

And here's how the test goes. Imagine that I say Danny, in this case, is either a monkey or an infant, and you're watching a display of me acting on the world. Later I'm going to ask the predictions you make about my behavior. Imagine, if you will, that I have a PowerPoint that shows an image of me with two different boxes where I'm hiding objects. So Danny, just this casual observer, will be watching as I hid an object in one of these boxes—I'll hide it in the box on the left. The question is where do you think I'm going to look? Well, if you were correctly representing that I had a true belief, you might expect me to reach over here. However, you might find it surprising, if you understood my true belief, that I would reach to the second box that didn't have this object that I desired.

It turns out that both 15-month-old infants and, in our case, Rhesus Monkeys show that effect. If you monitor how long they watch this event, or quasi-measure their attention or their surprise, they look longer at this event where I reach in the wrong spot. So we're just saying they're tracking information about what I might know about the world and how I'm going to act. The question is what happens in this critical case of a false belief, where reality should be telling me to do something, but my belief, if you understood it, would be telling me to do something else. This would be a case where again two boxes, I would place an object in one box, and as I wasn't looking, the object would move to the other side. Now if you're tracking my belief you should expect me to go to the box where the object was, but if you were analyzing my actions just in terms of my behavior, you might expect me to go where the object is because of course that's where it is.

What do 15-month-old infants do? The 15-month-old infants, when they see me put an object here, they see it move to the other side, they expect me to reach over here, and they're very surprised if I reach in the correct box. And this was some of the first evidence that within the second year of life babies might be tracking another individual's false belief, published in Science, this was a great thing. What we did was to say, "Ah, this is a fantastic test. Let's apply this to our macaque monkeys." I have to be honest, when we first did this test, I assumed if 15-month-old infants are tracking this information, that's exactly what the monkeys are going to do.

So, again, the test is put the object here, person's not looking, object moves over here. And as I hit the button on our stats package, and looked at all our data, as I hit the button to generate the means, I thought we're going to see one of two patterns.

The thing that was really curious was that we didn't see either of those two patterns. What we saw was that in both cases the monkeys looked very little at these different options. They looked very little when I put the object here, it moved, whether I reached here or whether I reached here. And that was really different than what we'd seen in the other case. And they said, "Why?" What it looked like is that the monkeys aren't just behaviorists, in this sense. They're not just tracking what my behavior is. They don't expect me to reach where the object is. But the monkeys might not have a full-blown representation of another person's belief, the content of where it is.

What it seemed that they were doing is they were tracking our visual access. We, as researchers, keep referring to this as knowledge, although we take it that this is not what philosophers refer to as knowledge, but monkeys are tracking our historic vision access and expecting individuals to act on it. What happens when you lose that

visual access is that all the representations go away, all bets are off. You're just ignorant. And the monkey might expect you, or Danny in this case, might expect you to look on the moon, because you don't actually know where the object is. This was surprising to us, because it wasn't the kind of result we expected. As we followed up on this, it turns out that the monkey system for thinking about how we act seems to, again, not have any representation of other's beliefs, but it seems to be relatively sophisticated in its own right.

By 15 months of age babies seem to be tracking other individuals' false beliefs, but this raises this question of whether or not they have this other same system that's going on under the surface, that's also tracking this visual access, too.

Well, the first thing we've learned is that it seems to take into account what other individual's inferences are, and this is work not by me, but Mike Tomasello and his colleagues looking at the kinds of simple inferences you might make about where a piece of food is hidden. So they did this clever experiment with chimpanzees, where they had a delicious piece of food that they hid behind a screen, and when they lifted the screen, there were two pieces of cloth on the table, one that was totally flat, and one that was beveled exactly in the shape of the food. They asked: can chimpanzees smartly make the inference that the food has to be hidden under the beveled thing? The answer is yes. Not so surprising. Chimpanzees are pretty smart.

The surprising thing is that chimpanzees can also represent in another chimpanzee that same inference. So if they watch a different individual have this test where they see a piece of food hidden, one is upright, they have the same intuition that the chimpanzee should search in that spot.

The second, even more surprising thing we found is that the way the monkeys seem to shut off their inference about whether you have visual access or whether you have knowledge seems to actually be pretty sophisticated, and seems to not bear on what you may expect from behavior. So here's this test that we ran. Again, in one of these situations Danny would be watching me hiding different objects. You'd watch as I hid the object in this location, and just as I couldn't see, it popped right out and went right back in. So all the features of the world should tell where I'm going to reach, I should reach over here. But this is not what we find in the nonhuman primates. What we find is that they, again, say well, you lost your visual access. You should be

ignorant. You can search on the moon. So even though all these features of the world are telling them the way we should behave, we seem to have this interesting disconnect.

Why am I telling you all this stuff? First, I think we're finally getting some important insight into this age-old question that Premack and Woodruff gave us about whether or not other animals are mentalists. And I think the answer is that they don't seem to have representations of other's false beliefs, but they might not be as tight a behavior as we thought in the first place.

The second insight, and the reason I think this bears importantly on human nature, is it seems like we have a phylogenetically old system to track information about an individual's visual access that seems to be present in monkeys, and we have no idea yet whether or not it's present in humans as well. By 15 months of age babies seem to be tracking other individuals' false beliefs, but this raises this question of whether or not they have this other same system that's going on under the surface, that's also tracking this visual access, too. And I think that makes some interesting predictions about whether you get some disconnects between cases of these two systems, cases where what you're tracking with the sort of phylogenetically earlier system tells you something different. I think those kinds of questions would be very interesting to explore, and might redefine the way we're thinking about how other animals track other minds.

So that's set of questions number one, which I, in part, wanted to tell you because I think Mike Tomasello should win the Noble Prize; he'd certainly be getting my vote. The second set of studies I wanted to tell you about I think are even more relevant for some of this stuff we've already talked about, because I think in some ways they fall out of this case of us being a species that has a phylogenetically relatively recent system for representing other's beliefs. And the possibility I think is that when natural selection builds in new systems, they tend to be a little bit kludgy and they might actually have some problems inherent in them.

This raises the question of how we deploy our systems for representing other's beliefs. How is that we look out into the world and think that Danny might have a certain belief about something, but he's ignorant about something else. How quickly do we deploy these things? And there's a couple different options. One is that we're kind of cognitively lazy. We should only deploy these kinds of complicated systems in these cases where we really, really need to. So if Josh were to give me some complicated moral scenario about some guy knew something, but somebody had another belief, I would have to turn on all this machinery to make sense of it. But I shouldn't be kind of doing

it haphazardly, just when there are kind of random things around the screen.

The second set of results I wanted to tell you is that it seems like that's not actually the case. It seems like there might be some interesting automaticity to the extent to which we turn on our mindreading abilities. And it seems like this automaticity might be different in nonhuman animals. This comes from a study that came out by Agnes Kovacs and her colleagues recently in Science, where she was asking, again, about the automaticity with which we start thinking about another individual's beliefs.

And it must have been the most boring study ever for subjects to do, because what it involved is just a subject, say Danny, in this case, is just tracking an object that's moving behind an occluder and all Danny's task was is to say when the object fell behind the occluder and the occluder went away, does he think the object is there or not. Just a basic visual detection test. And, of course, since we're tricky cognitive scientists, what we'll do is have some trials in which the object looked like it went back there, but when the screen falls, it's gone. And, of course, even though Danny is a fantastically smart person, he's going to make errors and be slower when I mess with him in that way. And that's just what you find. No surprise there.

The question is what happens in the case where there's another individual who happens to also be watching the scene, who has a different perspective than you do, who might even have a different belief about what you're seeing than you do, and the way Kovacs and colleagues tested this was to put a cartoon Smurf head on the screen, so the Smurf is on the screen while Danny is doing this task. It's completely incidental. Subjects know the Smurf doesn't matter at all. But it sometimes shares Danny's belief. Sometimes it sees it go back there just like Danny, and the screen drops, and it's gone. And sometimes it actually has a different belief. Sometimes it turns away at this critical moment where the object moves.

And the question was, even though this is a cartoon Smurf, even though it's completely incidental from the task, does it affect the way Danny responds? And I think the surprising answer is yes. What you find is that if the Smurf thought something was back there, even in the case where Danny didn't, he speeded up. So he doesn't take a reaction time pause for a belief that he would have had that was false. There's another individual in the scene who has that belief.

What does this mean? Well, it means a couple things. One is that we might be implicitly tracking the perspectives and beliefs of a variety of other individuals around us. This is the thing that Ian Apperly and his colleagues refer to as altercentric interference. We might be

getting this interesting interference by other people's beliefs, other people's contents, even though we know them to be different from our own.

Why am I, as a comparative person, telling you this? Well, we've recently been able to run a study like this on nonhuman primates, and what we find is that the monkeys are a lot more rational than people in this sense. They don't seem to be automatically computing other individual's visual perspective, and they don't seem to get messed up. In this sense the monkeys react as though if they saw the object back there, it's back there. If they didn't, it didn't. Okay?

What are the implications for some of this stuff? Well, I really wish Fiery was still here, because one of the implications, I think, is that we might have automatic systems for tracking what other individuals know, and speculatively I can extend this to what other people intend, what other people's attitudes are, and so on. These things might deploy automatically and be relatively under the hood in a way that we might not expect, but that's exactly the kind of mechanisms you might need for the sorts of uniquely human things Fiery was talking about—namely things like social learning, namely things like picking up on other's reinforcement histories, all the kinds of things that humans do that we think of as unique —might rely on this kind of kludgy mechanism, where we just get interference with the contents of our own mental states versus somebody else's.

Is this really true? Do we see any data that something like this might really be happening? This is an extra third line of comparative studies that are coming out that I'll tell you about, which is some interesting work on the cases in which other animals can socially learn from us, and cases in which humans might learn from others in a way that's less rational than other animals.

