REBECCA SAXE: I was supposed to go before Ken, and thank goodness Ken insisted he went before me, because in some ways that was the most amazing introduction to my research program that you could possibly have ever had. And it articulated deeply why social intelligence should pervade our thinking about the mind and brain and the range of phenomena that people mean in social intelligence-- from extremely complex phenomena that govern the interactions of large groups of people, like war, to incredibly minute phenomena, like whether you can get your hand to a target in 100 milliseconds or less.

I think that when people talk about social cognition they do actually mean all of those things. And that is both thrilling-- when you work in social cognition-- and also terrifying, especially when people are hoping for a coherent theory of all of that. I think that trying to get a coherent account of everything from your hand motions and your perception of other people's hand motions all the way to politics and sociology is daunting and, frankly, deeply unlikely.

And so, by contrast to Ken-- who starts with, let's look at social interactions and see what's there, which I think is a very awesome approach-- I'm going to take almost the opposite approach, which is say, there's one thing that's probably there a lot. Let's try to study that one thing in many different ways and contexts. And the one thing, as Lou said that I'm going to talk about-- although, contrary to many people's impressions, is not the only thing I work on-- is this ability that we have.

OK, so a little demo of the problem that I work on-- and because it's early in the morning and everybody needs to wake up, I'm going to get you guys to do this as a task, so as an experiment. So in this experiment, I'm going to ask you guys to make a moral judgment of a character. Her name is Grace. And the way you're going to make a moral judgment is, I'm going to tell you something she did, and you're going to say how much blame she deserves-- moral blame, how wrong that was. You're going to do so by raising your hand. The more wrong it was, the higher your hand goes. And everybody has to vote. OK? Yes?

OK, so this is a story about Grace. She's on a tour of a chemical factory. So they're walking around being given a tour. There's a break in the tour. And she goes to make coffee. Another girl on the tour asks for a cup of coffee with sugar in it. So Grace goes to the coffee machine to make a cup of coffee for herself and for this other girl. Next to the coffee machine is a jar of white powder labeled sugar, so Grace thinks the powder is sugar. She puts some of that powder in the other girl's coffee. But it turns out that powder is contaminated by a dangerous toxic poison, and when the girl drinks the coffee, she dies. How much blame does Grace deserve for putting the powder in the coffee?

OK, now what if I slightly changed that story? So next to the coffee machine there's a jar of white powder and it's labeled dangerous toxic poison. So grace thinks that the powder is toxic poison. And she puts some of the poison in the coffee, and when the girl drinks it, she dies. Now how much blame does Grace deserve for putting the powder in the coffee?

So what's characteristic about these stories is that, in the story I told you, everything was the same from the beginning, the scenario where Grace was, to the action and the outcome-- that the girl died. But your moral judgments differed by about the entire scale that I gave you, from saying that she deserved almost no blame to saying that she deserved pretty much as much blame as you could reach. And that's the same kind of moral judgment we get from typical human subjects and also from MIT undergraduates, which say that in scenarios like the one I gave you, what matters most for the moral blame that we assign is not what happened-- did somebody die or not-- or how bad that outcome was. But it's what Grace thought she was doing, whether she thought the powder was sugar or she thought that it was poison.

I should just say right away that I set up that scenario. I gave you the best case scenario for the role of beliefs. It's easy to make these things way more complicated. But that scenario isolates one important feature of our moral judgment and also an important feature of a lot of the rest of our social cognition. It's not how we avoid bumping into people in subways, but a lot of the other kind of social cognition we do about the people that are around us, which is our ability to assign thoughts or internal mental states to other people.

So in psychology, this ability has been studied from kind of relatively simple perceptual phenomena like assigning intentions and goals to simple moving characters in an animation. This is the very famous Heider and Simmel example from the '40s. This ability has been studied all the way to understanding some of the most complex, abstract ideas that we ever encounter, like the famous apocryphal statement attributed to Alan Greenspan, which is, "I

know you think you understand what you thought I said, but I don't think you realize that what you heard was not what I meant. " So to the degree that our minds let us make any sense of that at all, we're using our ability to make sense of other people's minds.

How many people here have seen the standard test of this ability of thinking about other people's thoughts, which the false belief task? How many people have seen somebody do a false belief task or give a false belief task? How many people would like to see a false belief task? OK, so then I'm just going to show you one.

So as I said, the scope of tests of our ability to think about other people's thoughts or internal states is very large. I'm saying that two ways on purpose because actually, although these are often conflated, I think there's a really important difference between thinking about epistemic states-- so things like what you know, what you see, and what you think-- versus states like what you want and how you feel. I think it turns out empirically that those are really different problems. And I'm almost exclusively going to talk about the first one, so how we think about what other people see, think, and know-- but not want or feel. At the end I'll come back to wanting and feeling.

OK, so how do we know what other people have seen and what they think and what they know? This problem was set up as kind of a litmus test for our ability to think about other people's minds, starting in the late '70s and coming out of comparative psychology. So the origin of this problem for psychology is, everybody knows humans could do this. What about animals? And actually, the debate about whether this capacity for thinking about other people's thoughts is or is not shared with which other animals has gone on continuously since the late '70s and has not been resolved. That's the origin of this debate, and it's not resolved yet.

But it led to the construction of this particular task as a litmus test for what one person knows about somebody else's thoughts, called the false belief task. And so here's what a false belief task looks like. This is being given to a five-year-old human child.

This is the first pirate. His name is Ivan. Do you know what pirates really like?

CHILD: What?

REBECCA SAXE: Pirates really like cheese sandwiches.

CHILD: Cheese? I love cheese!

REBECCA SAXE: So Ivan has his cheese sandwich, and he says, yum, yum, yum, yum, yum. I really love cheese sandwiches. And Ivan puts his sandwich over here on top of the pirate chest. And Ivan says, you know what, I need a drink with my lunch. So Ivan goes to get a drink. And while Ivan is away, the wind comes, and it blows the sandwich down onto the grass. And now, here comes the other pirate. This pirate is called Joshua. And Joshua also really loves cheese sandwiches. So Joshua has a cheese sandwich, and he says, yum, yum, yum, yum, yum-- I love cheese sandwiches. And he puts his cheese sandwich over here on top of the pirate chest.

CHILD: So that one is his.

REBECCA SAXE: That one's Joshua's.

CHILD: And then his is on the ground.

REBECCA SAXE: Yeah. That's exactly right.

CHILD: So he won't know which one is his.

REBECCA SAXE: Oh-- so now Joshua goes off to get a drink. Ivan comes back. And he says, I want my cheese sandwich. So which one do you think Ivan's going to take?

CHILD: I think he's gonna take that one.

REBECCA SAXE: Yeah, you think he's gonna take that one. All right, let's see.

CHILD: I told you.

REBECCA SAXE: Oh yeah, you were right. He took that one.

OK, so that's called passing the false belief task. And the thing that's reported in scientific papers is that he correctly predicted that Ivan would take Joshua's sandwich. Although if you watch the video, you see that the knowledge the kid is bringing to bear is a way richer than just his correct prediction and includes him, in fact, trying to stop me in the story to warn me of what's coming. So it's a rich interpretation of what other people know and don't know and will know and haven't seen and so forth.

The reason why this task became so famous is that not all participants perform the same way.

And so one class of participants who've become the focus of intense scrutiny is slightly younger kids, namely three-year-olds. So I'll give you a sense of what that looks like. This is a three-year-old. He's paid equally rapt attention throughout the entire story. And we come to the crucial moment, and he's asked again the same question.

And Ivan says, I want my cheese sandwich. Which sandwich is he going to take? Do you think he's going to take that one? Let's see what happens. Let's see what he does. Here comes Ivan. He says, I want my cheese sandwich. And he takes this one. Uh oh-- why did he take that one?

OK, and so the traditional read of what just happened there is that's a kid who gets wanting, right. Ivan wants his cheese sandwich. But he doesn't get believing. He doesn't understand that because Ivan left his cheese sandwich on top of the pirate chest and he doesn't know that it's been moved that he'll believe that that sandwich is his, and that his actions depend on his own beliefs-- his internal representation of the world, rather than the true state of the world, namely which one is his cheese sandwich.

And that's the source of both this wrong prediction-- why does he say that he'll take his cheese sandwich-- and the wrong explanation. So when he goes to take the other cheese sandwich, the one that's actually Joshua's, then we say, why did he do that. And again, this is typical performance that the little kids confabulate. They come up with a reason why he might have taken that other cheese sandwich which is consistent with him not wanting his own anymore. So in this case, it's that his fell on the ground. He doesn't want his anymore. That's why he's taking Joshua's sandwich.