One of the leftover empirical results from the 1990s is often folks think that other animals can't imitate. It's not true. They can actually follow our own actions and imitate, but they tend to do it in relatively select situations. What are those situations? Well, it tends to be situations in which they, themselves, don't know how to do something. So if you give chimpanzees an opaque puzzle box and they have no idea how to open it, what they will do is they will watch how you open it, and they will follow exactly what you do. If you give, in contrast, chimpanzees a transparent puzzle box and they can kind of figure it out, they just go on the basis of what they know.

The critical question is what I've just told you predicts that humans might do worse at this task, and this is what Vickie Horner and Andy Whiten tested, where they gave these opaque puzzle boxes and transparent puzzle boxes to chimpanzees and children, and they gave

them a demonstrator who wasn't a smart demonstrator, but who was doing something dumb. So imagine you see a puzzle box, you don't know how it works, but you see me take a tool, and I probe into the top of the puzzle box in this little opening, and then I use the tool to open up a door in the front and I take candy out. What you do is you then give this to children and chimpanzees. It's an opaque box. They don't know how this works. They do exactly what the human demonstrator did, they probe into the top and use it to open the box.

Now, the critical test is you bring out a transparent box and you can see that the box is just empty. All you can do is open the door and there's the candy. But you see this demonstrated. He painstakingly sticks the tool in the top, opens the thing. What do you do? Well, chimpanzees just cut to the chase. They just open the door and take the food. What do human four year olds do? They slavishly copy exactly what they see the human do. And you might think, well, the kid doesn't want to, you know, annoy the human adult, who's just been teaching them. A graduate student at Yale, Derek Lyons, did a whole variety of controlled conditions to show it's not that the kids think that this is normative. Watching an adult demonstrator has changed the way the kids think about the causal mechanism of this box. They think somehow, I don't know the causal mechanism, but you have to do this thing at the top, or else you can't open it.

This is very profound, and, again, it suggests that in some ways animals, in their noninterference or cross-mental states, might be more rational than us, but I think this provides a powerful mechanism for teaching, a powerful mechanism for the kinds of rewards structures that Fiery has talked about, and potentially a powerful mechanism to solve the chicken and egg problem that I was asking Nick about earlier, which is if we want to know why these crazy things transmit through networks, things like our attitudes, or whether or not we smoke, or whether or not we're obese, and so on, it might be that if we're constantly walking around automatically having interference between other people's attitudes and beliefs, that's a really easy way for just being around some friend to transmit these kinds of things.

All of this stuff I talked to you about at the end has been pretty speculative, but this is exactly the reason I wanted to talk about this stuff in front of you guys. I'm not sure, if you followed John's marching orders, you get deep insight into human nature just yet, but I think these new kinds of findings where we're seeing differences are pointing us to new directions not just in the ways that humans might be unique cognitively, but the way these different cognitive mechanisms might play out in a broader context, allows us to do all kinds of human thinking things, like culture, and so on.

Just to kind of round out the discussion we had last night at dinner, I hope I've posed some interesting new questions for you, given you some zany speculation, and talked to you about some spots where the jury is still out. Thank you.

DENNETT: My bumper sticker these days says, "Competence without Comprehension." The idea is that human comprehension is built up out of competences which are themselves relatively uncomprehending. The Whiten result fits beautifully into that, in that it even permits you to speculate that it's an adaptation for cultural transmission that we ape more than apes do, and this opens the gates for all sorts of advanced techniques that we can acquire, and then have in our toolkit, that we don't yet have to understand. They bring us benefits, and then we can build other things out of them, but we don't require any level of comprehension in order to take them on, and then they can help us develop comprehension later.

SANTOS: Yes. Although I think with some of the other over imitation results you might need to amend the bumper sticker to, "Competence, not comprehension, but then later comprehension," because I think the powerful thing about some of these results is not just that the kids followed the behavior, it's that they developed rich causal explanations based on the fact that somebody had an intent to do something. And so the thing that I find most fascinating is it's not just the behavioral transmission, what goes with it when you see an intentional human do something, it's the fact that it must have been done for a reason. There must be this explanation, and kids, based on this social input, are completely willing to override the physics.

One of the powerful results that Derek had is he asked the separate children how this object works, and all of them are sharp enough to exactly know the physics of how this object works. You see a human do a dumb thing on this object, this kind of strange thing that you wouldn't do. All of the kids override what they saw before, not that you just have to do it, but that this is how the object causally works, and they spin a ton of different interesting stories that don't make any physical sense to come up with how this works. So it's not just that you can get these things without comprehension, but seeing it build in a comprehension that may or may not be accurate, based on your knowledge.

CHRISTAKIS: The thing that you're saying now that prompts a thought in me, and it was also prompted by something June said earlier today was, of course, experimentally these are fascinating things, right, to think about the way you're describing them, and the

experiments are so fiendishly clever, like it makes me want to switch fields, and do the experiments, and there's so much thought and creativity in making them. And, of course, when we do experiments, we isolate down to particular actions, and so forth.

But maybe it's the case that while it's seemingly "irrational" for the baby to behave this way in this clear puzzle box, in aggregate it's better for the organism to do what the adults do. And, of course, you know, it's like genes and competition, right? I mean you could have, you know, "dysfunctional genes," or emotions earlier we were talking about. I mean there may be ways in which across time it makes no sense for you to be happy when the world is collapsing, when we look at a single packet of time, but maybe on average, in fact, across time maybe it's good for you to feel happy no matter what's happened. I don't know. I'm making it up. But the point is if you expand your horizon maybe it's no longer as crazy. Maybe it's my resistance to not wanting to think that chimps are smarter than I am, but, you know, when you describe it like they are behaving more rationally, yes, in this particular case, but maybe more generally that's a price we pay for …

SANTOS: So it makes a prediction about the kind of extended phenotype in which we humans find ourselves in, which is that the social information we get is often pretty accurate …

CHRISTAKIS: And sometimes we're led astray. Yeah.

SANTOS: And sometimes we can get led astray. And for the kinds of physical environments where we, at least as modern humans find ourselves in, that's, for sure true, right? If I were just to use my physical intuitions to try to figure out this iPad, I would be completely screwed, but as soon as Josh hits one button and does it, then I have insight into this.

When Derek Lyons talks about some of these results, he always starts his talks with the latest whatever the winning new Rube Goldberg experiment is. He puts that up, and then a coconut, and he says the coconut is the most complicated thing in the chimpanzee world, this is the causal thing that they cannot figure out, whereas we deal with these causal systems that are incredibly complicated. And he chooses Rube Goldberg to say the beauty of these is that you can, with your naive physics, understand all this stuff, but that's the teeniest, tiniest crazy causal system that we have to deal with as humans. We're constantly faced with these causal systems that we just don't have the ability to understand, but other people do.

And I think the interesting thing, the reason I think this relates to Fiery's stuff, is that it might not at least be for complicated causal systems, it might be for elaborate social reward structures, elaborate sets of goals and behaviors that you want to link together, but you, yourself, haven't done yet, and I think we really need to look en mass atthose kinds of cases and ask the question: Do these kinds of low-level mechanisms work in all these cases, and do they ultimately derive the kinds of smart answers you're talking about?

CHRISTAKIS: Have you seen that YouTube video that went viral a year or two ago of a little baby, I don't know how old, holding a magazine, like an old-fashioned magazine, and going like this to try make the picture bigger? They have to learn, you know …

SANTOS: The other thing that you get out of this is the power of your social input, and one of the things you learn if you hang out with toddlers who have access to iPads, it's just how incredibly reinforcing these structures are. I think part of it is that they're reinforcing for that reason that Fiery is talking about, that they're getting incredible social input that this is a reinforcing thing. They see their parents and caregivers around these objects, interacting with them in a way that says this is the more important thing, than any food or anything. They're like the rats that were getting the cocaine except they're like the rats that are getting the iPad. But the key is the kids don't have to do that themselves. The inputs we're providing are getting sucked in in different ways.

KNOBE: It seemed like the answer you were giving to Nicholas was that what we really want to do is understand the causal structure of some object, but luckily there are people around who already know it. We're just kind of using them as a means to do this other task. But I wonder if there's any evidence for that view as opposed to another possible view that it's not really as crucially important for us to get the right answer about the causal structure of this object. It's just to get along with other people.

What we're really concerned with is not using other people as a means to correctly understand the causal structure of this object, but interacting with other people and working with other people in certain ways, and even if we get the causal structure of the object wrong, if we connect up with other people by doing it the same way they do, and we're better off.

SANTOS: Right. So if we make really strong cooperators, even if we don't understand how the box works, we still don't get attacked or whatever.

KNOBE: Suppose I have the option of getting it right, but everyone thinks I'm a weirdo, or getting it wrong, but everyone things I'm good. Maybe I'd be better off getting a wrong answer.

SANTOS: Yes. Well, the question is why it has to go with the wrong answer, though, right? So you could imagine a whole set of mechanisms of conformity that didn't go with the competence, plus comprehension, plus comprehension, extra part, right? You could imagine a whole case of conformity that was of the form, "Wow, like Josh is such a weirdo when he's opening the thing that way. I'll open it that way in front of Josh so he won't hate me, but as soon as Josh leaves, like that's it, because I know that that's not the way to do this. I just won't test the waters of being a jerk in terms of my conformity."

But that doesn't seem to be what kids are doing. So the fact that their causal analysis goes along with [what others do] that suggests that it's not just about relating, or something similar, or setting up your ingroup to do actions in the same way. The fact is that what goes along with it is a rich causal analysis that goes beyond what I think just if we're trying to get long. I mean maybe that might come along for the ride, and so on, but I think we need an extra thing to explain why that part comes, too, and I think that's the nice thing about some of these studies, is that they've kind of controlled for that possibility.