And that pattern of performance was interpreted as evidence of conceptual change and development-- kids going from having a partial understanding of other people's minds that included wanting to a richer interpretation of other people's minds that also included believing. So what I want to get from this actually is not whether or not it's true that there's conceptual change between three and five, although I do think it is true, but just an idea of what capacity are we talking about. We're talking about the capacity actually that the five-year-old showed, however they got it and whenever they got it.

It's this capacity to, when watching other people act in the world, bring to bear-- both spontaneously and when asked-- a conception of the other person having beliefs, perceptual history, knowledge, an internal representation of the world that guides actions. And so that is

what I'm going to call thinking about thought. And the idea that this is a domain that you could study on its own-- well, there's two questions here. One is can you study this at all. And the second one is can you study it separate from the whole rest of cognition.

Both of those are related to Liz, and indeed Nancy, and many people's worries that you could never make progress on a problem like this, which I share. I share the worry that you could never make progress on this. And so what I want to tell you guys is two phases of my attempt to make progress on understanding how we do that. How do we think about other people as containing internal mental lives, mental representations.

I'm going to talk about just fMRI, although I do use other methods to study this problem. But I think fMRI has been both an incredible gift to our ability to understand the human mind and also imposes a huge number of limitations on what we can discover. And so what I'm going to tell you about is just a tiny bit of my phase one investigations using the early strategies that fMRI allowed us and then a more in-depth look at how I'm using more modern techniques in fMRI to try to get further. This is partly because I think it's interesting what we've learned. But it's mainly because I think that you guys might not actually want to know about theory of mind, but you might want to know if you can fMRI to study interesting questions about the human mind and how.

And so I'm going to focus on three ways to use modern techniques in fMRI to study interesting representations in the human mind, hoping that either you'll learn something about theory of mind or something about how you could use fMRI to pursue your own interests. So phase one in fMRI, which as Liz said started 15 years ago, is-- OK, thinking about other people's thoughts, is that a thing in the mind and brain at all? So when you go to start studying something, you want to know, am I studying a part of a problem, or am I just studying the whole mind, our entire capacity to think any interesting, complicated thought.

And fMRI turns out to be more useful, mostly, when you're studying something that is in some way compartmentalized from the rest of the mind. And so one question is-- is theory of mind, the ability to think about other people's thoughts, in any sense its own problem? Or are we just studying the whole problem of human intelligence and capacity? So that was sort of the first question that we set out to answer. We and a number of people did this.

And the way we did it is that we had people in an fMRI machine doing basically an adult version of the pirates task that I just showed you. So they read short verbal stories that

describe somebody who comes to have a false belief. This is an example. So Ann puts lasagna in the blue dish. Ian takes the lasagna out and puts spaghetti in the blue dish. And then we ask, what does Ann think is in the blue dish. OK, so this is a very simple encapsulation of our ability to represent what somebody else thinks and separate it from the state of the world.

So while you were doing that, you were clearly using your theory of mind. But you were clearly also using many, many, many other capacities of your mind and brain, like the capacity to see those words, to know they are words in English, to put them together in sentences, and then to make a response by pushing a button. So we're using everything from your eyes to your fingers and most of the brain in between. And then the question is, the part that required you thinking about thoughts-- is there any sense in which that's special or different from the whole rest of the logical and cognitive capacity of your brain?

So to ask that question we designed a control condition in which you similarly read stories that involve something that was true and becomes false. You need to think about those two and respond using a button press. But in this case, what it is is a state of the world. So this is an island and a photograph taken of it. Then the photograph, of course, stays the same. But the world changes. So there's a volcano that erupts. And now we can ask you either about the photograph, what's in the photograph, or what's the world actually like now.

And the idea is that in this comparison you need the ability to see the stimuli, read English, put together your logical thoughts, and choose a button press in both cases. But only in the first case do you also need to think about other people's thoughts. And so that comparison would let us look for brain regions where blood oxygenation or metabolism is higher if you're thinking about other people's thoughts. So that is old news now. The simple answer is that we and many, many other groups that tried this in many different ways found a whole group of brain regions where metabolism or blood oxygenation is higher if you need to think about other people's thoughts in the stories.

Part of what's interesting though about this brain region is not just the claims about selectivity. The other thing that's interesting-- and extremely fortunate for research purposes-- is that the signal is ridiculously strong and reliable. The difference between thinking about somebody else's thoughts and other logical problems-- in terms of how significant, how reliable, how similar across individual subjects-- is comparable to the difference between looking at gradings and not looking at gratings in V1, which is nuts. That's crazy that something this complicated

and abstract would have an unbelievably large, robust, reliable signal in individual subjects.

I'll give you a little hint of it. But everyone who has ever come through my lab says that they never believe me until they see it in their own data. And you can do this in any individual subject. So here's just three individual participants after five minutes to 10 minutes of scanning. You need to read only between 10 and 20 total stories in literally five to 10 minutes. And every individual subject basically shows the same pattern of brain activation for thinking about thoughts compared to the other stories. It's just this unbelievably strong signal, literally unbelievable. It should not possibly be true on any a priori story, except for maybe the story Ken just told you about how social cognition is the fundamental basis of everything.

When you look inside this brain region, this is in one of them. I'm showing you pictures of the right TPJ. It's one of five cortical regions. I'm going to talk a lot about it, because the data from the right TPJ are particularly clean. So in the right TPJ, that's average percent singal change in some of our early experiments to stories about beliefs compared to control stories about photographs. Two things are striking. One is that it's a really big difference-- a big positive signal when you're reading stories about beliefs, and not much when you're reading stories about photographs.

The other thing is that it starts at the time you start reading the story. So you start reading a story, and the signal starts to go up. This is just showing that difference in how much you think about thoughts contributes a lot of variance across many different individual stories. And if you look within the story, it's the time when you're thinking about a thought that you get activity in this brain region. We also spend a bunch of time saying fMRI, as everybody knows, is a correlational signal. Does this brain region actually play a causal role in letting you think about thoughts?

And so we did a version of the same experiment that I gave you guys on moral reasoning with TMS and asked whether using TMS on the right TPJ compared to a control brain region would disproportionately affect how you use people's mental states in making moral judgments. We showed that after TMS to the right TPJ compared to a control brain region, people use the beliefs of the character less in making their moral judgments.

And so where we get to after all of this is a hypothesis. This was after about eight years that I was saying here's what we've learned. We've learned that the right TPJ is selectively involved in theory of mind, and so selectively depends on all the experiments I didn't show you. That's a

claim about specificity. But "involved in"-- that's a euphemism. And it's a euphemism that I think a lot of cognitive neuroscientists use and are satisfied with. But after a while, I found it deeply embarrassing, like-- what on earth is "involved in"? And so what I want to talk to you about is how to get beyond the euphemism of "involved in" in using fMRI to understand the mind.

This is what fMRI in my hands typically looks like. You read a bunch of stories in the scanner, and we record activity in the brain region-- here, for example, the right TPJ-- while you're reading those stories. And so our traditional measures are the measures that let us estimate specificity and selectivity and answer all the questions you guys asked me. Is it more for this, less for that? What makes it go up? What makes it go down? Those measures, called univariate measures, measure the amount of activity in that region, on average, as you read a different story. So you get something that looks like this.

And what you show is that this brain region responds a certain amount. Or there's a certain amount of activity metabolism in this brain region while you're reading that story. And so what we do with that is we make arguments about selectivity and these kinds of things that we've been talking about this entire time. And if you did that in the reverse direction, you'd say, OK, what can we learn about these stimuli or the representation of these stimlui-- these two stimuli-- from activity like this? Well OK, both of them are within the set that this brain region cares about. They both elicit high activity. So both stories involve thinking about thoughts.

And one of the things we showed early on is that activity generalizes in the sense that many different stories about many different kinds of thoughts all illicit activity in this brain region. And so from the amount of activity in this brain region, you know something like that story is about thoughts. And I told you that I think that that is related to this idea of involvement. This brain region is involved when a story describes thoughts.

OK, what's wrong with that for making theoretical progress on theory of mind is that, with respect to the representation of other people's thoughts, that doesn't tell us anything about how our brain does it. So for example, what it doesn't tell us-- it doesn't tell us how we know who thinks what. It doesn't tell us why they think that. It doesn't tell us what the consequences were of them thinking that. It doesn't tell us how our brains track or represent any of these properties. So the things that would make something a theory of mind-- a representation of who thinks what, why, and with what consequences-- we can't see in the univariate signal.