The way Derek did it was really elegant. So a child comes in, and they learn this task. They see the experimenter do things and Derek, as the experimenter, convinces the child that the experiment is totally over. The child is like, oh, the experiment is over, the kid gets their prize, everything is fine. And then Derek convinces the child that some emergency has happened. The emergency is there's a new child there, and we all forgot to check if the object was back in the thing. So somebody needs to open the object as quickly as possible while Derek leaves, and nobody is going to watch, but it's got to be incredibly fast. Nobody is going to watch you, and like it's very, very urgent. Derek runs out of the room, and what you see is not that the child stopped doing the stupid thing, they just do it really fast, and really urgently. It doesn't seem like it's about just relating. It seems like it's really changing their comprehension, and I'm not sure why you get that, but you do.

CHRISTAKIS: That's so clever.

~~~
~~~

What is the field of experimental philosophy? Experimental philosophy is a relatively new field—one that just cropped up around the past ten years or so, and it's an interdisciplinary field, uniting ideas from philosophy and psychology. In particular, what experimental philosophers tend to do is to go after questions that are traditionally associated with philosophy but to go after them using the methods that have been traditionally associated with psychology.

JOSHUA KNOBE:
Experimental Philosophy and the Notion of the Self

I'm going to be talking today about some recent work in the field of experimental philosophy. But before I talk about what this actual recent work has discovered I want to say something briefly about what this field is.

What is the field of experimental philosophy? Experimental philosophy is a relatively new field—one that just cropped up around the past ten years or so, and it's an interdisciplinary field, uniting ideas from philosophy and psychology. In particular, what experimental philosophers tend to do is to go after questions that are traditionally associated with philosophy but to go after them using the methods that have been traditionally associated with psychology.

If you want to get a vague sense of what this field is like, you might consider the analogy of neuroeconomics. If you open up a typical paper on neuroeconomics, you see this experimental methodology and statistical analyses that would be very much at home in just any other kind of paper in cognitive neuroscience. But this work is infused with this tradition of theories and concepts from this much older tradition of economics.

In much the same way, if you open up a typical paper in experimental philosophy you see these experimental methods and statistical analyses that look kind of the same as any others that you find in psychology. But this work that people are doing is being informed in certain ways by these theories, by questions, by concepts in this much older tradition of philosophy.

Over the past few years there've been all sorts of work in experimental philosophy on all sorts of different questions—the concept of knowledge, on consciousness, on morality.

But here I'm going to be talking about one specific thing that's really been exploding in the past couple of years and this is experimental philosophy work on the notion of the *self*. This is work on questions about what is the self? How does the self extend over time? Is there a kind of essence of the self? How do we know what falls inside or outside the self? I'm going to be talking about two examples of this type of work.

The first is about this question that philosophers have called the "question of personal identity." It's a question in philosophy that goes back, at least, to the time of John Locke. It's one that philosophers are still talking about up until the present day. You can get a sense for the question pretty easily just by thinking about a certain kind of initial question, and it's this:

Imagine how the world is going to be a year from now. A year from now there are going to be all these people in this world, and one of those people is going to have a very special property. That person is going to be *you*. So, with any luck a year from now, there'll be someone out there who's you. But what is it about that person that makes that person you?

At this moment you have a certain kind of body, you have a certain kinds of goals, and beliefs, and values, you have certain emotions. In the future there are going to be all these other people that are going to have certain bodies, they're going to have certain goals, certain beliefs, certain emotions. Some of them are going to be, to varying degrees, similar and, to varying degrees, different from yours; and one of those people is going to be you. So, what makes that person you?

Philosophers have discussed this in great detail and the way they usually discuss it is at a very abstract level and often with recourse to seemingly absurd, insane science fictional thought experiments. But although this work might seem initially to be so abstract that it could never have any bearing on how human beings actually think about any question, I think that this work in philosophy has actually led to some really interesting insights.

We're going to consider just one crazy thought experiment from the philosopher Derek Parfit, and this is the way it goes: Imagine that Derek Parfit is being gradually transformed molecule by molecule into Greta Garbo. At the beginning of this whole process there's Derek Parfit, then at the end of the whole process it's really clear that Derek Parfit no longer exists. Derek Parfit is gone. Now there's Greta Garbo. Now, the key question is this: At what point along this transformation did the change take place? When did Derek cease to exist and when

did Greta come to exist? If you just have to reflect on this question for a while, immediately it becomes clea that ther couldn't be some single point -- there couldn't be a single second, say -- in which Derek stops existing and Greta starts existing. What you're seeing is some kind of gradual process where, as this person becomes more and more and more different from the Derek that we know now, it becomes less and less right to say that he's Derek at all and more and more right to say that he is gone and a completely other person has come into existence.

A year from now there are going to be all these people in this world, and one of those people is going to have a very special property. That person is going to be you. So, with any luck a year from now, there'll be someone out there who's you. But what is it about that person that makes that person you?

So far we're talking about this seemingly crazy level of a weird science fiction experiment. But now try to think, in light of everything I've just said, about your own life. Imagine what things are going to be like in 30 years. In 30 years there's going to be a person around who you might normally think of as you, but that person is actually going to be really, really different from you in a lot of ways. Chances are a lot of the values you have, a lot of the emotions, a lot of the beliefs, a lot of the goals are not going to be shared by that person. So, in some sense you might think that person is you, but is that person really you? That person is like you in certain respects, but just like Derek on his gradual transformation into Greta, you might think that person is kind of not me anymore.

Once you start to reflect on that, you might start to have a really different feeling about that person—the person you're going to turn into. You might even start to feel a little bit competitive with that person. Suppose you start saving money right now. You are losing money and he or she is the one gaining the money. The money is being taken away from the person who has the values, the emotions, and the goals that you really care about and going to this other person.

Experimental philosophers began thinking about this kind of problem and thought that maybe this kind of work—work coming out of this very abstract tradition of philosophical reflection—can actually shed a certain light on how people think about their own future selves. There's been a whole bunch of different experimental studies on this but I'm just going to give one example. It's a study by Bartels,

Kvaran, and Nichols. It came out recently. In their study, participants were randomly assigned to get one of two different pieces of information about the self.

Participants in one condition were told: "Scientists have studied the self in great detail and what they've determined is that the self is really surprisingly stable over time. Even far into the future you're going to be, on a really deep level, fundamentally similar to the person you are today. Of course, certain superficial things might change here and there, but the person you're going to be in the future is going to be shockingly similar to the person you are right now."

Participants in the other condition were given the opposite kind of information, they were told: "Scientists have studied the self and what they discovered is this really surprising thing, that the self changes radically. Just a few months from now, many aspects of who you are now and the way that you think of yourself are going to be different. By 30 years from now, you're going to be completely different, utterly different from the kind of person you are now."

Then participants were told, after they had gotten this information and answered a few questions about it, "Guess what? We're giving you a special bonus for participating in this experiment. We're giving you some extra money for participating in this experiment, a special bonus, and now you have a choice. You can take any percentage of it for yourself or you can give any percentage away to this charity, Save the Children. So, you can take the money 100 percent for yourself, 100 percent for them, or any percentage of it you can give away to them."

But now here comes the trick. Participants in the study were given random assignments to a time at which they or Save the Children would get the money. So, participants in one condition were told: "In one week either you're going to get the money or Save the Children's going to get the money. So how much do you want to give to either of them?" Participants in the other condition were told: "In one year either you're going to get the money or Save the Children's going to get the money, so how much do you want to give in each case?" What you see here is something really interesting. In the condition where participants were told that either they or Save the Children were going to get the money in one week, the manipulation about the information of the self had almost no effect. It doesn't matter whether you're told that the self changes radically or the self is remarkably continuous.,Either way they gave roughly the same percentage away to this charity.

By contrast, in the condition where they were told that they're going to be receiving the money in one year, the information about the self ended up having a quite substantial and significant effect. So, when participants were told that they in a year were going to be radically different from the person that they are today, they were willing to give away a larger percentage of the money to charity.

What we see here in this experiment is that people's judgments about how much the self is changing is having an impact on how much of a difference they see between themselves and other people. In the condition where they're told that their self is remarkably stable, they think of themselves as being fundamentally different from other people—this person in the future has a special kind of connection to me that no one else could have. By contrast, in the condition in which participants were told that the self is going to be very, very different in the future, participants thought, "You know, I guess that person who exists in the future, is more similar to me than other people around. He has a little bit of a special connection to me, but every other human being also has a certain connection to me. The sense in which that person is really specially connected to me in a way that no one else can be has been diminished." So, what we see in this first example is how this very abstract notion coming out of this tradition in philosophy stemming from John Locke can actually be applied to understanding human behavior and to manipulate it to the degree at which people show generosity to others. That concludes our first example. Now, let's consider one more example.

The second example is very different philosophical question and the question is: Is there something like a core of the self, anessence of the self, the true self, who you are deep down? One thing you might think is: there's all sorts of stuff going on within our minds. We have all sorts of beliefs, goals, values, and emotions, but not all of this is equal. Some of these things represent our true self—the person that we truly are deep down inside – but of course, you might alsohave all sorts of other beliefs that you must manage to pick up in one way or another. Maybe you picked it up from some clever advertiser or something on TV, from the way your parents were, but those things aren't representing your true self. Rather, if you could get rid of those things, if you could get rid of these parts of your psychology, you'd be able to more truly reveal the person that you were all along.

For thousands of years philosophers have been interested in this question: What is the true self? In particular, they've been interested in the question: Of all the parts of you, which ones are the true self and which ones are this kind of superficial layer—the part of you that isn't really your core? If you go back to the ancient Greek philosophers, for example, to Plato and Aristotle, you find this view that our capacity for reasoning, for reflection, that is our true self. So,

the view that these people developed is if you really reflect on certain matters clearly and deliberately and you think on reflection, "This is fundamentally what I should do," then that answer—the belief that you come to on reflection—that is your true self. Of course, you might not do it. You might not actually do the thing that you arrive at on reflection, but when you don't do it, you're just failing to act on your own true self. You're not doing that thing that reflects the person that you yourself really are deep down inside.