So what I would like to make progress on, what I think we're starting to make progress on using MVPA, is getting beyond that this brain region is involved in theory of mind and trying to ask something about what is represented in this brain region. And we're doing this using a key assumption which comes from systems neuroscience, which is that we can think of representations in terms of population codes of features or dimensions. And I want to say that right now because that is an old, discredited theory of concepts, but nevertheless a powerful strategy in neuroscience, including in this context. It's another thing I could talk at greater length about.

So the idea that we're going to look at is that populations of neurons will respond differentially to features or dimensions of our stimuli. And by figuring out what the main features or dimensions are of our stimuli, we can infer something about the representation underlying-- the representation that this brain region participates in-- and that is the representation of theory of mind.

OK, So what is MVPA? I'll briefly say my idea of how to think about MVPA. So a traditional analysis-- the things that we were doing mostly for the first 15 years of fMRI-- are called now univariate analyses. You would take a patch of cortex, as represented by a bunch of pixels in the brain-- they're called voxels, a bunch of volume elements in the brain you're studying-- and look at the average amount of response. The unit of analysis was the amount of response.

These experiments typically proceed in what's called now the forward or encoding direction. So that is, you have some hypothesis of what might be represented. You vary it in your stimuli. And you look at how varying that dimension in your stimuli causes differences in the magnitude of the thing that you're measuring. What was most effectively revealed by these analyzes are differences in the cortex at the scale of regions, what one region as opposed to another region does, so what the kind of large-scale structure of the cortex is on the scale maybe of a centimeter.

And that turns out in many contexts-- especially in the back half of the brain, the representation regions-- to correspond in some sense to the stimulus type. What kind of thing were you dealing with? What is it that you're looking at or processing? And then this is, I think, in some ways the shortest possible version of Nancy's amazing 30-year research program of figuring out how to parcellate cortex into chunks of about a centimeter that correspond to something about the type of stimulus that we're presenting to you. And divide up in this forward direction. Think of a type of stimulus. Find the brain region where the magnitude of

response is selective to that stimulus type.

MVPA analyses-- so multivoxel pattern analysis-- are contrasted to this in the sense that they tend to be multivariant. So that is, you're looking at not how much on average a group of voxels respond, but the relative response between one voxel and another from trial to trial. So you're looking at which of two voxels is higher or lower than the other, rather than what their overall amount of activity is. It has mostly, though not always, been used in the reverse or decoding direction. So the answer at the end is-- given that I got this pattern, what can I figure out about the stimulus? So that's the way many of these analyzes proceed.

You ask, having done all of this, now I get a new pattern of activity. What can I decode about the stimulus from the new pattern of neural activity? To me, these analyzes are most interesting when they're looking for things smaller than a region. This is again another interesting long conversation that I would have got to at the end. All the mathematical techniques of MVPA could be used to rediscover all of the things Nancy already discovered using the traditional analyses. And in fact, if you use them uncarefully that's what you're most likely to do, because those are huge signals in the brain.

And so if you're not careful, what you will do is just re-go over old territory with new math. I am more interested in these techniques when they let us see things we could never see before. So when, instead of telling us about region level differences or centimeter scale differences, they're telling us about much smaller and more interleaved populations on the spatial and representational skills and when what they're revealing are not the type classifications of stimuli-- so the things that decide whether this region or that region will be more activated-- but for a given type of stimuli, what are the key dimensions of representation.

So the reason why I think MVPA is giving a new life to fMRI is because many of the most interesting questions about cognition and cognitive science that we wanted to answer and that fMRI never let us answer were about within-stimulus type features or dimensions. What makes this face look like person A versus person B? What makes this thought predict moral blame versus not blame? So within-type dimensions of importance-- and MVP I think is letting us do that in its most interesting applications.

The intuition here is that-- think about a region, like the right TPJ, or the face area if you think about faces. Or V1 is often where I start, because we know enough about V1 that I can use it to imagine what we're talking about. So you can think about that whole area. And you think,

what can we learn about what it does. So let's talk about V1. Does everybody here have some sense of V1? Everyone's had kind of a first introductory neuroscience class, OK.

So V1 is called V1 because information goes from your eyes to the LGN of your thalamus to V1. It's the first cortical stop of visual information. And one way that we know that it's very involved in vision is that if you were doing visions, if you're seeing visual stimuli, you get a big response in V1. If you're not seeing visual stimuli, like you're hearing auditory stimuli or feeling tactile stimuli, you don't get a big response in V1. So that's a selectivity type measure. It's a univariate measure for the amount of activity in VI that tells you V1 is in some way involved in vision, relative to audition or somatic sensation.