Other thinkers in later centuries ended up having almost the opposite of that view. Many people thought exactly the opposite of this first view; that your true self is the thing that comes in your sudden impulses, your hidden urges, these flashes of emotion. It's not that your true self is going to be revealed when you think carefully and calmly about something. The opposite is true: to the extent that you're carefully and calmly thinking about something., that just obscures your true self. Your true self is that thing that comes out when you're overcome with emotion, when you're completely drunk, it's in those moments that really your true self is going to come out.

Experimental philosophers have been interested in this question as well, but the question that experimental philosophers have been interested in is maybe slightly different. Experimental philosophers aren't trying to ask the question, "Do human beings really have a true self and, if so, what is it?" Rather, the question has been, "Do people think of themselves and other people in terms of a true self and, if so, how do they decide which part of the self counts as this true self?"

In just the past couple of years there's been a surge of work in this area, including some really fantastic experiments by the philosopher Chandra Sripada. But here I'm going to talk about one particular study that I think really gets at that initial philosophical question. This is a study that was recently conducted. The lead author is George Newman, and then in addition there are two co-authors, Paul Bloom and myself. We were interested in this question about the true self and we assigned participants randomly to one of two conditions. So, participants in one condition received the following story: Imagine a person named Mark. Mark is a secular humanist and he travels around the world teaching people that there's nothing morally wrong about being gay and, in fact, he coaches people on techniques they can use to avoid being prejudiced against gay people, to overcome their tendencies to stereotype or be prejudiced against gay people. But Mark has a problem. Mark's problem is that he himself ends up having certain feelings of disgust toward gay people and he openly acknowledges this, just sees it as part of his own personal struggle.

In the story I've just told you there's a kind of conflict between Mark's mind; it's a conflict between System 1 and System 2; it's a conflict

between his more automatic emotional self and another part of him—his more reflective beliefs; and in particular, his more reflective beliefs are telling him there's nothing wrong with being gay. But on a more automatic visceral level he's having this emotion that he himself would reject. So, now the question is, considering these two parts of himself, which is the true self? Let's just try running the experiment right now. So, consider the belief he has, the belief that it's not morally wrong to be gay. We want to know, is this really part of his true self or is this just some other thing within himself such that if he could get rid of it, he'd be able to more fully reflect his true self? How many people think that belief is part of his true self?

DENNETT: We can pick one or the other?

KNOBE: Okay, there's an emotion and a belief and the emotion and the belief are in conflict. The question is about the belief. Is his reflective belief part of his true self? How many people say yes? And how many people say no?

KURZBAN: So, if we reject the ontology of the notion of a true self?

KNOBE: If there's no true self, then it cannot be part of the true self.

CHRISTAKIS: What if we think that both are part of the true self?

KNOBE: Both. Then the answer is yes. Should I try the vote again?

MULLAINATHAN: Yes.

KNOBE: Okay, how many people think it's part of his true self? How many people think not? So, participants answered this question and then after answering this question about the true self, they were given a very simple individual difference measure. The individual difference measure is just one item, and the one item is: Would you describe yourself as a liberal or a conservative? So, now we can look at answers to this question among liberals and among conservatives and what we find is this: Among liberals, the overwhelming majority say exactly the same thing that you say. They say, "That belief is part of his true self. It's the voice of his true self speaking to him, telling him, don't be prejudiced against gay people." As for conservatives, they say, "At the very core of himself, speaking to him is this voice of his emotion of disgust telling him this is morally wrong. Then he just picked up this thing from our present politically correct culture. It's kind of leading him in the wrong direction. If only he could rid himself of that, then his true self would be able to be revealed."

But, of course, there's also another condition. In the other condition participants get a story about someone who's conflicted but just in the opposite direction. So, here's the new story: Mark is an evangelical Christian. He travels around the world preaching to people his message—the message that homosexuality is a sin -- and he teaches people, coaches them in techniques that they can use to avoid committing that sin, having the self-control not to sleep with other people of the same sex. However, Mark has a problem. His problem is that he himself is gay. So, he himself has a desire to sleep with other men, and he overtly acknowledges this to people and describes it as part of his own personal struggle. So, then participants were asked the exact same question. You have this desire and this belief. Now consider the belief. This is a belief that homosexuality is morally wrong, that he should not do the exact thing that he viscerally wants to do. Is that part of his true self? How many people say yes? And how many people say no?

Once again, participants were asked whether they were liberal or conservative and now we see the whole thing flipping. So, in this latter condition the liberals tend to say, "His emotion, the seething urges he has, that is his true self speaking to him, telling him about this other form of life he could have where he could fall in love with another man, but that's been papered over by this thing on top of that that's getting in the way -- his Christian belief that he should not sleep with other people." By contrast, the conservatives say, "At the very core of himself is this belief, this moral belief in this Christian vision that he should not sleep with other men. But, unfortunately, he's just acquired from other people or from society this desire. If only he could rid himself of that, then he'd be able to more fully reflect the person he really is deep inside—this Christian person."

We seem to be seeing coming out of this data this surprising result. It's not that people think on the whole that reason is the true self and it's not that people think on the whole that emotion is the true self. They think that the true self is whichever part of you is morally good. So, when other people are looking at you, they think that certain parts of you are good, certain parts of you are bad. Depending on who you are, it might be that they think your reasoning is good and your emotion is bad or they might think your emotion is good and your reasoning is bad. Whichever one they think is good, they seem to be seeing that as the essence of your self and this other thing as something around that—something just covering it over such that if you could only get rid of that, the person you really are deep down inside will be revealed.

In this first study I just mentioned, the question people are asking is directly about the self, but a lot of the recent work on this topic has proceeded in a different way. By looking at these issues more

indirectly, by asking people questions that, on the surface, are about something else and then arguing that this question about the true self can actually shed some light on that, can help us understand how people are thinking about these other issues. So you ask questions about what it means to be truly happy, what it means to be in love, under what conditions someone's blamed for what they do. And now the thought is that people's judgments about what is the essence of the self can actually explain people's answers to those kinds of questions.

What's happening increasingly is that the line between what people are calling philosophy and what people are calling psychology is just increasingly blurred.

Just as one example of this sort of phenomenon, I wanted to mention a study by David Pizarro. Pizarro looked through people's intuitions about agents who act impulsively and do something either morally good or morally bad. In one condition participants are told to imagine a person who is overcome by this rage and because of his rage at this person in the car in front of him, he just smashes the person's windows in. Now, in the other condition participants are told about someone who's overcome by feelings of compassion and because she's so overcome by these feelings, she ends up helping a homeless man by giving that homeless man her own jacket.

Now the question in each case is how to evaluate this person morally. Here you see the striking asymmetry. When people are so overcome by emotion that they do something morally bad, people tend to see that as an excuse. They take away the blame that you would normally give that person. So the person who smashes someone's windows in because he's enraged is going to be given less blame than someone who smashed the other person's windows in after a cool, calm, careful reflection. Now, by contrast, in the other case, you don't see that effect at all. In the case where someone's doing something morally good because she's overcome by compassion, she doesn't get any less praise than if she did the morally good thing after a cool, calm, careful reflection. Why is that?

What Pizarro and colleagues showed is that that effect is mediated by judgments about the true self. Consider the impulse to do something bad and the impulse to do something good. When you have a sudden impulse to do something bad, people think that's not your true self. Your true self is this more reflective part. To the extent that you yield to that temptation, then you're not reflecting the person that you

yourself really are deep down inside and so you're not fully to blame for it. In the condition, by contrast, where you're overcome by compassion and you can't help but do something good, people think that that compassion you feel is your true self. And so, as a result what you're doing is a reflection of your true self and you deserve full praise for it.

What we see in the study by Pizarro and also in the earlier study by Bartels and colleagues is something that we're finding much more generally in this field of experimental philosophy.

Experimental philosophy is a movement that was started by people who were really deeply steeped in this kind of philosophical tradition. These are people who had spent years and years thinking about Aristotle, logic, the problem of free will, and they wanted to do some experimental studies that could help them get a deeper insight into those kinds of questions. And you could've imagined what would happen in that case is that as you went deeper and deeper into those questions experimentally, you'd be moving into more and more technical territory that was farther and farther away from anything that non-philosophers could understand.

What has actually happened is exactly the opposite. As we develop deeper and deeper insights into these philosophical questions, what we're finding is that we're coming closer and closer to the rest of the sciences of the mind. What's happening increasingly is that the line between what people are calling philosophy and what people are calling psychology is just increasingly blurred.

GREENE: Really interesting. I want to ask about the deep true self-experiment and I see two possibilities here. So, one is that what people judge to be the true self is affected by normative considerations, right? And the other is that it's the heart of the matter. I mean, you can imagine a store that sells mattresses and they say, "Well, if you buy a mattress we'll give 10 percent of the revenue to charity" and mattress sales go up. You wouldn't conclude that people buy mattresses in order to help the poor or something like that. Obviously there's something closer to the core of the problem than that. But is it the whole story? One way you can get at this, and maybe you've already done this, is have cases where it's relatively well-matched. So you have Josh, the philosopher, and you feel like that's your true calling, that's what you always wanted to do, but this life that you live as a dancer is what you did to please your parents. Or you could reverse it where you have it so that being a philosopher or being a dancer are a priori rated is about equally good or equally bad. If you look at things that are neutral, is there a tendency to think of the brute desire versus the more reflective desire, one is more the self

than the other, or does it depend on context? I'm curious if you've thought about that or if you already have the answer.

KNOBE: Excellent question. What we see in these data are two separate effects. So, one is the effect that I just mentioned, which is just whatever you think is good, you think that's the true self. The other is the anti-Aristotle effect. In a case in which you have an emotion going against reflection, there's a general tendency to think the emotion is the true self and the reflection is not the true self. Now, putting those two effects together, I think we can have a pretty good explanation of why people initially thought that reflection was the true self. The reason is not that there's something within us having the intuition that whatever is coming from reflection is the true self. It's just that, generally, in cases in which your emotions and your reflection are pulling in opposite directions, other people are going to think that the thing you choose on reflection is the more morally good thing.

If we sampled all cases, in general, there would be a tendency that the thing that people are doing on the basis of reflection is more often going to be seen as their true self than the thing they do on the basis of emotion. But if we can control for morality, then people think System 1 is the true self. So, in the study I just told you about, for example, even though there's an effective political view, there's also just a main effect of just wanting to say that the emotion is the true self.

SANTOS: I have one comment and one question. So the fast comment is I don't think we should shortchange experimental philosophy by comparing it to neuroeconomics. In my experience, experimental philosophers are typically both good at the experimental side and philosophy side and I'm not sure the same is true for both sides of neuroeconomics. But the question is, in the Bartels study you can imagine two reasons you get the effect that you do. One is that when you tell me, "Oh, my future self is so similar to myself in the future," it helps me in perspective take better care of my future self. And I'm like, "Oh, my future self is totally going to want rewards, I'll take the rewards." The other effect is that when you tell me, "Look, your future self is going to be totally different than similar to any random person," then that makes me in perspective take better care of any random person and, therefore, I'm happy to give the money to the charity than to any random person. So do we know which is doing the work? Is it thinking of your future self as more similar or is it thinking of your future self as basically the anyman and, therefore, you give the money to charity?

KNOBE: This is a really interesting hypothesis, which I hadn't thought of. The idea is that to the extent that you think of your future self as

dissimilar, it's not just that it decreases your interest in your future self, it's actually increasing your interest in other people. So, the answer as to whether that second thing is true, I don't know. But there is evidence that the first thing is true. So, if you don't make it a choice between your future self and other people obut rather, between your present self and your future self, then the more you think that your future self is different, the more you're going to be in the technical sense, impatient. You're going to exhibit more temporal discounting. But the fact that that first thing is true doesn't mean the second thing is not true. It could be that as you start to think about your future self as being somehow very dissimilar from your present self, you start to feel a greater communion with other people. You think, "Other people are more like me. They're people I should really care about."

KAHNEMAN: Have you seen the experiments where people have so much money they will save and this is affected by seeing their face morph into the face of an old man? So people contribute more when they have seen the morphing and they have seen what they'll look like as an old person.

KNOBE: One thing that's really striking about this other kind of study that both of you are bringing up is there's a tendency, when you hear that first result, to associate this idea of continuity with a certain kind of value judgment—that is, to say, 'Clearly we think of temporal discounting to the extent that people do it as bad, and we think patience is good. So, to the extent that people are thinking of themselves as continuous, that's more good. To the extent that they're seeing themselves as discontinuous, that's bad.' But once you think about this choice that you're making, not just between your present self and your future self, but between your future self and other people, then the idea that there's some specific value judgment associated with this continuity becomes suspect.

MULLAINATHAN: It's interesting in your example about durability of the self that you've chosen long-range durability. And the kind of problems that I've often looked at and I think a lot of people have looked at, it's much less long-range durability than very short-range durability: "Will I go to the gym tomorrow?" I mean, "Yeah, I'm going to the gym tomorrow morning." But tomorrow morning, some other dude's there and for whatever reason, that dude is not interested in going to the gym. So, it feels like this conflict actually happens at quite high frequencies, and I know people have talked about that, but is that simply a special case of the long-range durability that you're talking about? Is that fundamentally different? Do you see what I'm getting at? Here we might've called the self-control, but is the problem of self-control, the problem of dual selves, hyperbolic discounting. Is that something very different?

KNOBE: That's a really interesting question. So, when people think that the conception of the self affects this issue of temporal discounting, how much we care about our future self, the way that they're often thinking about it is this idea that in a year you're going to be really different from the way you are now. You really will have deeply different goals. And then you might think that person isn't really me in some sense. The idea that the next morning you'll already be different from the person that you are now, and that you could start to feel jealous of your next morning's self—that's not the first thing you think of. But it'd be interesting to try to extend this kind of prior research into judgment of that type. Maybe we'll even thinking, "I don't know, on Sunday I'm going to be a little bit different from how I am now. Maybe I should choose something that makes me get the goodies and not the Sunday self."

CHRISTAKIS: I just recently for various reasons have been really angry at my 20-year old self, because I made all these decisions in my twenties that are still binding me today. Well, who the hell was that asshole that did all that stuff 30 years ago? And I can imagine the set of experiments where you ask people retrospectively what they think of their 20, or 30, or 40-year old selves, depending on their age.

KNOBE: In a really interesting way this ties in with the topic that Rob Kurzban was bringing up earlier. Rob was talking about this problem in our field, which is that people adopt a certain view and then they stick to that view despite increasing amounts of evidence that that view is false. But maybe now in light of this we have a simple manipulation. In some sense that person five years ago ...

DENNETT: Was a different person.

KNOBE: But you shouldn't really feel ashamed of that person ...

DENNETT: I was brought up short a few years ago when I was talking with a philosopher and I mentioned that I really didn't want to go on living when I became a foolish old dotard and was an embarrassment and all that; time for somebody to push me off a cliff. And she said, "Wait a minute. If you had a brother who was embarrassing to you would you feel you had the right to push him off the cliff?" "No." She said, "Well, what makes you think you now have the right to make arrangements about that that future self who wants to sit there watching Bugs Bunny cartoons. " And I must say that did give me pause.

CHRISTAKIS: Just like the Odysseus ...

KURZBAN: Josh, so I do want to push this a little but not that anything turns on this in terms of the psychological or the experimental philosophy. This notion of a true self. So there are those of us, and I thought that Dan Dennett was among those from some passages in some of his earlier work about one of the worst ideas that bedeviled psychology about this notion. So, do you think there is such a thing as a true self? If you do think that, what would the ontology of that be? Let me just end with that. So as a modular person, I believe the mind has lots of different bits and pieces. It's just weird to think that you ought to privilege some bit of it over another.

And just as an aside, I don't know if you've seen people who wrestle with the results of IAT work, where they have this split, which is almost exactly the homosexual case that you're talking about, right? So, if you look at my reaction times, I don't like black people, if you look at what I say, I say, "I love everybody equally," right? And then if you look at the psychologists who wrestle with this, they explicitly, "We're not saying that this implicit guy, that's you, and this explicit guy, that's just a veneer." They're saying, "There are just two different representations in there and we don't have any reason to privilege them." So what are your ontological commitments and do those ontological commitments matter for what you're about?

KNOBE: The question is a really good one. The question is, given that there's no good psychological theories that involve an actual true self, why do people think that there's a true self? These kinds of theories that we're developing about human cognition can explain why, in the absence of any evidence for this kind of strange phenomenon, we would believe in it. So, what is the thing that's making us believe in it? Right now, we're working on this question, we don't know the answer to it, but one thought that we have is that the belief in something like a true self is the application to the self of a more general capacity we have to think of something like essence. So we have the idea of essence and compare our idea of essence to many different things. And then when we apply it to the idea of the self, we get this notion, the notion of the true self. And what we're seeing in the case of judgment of the true self is this kind of byproduct of our general way of thinking about things as having essences.

If we thought about other kinds of cases in which we might apply this notion of essence, we seem to apply this notion of essence using similar kinds of techniques but we wouldn't ever think in these other cases that the essence of something is actually, literally, a part of that thing. So, suppose you were thinking about a band, say, The Rolling Stones. You might have a certain notion that there's something like the essence of the Stones—what the Stones are really about. Then you might have this idea, you know, all the music that they've been doing since the late seventies is just a betrayal. So, the last 30 years of the

Stones is just a betrayal of this thing, the essence of the Stones, like, what the Stones really mean—that's what came out in "Exile on Main Street."

But when we think about it, we're not thinking that the essence of the Stones is something like a certain *part* of the Stones. Say that the Rolling Stones were in front of us, it's not like we could point to a certain part of the band and say, that is its essence. The essence is this normative notion that if you saw their complete works, you could pick out this thing that's what makes them of value.

Now, with human beings we also apply this notion of essence and it seems like the criteria we use to figure out what is your essence are the same criteria we'd use to figure out what's the essence of the band. We look at all the different things you do, then we try to think, what is the most value in all of things that you do? And we think that is your essence. But then when we try to interpret what it is that we've come up with when we do this, we don't think of it in that way that we would naturally think of the essence of a band, or the essence of the United States, or the essence of social psychology. Instead, what we think is that there's actually some thing in you, like the true self module; it's sending signals to other parts that are being overridden. And it's maybe that that gets us into trouble. When we think of this notion of essences, it's almost something like a psychological theory.

~~~
~~~

We had people interact—strangers interact in the lab—and we filmed them, and we got the cues that seemed to indicate that somebody's going to be either more cooperative or less cooperative. But the fun part of this study was that for the second part we got those cues and we programmed a robot—Nexi the robot, from the lab of Cynthia Breazeal at MIT—to emulate, in one condition, those non-verbal gestures. So what I'm talking about today is not about the results of that study, but rather what was interesting about looking at people interacting with the robot.

DAVID PIZARRO:
The Failure of Social and Moral Intuitions

Today I want to talk a little about our social and moral intuitions and I want to present a case that they're rapidly failing, more so than ever. Let me start with an example. Recently, I collaborated with economist Rob Frank, roboticist Cynthia Breazeal, and social psychologist David DeSteno. The experiment that we did was interested in looking at how we detect trustworthiness in others.