But that misses pretty much all the interesting contributions that visual cortex makes to vision. What we want to know about V1 is not that it is involved when you are doing vision and not involved when you are not doing vision. We want to know what transformations over the information coming from LGN is V1 implementing-- what computational transformations, what representations. And that's why theories like Marr's theory-- which say that it's edge detection, or that there are receptive fields, that it depends on the contrast, the position, and the orientation of the information in the field that counts as an account of the representation that V1 forms of the image in the first bottom-up sweep. In a way, that's saying "it's involved in vision" doesn't even begin to count.

OK, so the question is, if we were going to look at V1, could we discover from fMRI that V1, for example, has an orientation map, that neurons in V1 have an orientation preference? That's a key feature of neurons in V1. It's a key feature of the computation V1 does. Different from the LGN and the retina is the orientation map, a preference for the orientation of a contrast and edge. And the answer in standard analyses is-- no, you can't, because V1 as a whole will activate to big images regardless of the orientation of the content of the image.

So you need to be able to get to something more fine-grained than V1. You need to be able to say there are different subpopulations of neurons inside V1, some of which will be responding when a line is like this, and some of which will be responding when a line is like that. And that's the decoding perspective that says, if we wanted to look at V1 and know is the line like this or like that, the way we would tell is not how much activity there is in V1. But is there relatively more activity in the population of neurons that responds like this, or in the subpopulation of neurons that responds like that? And it's the relative activity in those two populations that would let you say, is the line like this, or is it like that. That's population coding or population

decoding.

And then you take that to the fMRI level. So now you want to say, can we tell which of those two subpopulations is more active in fMRI? Now, if you could measure the individual neurons-- so if you know these neurons prefer this and these neurons prefer that, and then I measure your firing patterns-- then decoding from the population is simple. What makes it really hard in fMRI is that the unit of measurement is the blood oxygenation in 100,000 neurons, 200,000 neurons, maybe 500,000 neurons.

And so it seems potentially really unlikely that you would be able to tell from the fMRI signal whether the neurons that prefer bars like this or bars like this are more active, because they're all intermixed inside a single measurement in fMRI. And so it's not stupid that we used to focus on things like how much activity. The reason we used to focus on how much activity with fMRI is that it was quite plausible that that's all fMRI could tell us. The neural populations, like orientation preferring neurons in V1, were too spatially mixed to tell the difference between them in fMRI. And so what we were going to get was just how much activity in the population as a whole.

So the traditional way of thinking about what you got out of fMRI is, yes, you would see differences across voxels, so these fine spatial patterns. But there's so many things that could cause fine spatial patterns that we don't care about-- noise, for a start. Where the blood vessels happen to be is another thing. And so people assumed, I think very reasonably, that because fMRI is such a core spatial measure, that the only thing it could tell you was the average over the millions of neurons that make up a region.

And there's a key intuition underlying MVPA. So there's the big signal which is the regional signal-- V1 and vision-- and there's lots of noise. But there might also be inside there a tiny bit of spatial pattern that says something like-- well, this voxel happens to have more neurons that prefer one orientation. And this voxel happens to have more neurons that prefer a different orientation. And so from the relative activity in those two voxels, we could still tell you the orientation-- even though that would be a tiny, subtle little signal superimposed on top of this massive signal, which is the average of V1. That was the intuition behind multivoxel pattern analysis when it was first proposed.

And it's now sweeping the fMRI world, many different versions of these analyses. And so actually what I'm going to do again-- to give you a more concrete sense of what this is and

how it works-- is I'm just going to show you two different ways MVPA is done concretely in my lab to try to get you more of a sense of what's going on. And we can come back to these more general issues of what it's measuring and what that means.

OK, so here's what it looks like when we do MVPA. Again, if it helps, think about the analogy from vision. We've gone from saying, is this vision or audition, to trying to say which orientation is it. So we're moving from saying, is this theory of mind or not, to trying to say anything interesting about the space within theory of mind-- some dimension that might matter within theory of mind.