We had people interact—strangers interact in the lab—and we filmed them, and we got the cues that seemed to indicate that somebody's going to be either more cooperative or less cooperative. But the fun part of this study was that for the second part we got those cues and we programmed a robot—Nexi the robot, from the lab of Cynthia Breazeal at MIT—to emulate, in one condition, those non-verbal gestures. So what I'm talking about today is not about the results of that study, but rather what was interesting about looking at people interacting with the robot.

Nexi is a cute robot that has a very, very limited set of facial features and range of motion, and in fact, has wires coming out at the bottom, and moves in a little bit of a weird way. But within seconds participants were interacting with her as if she were a person (I say "she" because of the voice that we used. Nexi, as you might imagine, doesn't actually have a gender that I know of. We didn't get that far). This is not a novel finding. It's not surprising that people adapted so quickly. Within 30 seconds, people were actually talking to Nexi as if she were a human being, in fact, were saying things that were quite private. At the end of the study some people were convinced that technology had actually advanced so much that Nexi really was a robot that was talking, when in reality there was a graduate student behind the curtain, so to speak.

I'll perhaps bring up the oldest experiment that has been talked about. In 1944, the psychologists Heider and Simmel actually used animated

figures to display very minimal actions on a screen, and what they found was that people spontaneously described these actions as intentional and animated. They ascribed motions and goals to these small animated figures. One of the great discoveries of psychology, in general in the past 50, 60, 70 years, has been that we have a very basic set of social intuitions that we use to interpret the actions of other human beings, and that these are good for us. This is how I can even be here talking to you guys, looking into your eyes and thinking that you might understand what I'm saying. But the cues that we use to apply these intuitions have to be very minimal, right? All you need is a couple of eyes and a face, and you put that in your Email, and I know that something social all of a sudden is cued in my mind, and I apply my social intuitions.

...we see intentionality in agency where there is none at all. So we're quick to think that even a machine—a vending machine that doesn't deliver, that doesn't dispense what I order is angering me—and in some way I am making a judgment of moral blame, when, in fact, there is absolutely no intentionality there.

We have intuitions, very basic intuitions about the physical world, like one is about causality, but we also have social intuitions about intentionality and agency. These build into our moral intuitions, such as who deserves blame, and some of the work that's built this whole view of how we make these moral judgments comes from people in this room, like Joshua Greene, Josh Knobe, and others. So we know that one of the ways in which we use these social intuitions is to generate moral judgments about who did good things, and who did bad things, and who deserves blame.

But what's interesting about these intuitions is that they can easily misfire. In fact, Daniel Dennett nicely called this view of things the "intentional stance." And it turns out the intentional stance seems to be our default view. But what that means is that we see intentionality in agency where there is none at all. So we're quick to think that even a machine—a vending machine that doesn't deliver, that doesn't dispense what I order is angering me—and in some way I am making a judgment of moral blame, when, in fact, there is absolutely no intentionality there. So we're promiscuous with these judgments of intentionality.

We can even use our promiscuity with these judgments in clever ways. The psychologist Kip Williams and his colleagues have, in their

experiments on social exclusion, wanted to develop a paradigm for how to make people feel excluded socially. So what they did was say, "Well, let's just do a simple game. We'll have three people in the room, and they'll toss a ball back and forth, and two people will stop tossing the ball to the third person." Now, the two people who are tossing the ball to each other are competitors, but the third person is actually the subject. People feel really, really bad when that happens, and it's a very simple game. So they said, "Well, maybe we don't need to have a physical ball game. Maybe we can just have three people in a room playing a videogame, right, and we can do it on the computer. It will be easier." Sure enough, that works. People feel really bad if they stop getting the ball. "Well," they said, "maybe we don't need to actually have the other two people in the room. We could just tell them that there are two people in the other room." People still feel bad. It turns out you can even tell them that it's just the computer that stops tossing the ball, and they seem to feel just as bad. Now, these people aren't stupid, presumably. They're college students, after all. But it's so easy to find intentionality in agency when there is none, and it's so hard to squash it that we generate these sort of weird errors.

Now, these are cute errors, and we can use them to do psychology studies on social exclusion, and we can learn quite a bit. In fact, it's kind of funny that you would kick a vending machine or that you would yell at your Windows machine when it gives you the blue screen of death, but they're increasingly failing to be that cute, because the more complex society gets, it turns out that these intuitions are some of the only intuitions we have to make sense of a social world that's quite different from the world in which we evolved. We've known this for quite some time.

Take, for instance, collective action. When we talk about the action of a company, we say, "Oh, Microsoft did this." Well, at some level we know that this is the actions of a whole bunch of people, maybe even stockholders or the board voting; it's not one person, but we seem to think of it and track it as an individual entity over time. So we can generate a moral judgment to say that Microsoft is evil. We do this with sports teams all the time. We say, "You know, I hate the Knicks because in 1983 this happened." But the 1983 Knicks may have absolutely nothing to do with the current Knicks, only by name are they the same, but we track them—as Josh was talking earlier about—we track them as if there is this essence, and it's continuous, and they're agents, and they do things, and they make us mad, and they shouldn't have done that. So those are instances in which it's becoming increasingly clear that our social intuitions may not have a good match with the actual social world in which we live.

Social media is another good example. I know what it's like to communicate to a group of people that I can see. You're giving me some feedback. I know the appropriate things to say and the things that I ought not say, I think, but now I have 600 Facebook friends. I have none of the cues that I would usually get from people in a crowd, and maybe I'm just thinking of it as talking to one person. So I say something, and all of a sudden I forget that I'm also friends with my grandma, I'm friends with my former advisor, and they all see it, and so our social intuitions don't work. It's the wrong kind of intuition to generate what we ought to do in much of today's social world.

Part of what I want to argue is that this is increasingly problematic, and it's not just the case that our social intuitions are going to fail and make it so that we're going to be embarrassed at what we say, but that, in fact, they might stifle real progress, and especially technological progress and innovation, because they're the only lens we have in which to interpret our social world, but they don't fit any more.

Let me give you an example. Algorithms that look in my Email generate personalized ads. Now, one of the first reactions that people have when they see an ad that has been personalized for them is: What the hell? Who's reading my Email? That's creepy. So "creepy" is a word that comes up quite a bit. The truth is no one's reading your Email. It's an algorithm, right? Somehow we feel like our privacy has been violated, even if we are assured that nobody, in reality, cares about your Email, but nonetheless the cue that we're getting is what would generally be a social cue—that is, somebody has generated a suggestion to us that normally would come from a friend.

So I have a new Google device, and Google now can do this very, very well. This is a service that goes through all of the information that I've given to Google—either explicitly or implicitly— and it generates these little cards that tell me, "Oh, by the way, David, you have a dentist appointment, and you better leave now because where you are it will take you this long to get here, because of the traffic." Now, imagine that somebody came up to you and said, "Hey, Josh, you've never seen me, but I think that your wife is worried because I was just over there at the house, and maybe you should call her, because you usually call her at this time, don't you?" That would be creepy. That would be extraordinarily creepy. But I love this service, because it gives me such useful information. In fact, I think I get a great deal of benefit from it.

What I fear is that, as technology progresses, and more and more good things can happen in the world—now, technologies

might actually give us anything from curing diseases, to preventing disease through genetic means—we're applying intuitions that are old, and we're making moral judgments that these are wrong and inappropriate, when in fact, we can still make those judgments, but perhaps we should be using something else.

My social intuitions are firing that there is a creepy person reading all of my Emails and looking at all of my appointments, but they're wrong, nobody is, it's an algorithm. But we don't have intuitions about algorithms, and I don't think we're getting any anytime soon. The image that I have sometimes is of a middle-aged man who's a few pounds heavier than he used to be trying to squeeze those jeans that he wore in high school onto himself. And so he squeezes and squeezes, and they just don't fit any more, but he can go to the store and get a new pair of jeans, and there's no intuition store for us, right? As technology advances, there is no way in which we can rapidly generate new intuitions. So what this means is that when we hear about self-driving cars, all of a sudden we get really nervous. Even though we're certain that percentage-wise this would reduce the number of traffic accidents, it just doesn't feel right; I'm not in control; I don't like it.

So what happens? Technologies get stifled a bit, because they have to match our intuitions. One of my favorite examples of this is BMW. BMW got so good at making the cockpit silent by developing new technologies to silence all external noise, that all of a sudden people started complaining that they couldn't hear the engine, and the engine provides actually really good feedback for many people, and they actually enjoy it. What used to be a side effect is something that people now enjoy. So what BMW engineers did is they spent hundreds of thousands of dollars, if not millions, to develop an audio algorithm that could generate engine noises to get pumped through the stereo that would be contingent on the conditions in which the person was driving, what gear they were in and how fast they were going. That is now in the BMWs, and there's no way to turn it off, it is not an option to turn it off. So here's a case in which this company had to bend over backwards to accommodate an intuition people had.

What I fear is that, as technology progresses, and more and more good things can happen in the world—now, technologies might actually give us anything from curing diseases, to preventing disease through genetic means—we're applying intuitions that are old, and we're making moral judgments that these are wrong and inappropriate, when in fact, we can still make those judgments, but perhaps we should be using something else. What that something else is is maybe

a question best used for philosophers and ethicists, but it's something we'd have to consider.

Those are the implications of modern society and our old intuitions. The implications seem to be for technology and for society, but is there anything that we now should conclude about the way that we study intuitions? Does this matter at all for the science of psychology? And I think that one way in which it does matter is because the normative theories that we use, that is, when we have to decide, is this decision good or is this decision bad, which has been a very, very fruitful way of understanding the human mind, and pioneered by Danny Kahneman and others, one of the ways we study human intuition is we say, "Well, let's see where people make errors." So we poke and prod people much like we use visual illusions. We look at when mistakes are made, and then we see the structure of the intuition, and we can say, and this is very useful, and it's very beneficial, we can say, "Under these conditions these intuitions misfired." We can actually now implement policy that says, here's the way to get people to make the right decision. But now what this entails is a proper understanding of what the right decision is.

In judgments under uncertainty, when we're making problemistic judgments, there are well-developed theories about what probability judgments you ought to make under human conditions. Should you be Bayesian? What information should you bring to bear on this decision? There's some controversy, but by and large people know when you're making an error.

In the field that I study, of ethical judgment, we've ported over some of those same techniques, and we used this error in bias approach to study moral judgments. We sometimes say, "Well, under these conditions it appears people are making errors in the moral domain." It's much trickier that way though, because the normative account of ethical judgments is much less certain than the normative accounts of problemistic judgments—that is, people still argue about this. But we can still say, "Well, look, we have thousands of years of philosophers who have developed normative theories of ethics that we can at least agree in some cases it's an egregious error, right?" And so many of my colleagues and I have looked at human moral judgments and compared them to normative theories, and concluded, look, your ethical judgment misfired.

As society has increased in complexity, and as some of these technologies have been innovated, though, even those normative theories are failing us. So it's unclear to me what the right answer is to whether the impersonal nature of drone attacks or robots in war is an immoral action. I'm just not quite sure whether simply removing agency makes it a more egregious violation. I want to work this out,

but what this means is that I can't use a proper normative standard to compare human judgment. So I think the implications here are that, as we proceed and as we study human intuition, and as the background of these intuitions changes because society and the complexity of technologies changing, we have to more and more act in concert with people who are thinking deeply about what the right answers are. Then we can start comparing our intuitions and the judgments that they generate, but it's essentially I think a call for a bit more of the working out of the normative side, before we simply start willy-nilly accusing people of committing egregious errors in judgment. I don't think that we know quite yet what an error in judgment is about many of these things.

KAHNEMAN: I think we do know something about errors. Take framing effects; you don't know which framing is correct, but here are two things that by some logical argument should evoke the same response, and they evoke different responses. And that actually, I think, is the more common way in which when you think that there is a problem, and the problem is that we have intuitions, and they're not consistent, so that you can trigger—that is, you have three intuitions. You have intuition A, you have intuition non-A, and then you have the intuition that they should agree. And that is really the standard problem.

PIZARRO: I absolutely agree. In fact, inconsistency, I think, is one way to determine if a moral judgment is an error. So one way you can do it is you can show people both conditions of the experiment. So you say, look, if they get embarrassed, and they admit that they made an error, but not all of moral judgment studies are like this at all. In fact, for instance, omission versus commission. So some researchers call this the omission bias. But now when you show people both conditions, you say, "Look, isn't it silly that you made this judgment that killing is worse than letting die. Don't you agree that this is an error?" They don't have the framing response. They say, "No. Of course not. I didn't make an error." In fact, they jump up and down and say, "I will make the same judgment over and over again." In those conditions is where I think we're having a little bit of problem.

MULLAINATHAN: To build on what Danny is saying and to go back to the first part of what you said, there's a book by Everett Rogers, I don't know if you've ever read this book, *Diffusion of Innovations*; it's a good book. He's got whole chapters in there on what he calls congruents. He basically reviews a huge literature on how innovations are adopted or not. And he's full of interesting stories.

One story is very relevant to what you're saying. It's about Indian farmers adopting tractors—tractors as a substitute for bullocks that would pull. It's interesting, because people who have studied this noted that after the farmers have adopted tractors, every night they would go to the tractor and put a blanket over the tractor. Great for bullocks. Actually, a little less than good for tractors. I mean at best, mutual. That is actually a theme that appears again and again in *Diffusion of Innovations*, that people adopt or fail to adopt technologies and use them in a way that's congruent with the intuitions they've developed and prior technologies that they'd had.

There's also, I think, a way out of it, which is related to this notion of you have multiple intuitions. One way you can get adoption or use of a technology is to actually just find an intuition for which this technology is congruent. So oftentimes when you see this mis-adoption, it's not that it's incompatible with all the intuitions you have. It's incompatible to an intuition, but it can easily be framed, so think of Facebook as X, or don't think of the Google guy who gives, you know, that algorithm, as a creepy dude, think of it as … and then all of a sudden it becomes totally understandable, people use it quite well.

So I wonder to what extent moral and intuition, social intuitions are also fertile enough, and different enough, and inconsistent enough that inconsistency is now a good thing, because now, as a framer, you have more things to choose from.

PIZARRO: Right. I think that's very insightful. In fact, that's a solution out of this, which is, okay, let's get another intuition going. As Josh Knobe and I were talking about earlier, one of the features of some of these reactions that we have is that it's not just that Google knows information about them, it's the social delivery. It seems to match the features of other forms of social delivery, that when a friend informs you, "Hey, by the way, you know, you have an appointment, right?" It's those that seem to get the intuition going. It's not, for instance, if my car is smart, and it measures the miles that I use, and then it says, "Hey, it looks like when you're driving in the city you use this this much…" That's great. I don't feel that violated. But that's not a social issue. So one way maybe that we can work on some of these problems is quieting the social cues that might actually get inconsistent intuition.

MULLAINATHAN: Or just to build on it. An example could be imagine that the application is explicitly bad early on, so that you could see it learning in a way. That's a situation where it might be more palatable to you, because you're like, "Oh, now I can see the process by which it's learning, and I'm actually growing close." This is just to give an example of how you might want to …

PIZARRO: In fact, actually, if you could involve the individual as an agent…You know the elevator buttons that are placebo buttons, if you could make it so that I just had to just remind me to touch this button so that Google now really knows, but it actually does nothing. I would actually feel a bit more involved, and so I would feel like they didn't just surprise me with this, like someone creeping in your window.

JACQUET: But it seems like the McClure Studies in science go directly against some of the things you're saying, where people are willing to accept unfair offers in the ultimatum game from machines, and they weren't willing to accept them or reject them from humans, even though the humans were also machine generated. It was just a face that changed the way they made the decision. So it might be that humans are willing to accept computers as competent, or even be compassionate towards them as objects, but they're not willing to accept that they have actual moral domains.

PIZARRO: Maybe. That's a good point.

KNOBE: One of the really interesting points that David was making is that it somehow has to do with the packaging of what you're doing. And there was an interesting follow-up on the study Jennifer was mentioning, in which it's exactly the same study except instead of saying it's just a computer, they said it's a computer with special artificial intelligence program. And in that condition people would lose money to punish the computer. So when the computer cheated them, they were actually willing to *sacrifice their own money to take money away from a computer program!*

PIZARRO: So sort of ramping up of the social cues. I think maybe one mistake that we make is well, let's make our robots more social, right? And in reality, a terminal might just be exactly what we want.

DENNETT: Many, many years ago Omar Khayyam Moore, at Yale—early pioneer in computer-aided instruction—really went on the warpath against phony anthropomorphization of programs. In those days it was you typed your name, and then it said, "Well, Johnny…" and he said get rid of all of that; get rid of that because you're squandering one of the great things about computers, mainly you're in the privacy of your own room, and there isn't anybody looking over your shoulder. There isn't any other agent in the picture.

Now, it seems to me that there are more positive steps, as recommended by O.K. Moore that might be considered. Since, when people adopt the intentional stance they invariably over-interpret, they always are charitable, they always interpret more understanding than is there. I mean that's just clear. But it might be good if we

deliberately built in self-exposing weaknesses and foibles so that when you start using a new app or something, you are taken through some things that it screws up on—It can't do this. It can't do that. It can't do that—so that you sort of unmask the thing before people start over-interpreting it. It might be a good idea.

~~~
~~~

Think for a moment about a termite colony or an ant colony—amazingly competent in many ways, we can do all sorts of things, treat the whole entity as a sort of cognitive agent and it accomplishes all sorts of quite impressive behavior. But if I ask you, "What is it like to be a termite colony?" most people would say, "It's not like anything." Well, now let's look at a brain, let's look at a human brain—100 billion neurons, roughly speaking, and each one of them is dumber than a termite and they're all sort of semi-independent. If you stop and think about it, they're all direct descendants of free-swimming unicellular organisms that fended for themselves for a billion years on their own. There's a lot of competence, a lot of can-do in their background, in their ancestry. Now they're trapped in the skull and they may well have agendas of their own; they have competences of their own, no two are alike. Now the question is, how is a brain inside a head any more integrated, any more capable of there being something that it's like to be that than a termite colony? What can we do with our brains that the termite colony couldn't do or maybe that many animals couldn't do?

DANIEL C. DENNETT:
The De-Darwinizing of Cultural Change

I deliberately asked to go last and I deliberately didn't plan a talk because I wanted to see what people were going to say so that I could sort of riff off that, and that's what I'm going to do. It's been a good choice because, boy, what a nice day full of ideas. I wouldn't be able to do this ifwe hadn't had all these talks before, which has sort of paved the way very nicely for it.

Think for a moment about a termite colony or an ant colony—amazingly competent in many ways, we can do all sorts of things, treat the whole entity as a sort of cognitive agent and it accomplishes all sorts of quite impressive behavior. But if I ask you, "What is it like to be a termite colony?" most people would say, "It's not like anything." Well, now let's look at a brain, let's look at a human brain—100 billion neurons, roughly speaking, and each one of them is dumber than a termite and they're all sort of semi-independent. If you stop and think about it, they're all direct descendants of free-swimming unicellular organisms that fended for themselves for a billion years on their own. There's a lot of competence, a lot of can-do in their background, in their ancestry. Now they're trapped in the skull and they may well have agendas of their own; they have competences of their own, no two are alike. Now the question is, how is a brain inside a head any more integrated, any more capable of there being something that it's like to be that than a termite colony? What can we do with our brains that the termite colony couldn't do or maybe that many animals couldn't do?

It seems to me that we do actually know some of the answer, and it has to do with mainly what Fiery Cushman was talking about—it's the importance of the cultural niche and the cognitive niche, and in particular I would say you couldn't have the cognitive niche without the cultural niche because it depends on the cultural niche.

What I'm working on these days is to try to figure out—in a very speculative way, but as anchored as I can to whatever people think they know right now about the relevant fields—how culture could prune, tame, organize, structure brains to make language possible and then to make higher cognition (than reason, and so forth) possible on top of that. If you ask the chicken-egg question—which came first—did we first get real smart so that now we could have culture? Or did we get culture and that enabled us to become smart? The answer to that is yes, it's both, it's a co-evolutionary process.

What particularly interests me about that is I am now thinking about culture and its role in creating the human mind as a process, which begins very Darwinian and becomes less Darwinian as time goes by. This is the de-Darwinizing of cultural change in the world.

Our ancestors, at the very early days of proto-language and language, were pretty clueless and they were not adopting language because they could see what it was good for; it was a sort of invasion. But once they had these words, this gave them competences they didn't have before and they began to be able to do things that they couldn't do before. Their brains became structured in ways that brains never were structured before. And so what you see is that, instead of thinking of human culture the way the people in the traditional social sciences and in the humanities want to see it—as high culture, art, and science, and religion, and literature—they're treating all of these as treasures that we bestow on our descendants and that we maintain, and that we preserve, and we have reasons for this.

Many of the trends and stabilities and patterns we see in the world of human culture are well-explained by a sort of economic model. We value these things, we treasure them, we trade them, we buy them, we sell them, we put money and time into maintaining them, and so forth and so on; that exists, that level exists. But that's just the most recent period of cultural history. It wasn't like that when our ancestors were first beginning to get the benefits of human culture.

If you ask the chicken-egg question—which came first—did we first get real smart so that now we could have culture? Or did we get culture and that enabled us to become smart? The answer to that is yes, it's both, it's a co-evolutionary process.

Now, if you look at it this way, then one of the nice things of this is that it means that I can still cling to one of my favorite ideas—the idea of a meme—and say where the meme's eye point of view really works, and really when it is needed is in the early days. The best example of memes are words. Words are memes that can be pronounced; that's their genus and species. Words came into existence not because they were invented, and languages came into existence not because they were designed by intelligent human designers, but they are brilliantly designed and they're designed by cultural evolution in the same way that a bird's wing and the eye of the eagle are designed by genetic evolution. You can't explain human competence all in terms of genetic evolution. You need cultural evolution as well, and that cultural evolution is profoundly Darwinian in the early days. And as time has passed, it has become more and more non-Darwinian.

I have an example that I use when I'm writing about this, well, two examples: One is Turing's computer. If there ever was a top-down design, that's it. I mean, they would not have given him the money to build the Manchester Computer if he didn't have proof of concept and drawings. This was the idea, the understanding preceding the physical reality. Just the opposite of, say, a termite colony, which is bottom-up designed, and although it's brilliantly designed, it's a product of little entities that are themselves non-comprehending but very competent in very limited ways.

What we want to think of is a space. Peter Godfrey Smith's book has wonderful diagrams with three-dimensional spaces, if I could use a diagram I'd have it up. If you think of this cube and in the lower left hand corner we have the early days culture, which is profoundly Darwinian, and that means very little understanding, really no understanding required at all, a very broad search space. It's a lot of randomness, a lot of trial and error, very little intelligence or comprehension coming in anywhere, and where order is local, not global. What do we put up in the back right hand corner? We put Turing, Gaudi, Einstein, and Picasso. There's a phrase of Picasso's, which I love to use because it perfectly epitomizes that extreme. What he famously said was, "Je ne cherche pas je touve"—"I don't search, I find." A perfect expression of the hyper-genius. Bingo, he just goes right to the optimal solution, just like that. No grubby trial and error, no messing around. He's just the perfect genius who goes right in for the kill. Baloney. He was bragging. It was a brilliant brag, absolutely not true of him. It's not true of anybody.

If we have Picasso up in that corner along with Turing, and Shakespeare, and Einstein, and those people, and truly Darwinian evolution—cultural evolution, not genetic evolution—down in the lower

left hand group, what's in the middle? That's where most of our lives exist, in what I like to think of as foible manipulation. We are imperfect, kludgy, semi-comprehending agents who are both in cooperation and in competition, exploiting the flaws that we discover in each other's kludgy operation.

Well, think of this. Have you ever played in a chess game where you made a move and only later realized what a smart move it was but not admit it? You were just lucky but you've got a great rationalization for it later. I think that phenomenon is actually ubiquitous. A great deal of the very well-designed behavior that we engage in, we only think we understand, we only think we have to understand. We, in fact, have only a very limited understanding of it and don't need to have the understanding that tradition would say we have.

How do you install a meme? Well, the first time the kid hears it, it's just a sound. The second time the kid hears it, it's a somewhat familiar sound and maybe there's something about the context that's the same. The third time the kid hears it, a little bit more. Pretty soon, by a process of gradual installation, a structure gets established, a little tiny micro-habit in the brain, which is then available to be exploited in various ways and, of course, not always well.

There's this humanistic tradition, which says the godlike mind is required to explain the adroitness with which we get through the world and the wonderfulness of our institutions; no. The wonderfulness of our institutions can be, to a surprising degree, explained with the same Darwinian mechanisms that we explain the wonderfulness of the design of organisms. Mainly, there's a long, long history of trial and error and the features that we have today in culture that we prize are not the products of human genius, they're products of a Darwinian trial and error process going on in culture over the years, which we like to think we invented ourselves but we are more the beneficiaries than the creators of those structures.

Apply that then to the human mind. That's what the human mind is; it's an organization, which is not just evolved genetically. In bringing up a child in a social world, what you're basically doing is installing thousands of apps and meta apps, and apps on top of apps on the hardware of the brain, which is profoundly unlike the hardware of your iPhone, because it's made of all those billions of obstreperous neurons. The trick is to see how the installation of cultural apps on this hardware takes place. But we know something about that, too.

I don't know if you've seen Deb Roy's spectacular work on the Human Speechome Project where he got this tremendous dataset of his own son's learning language. What he can now answer is, on average, how many times had a particular word of English been spoken in Davin's presence before Davin started trying to say the word. It's not very many, it's about five. And many interesting patterns emerged and are beginning to emerge from that dataset. But if you think about it, remember, a word is a meme. How do you install a meme? Well, the first time the kid hears it, it's just a sound. The second time the kid hears it, it's a somewhat familiar sound and maybe there's something about the context that's the same. The third time the kid hears it, a little bit more. Pretty soon, by a process of gradual installation, a structure gets established, a little tiny micro-habit in the brain, which is then available to be exploited in various ways and, of course, not always well.

Somebody mentioned the Stroop Test earlier today. A Stroop Test is a perfect example of where you have all these apps in your head that you can't turn off and they are firing; you're reading the words, you're saying the words, you can't prevent this app from firing, and normally it's a good thing that it fires but here's a case where there's interference.

Fiery's talk discussed the relationship between sort of controlled cognition and habit and the rude associations and so forth. He was on the right track there. He was talking about what the high level controlling mind is—a patchwork of kludges made up of the exploitations of the underlying habits, some of which are genetically encoded but most of which are themselves acquired by basically Pavlovian mechanisms early in childhood. I'll stop there.

MULLAINATHAN: I'm struggling to understand the interaction between the culture component and the things that are more ... I don't know what to call it ... more intrinsic to the mind. An example that's going through my head is a fact from child language development, which is that if you look at, say, pluralizations. Pluralizations that are awkward, like children rather than "childs." What you see is, that early on they get it right.

DENNETT: Then they go through a period when they've discovered some system and now they're over-systematizing it.

CHRISTAKIS: Right. And then they say, "childs" for a while until they learn the exceptions again.

DENNETT: "Holded" and things like that. Right. And then they get back.

CHRISTAKIS: And that dip is intriguing because it feels like the meme of the word, the copycatting feels good for the first part. Perhaps some adult said to them, "Hey, this is how you form it," but it feels more like, at least my understanding of literature, is that's a rule that's just inducted by the child at some primitive level, which is an interesting interface between while the culture is giving some things, the brain and all of our brains are discovering this other thing. I'm trying to see how you think of those kind of interfaces. It feels like not everything can be coming from the outside.

DENNETT: I'm inclined to be a real renegade about this and say that the innate components of the Language Acquisition Device are hugely overestimated and that what's really happening is that, for reasons we don't yet understand, there really are optimal solutions to certain sorts of communicative problems that are discovered by exploration.

Take the Nicaraguan sign language, where you have this brand new language, which gets more grammatical in a very short period of time. I don't think that shows us much about what's genetically encoded. I think it shows what happens when what's genetically encoded is an intense desire to communicate, and a circumstance where you have a bunch of kids running around who have a lot of time on their hands, they're playing and having fun and they're just being kids and they're very plastic, very labile, and they're just exploring the heck out of the possibilities, and they're turning their home sign into language and patterns are being created in front of your eyes.

And the fact that the patterns look very much like the patterns of actual natural languages doesn't speak to the fact that there's an innate mechanism that's biased that way, but speaks to the fact that it turns out that there's some reason why this is the efficient way of doing language. I've put it much too simple. But one ought to resist the innateness view more than is often the case because, after all, if you pass the buck to the biology and say, "Well, that's innate, " then we want to know how did the innate structures get genetically implanted in the brain? And there'd better be some answer in terms of how our ancestors' early explorations with language drove them into these patterns.

CHRISTAKIS: With birdsong we have good examples, right? There's some birdsong that is innate—the birds will produce the song even if they're never exposed to it, and others that are learned in different, often similar species. There's an example where we can actually distinguish, unlike in humans, between the two possibilities and find evidence for both. I suspect it's probably similar in humans with the

case of language. There are definitely going to be features of the mind that are encoded.

DENNETT: Oh, sure. Of course there are. But I would say a lot less than has been advertised.

~~~
~~~