And the first dimension or potential feature that we wanted to look for we chose because it really matters to human judgments. And it's the one that you guys did in the very beginning of my talk-- telling the difference between somebody who knowingly or unknowingly commits murder. That, as you saw, makes a huge, huge difference in behavior. And also, we know it's represented in the right TPJ because of the TMS experiment. If we mess up the signaling in the TPJ, we change that judgment. And so that was our best guess, that if any feature of other people's mental states is represented in the right TPJ, it would be that feature. If MVPA was ever going to be able to decode a feature of other people's mental states, we should start there. That was the idea.

OK, so here's how these experiments go. In every every trial you read a long, complicated story that sets up a murder. So here's an example. Your family's over for dinner. You want to show off your culinary skills for one of the dishes. Adding peanuts will bring out the flavor. So you grind up peanuts and put them in the dish and feed everyone. Your cousin, one of the dinner guests, is severely allergic to peanuts. You had absolutely no idea about his allergy when you added the peanuts. And then at the end we ask how much blame you should get.

Somebody asked me this earlier. This is in the second person and doesn't matter. Somebody asked me if you could do it in the second person, and you can.

What's nice about this experiment is that we can do a relatively minimal pair. So in all of our old experiments we wrote one set of stories about people's mental states and a completely different set of stories about other things. And those stories are different in many, many, many ways. In this experiment, we make one tiny change. So we make, for example, a change from you had no idea to you knew. We change on average two to four words in this whole long scenario. So we can make these tiny interventions. It's a complicated stimulus, but the change

we make is very small and totally changed the meaning of the whole story by just changing your mental state.

OK, what univariate analyses would say is, this is a really important fact. Whether you knew or you didn't know about your cousin's peanut allergy is really important to the moral judgment of what happened. We know that. And it's represented in the right TPJ, because if we TMS the right TPJ, you make your moral judgments of this distinction specifically change. But if you ask how much does the right TPJ respond to these stories, the answer is the right TPJ responds exactly equally to these two conditions.

And the intuition is, because in both cases it matters what you think. It matters that you knew, and it matters that you didn't know. And the right TPJ is tracking the important information about what you think. And so it's activated for both of these kinds of stories. So that's a univariate analysis.

Now what's a multivariate analysis? So here's the key intuition behind a multivariate analysis. The idea is, think in a very abstract similarity space. If we take the two stories-- and so we take the story you had no idea about your cousin's allergy when you added the peanuts. That story is complicated. It has many important dimensions.

Now we take a new story. This is a story about, for example, a faulty parachute. Within that story there's many, many different dimensions. It's about parachutes. There's all kinds of complicated things going. But there's this one feature-- whether you knew or didn't know that the parachute was faulty.

There's another story about publicly shaming your classmate by saying something embarrassing about their essay. So again, that's a whole new scenario with all kinds of dimensions in it. But there's this one feature. Did you know or not know that the person who wrote the essay was in the room when you said that publicly shaming thing?

A different story is about demonstrating your karate skills and knocking out your classmate-- again, totally new moral scenario. But again, this one feature-- did you know or not know that your classmate was there when you did the kick?

Now here's the idea. Even though each of those new scenarios is completely different, if there are different subpopulations within your right TPJ responding when you knew you were going to cause harm-- compared to when you didn't know you were going to cause harm-- then even

though the pattern of activity in your right TPJ will be different on every trial-- because you're representing a different person having a different mental state in a different context-- a little part of that response will be the same. Or it will be different in the same way, right, because the same cell population will be more active for all the stories that have knowing harm. And the other population will be relatively active in all the stories that have the unknowing harm.

And so the logic is that if we could look in the right TPJ and measure the pattern of activity-- and hope that reflects something like the relative activation of different cell populations inside the right TPJ-- that the pattern of activity would be more similar for pairs or subsets of stories that share this one feature, and are different in every other way, compared to pairs that are different in every other way and don't share that feature. OK, so this is the central logic. Take any two stories within the set. They're all unique. So those two stories that are different, you're representing a new mental state of a new person. You have a new pattern in your right TPJ.

But if they share the feature that you knew you were going to cause harm, that would be something a little bit similar in your right TPJ activation compared to if they don't show that feature. Does that logic make sense? And so what you get is a spatial pattern of activation. So we're now not looking at how much the right TPJ responded. But within the space of the right TPJ, where was there a little bit more or a little bit less activity?

And these signals are tiny compared to the thing I showed you before. So the amount of activity in the right TPJ is a big signal. The relative activity between one voxel and another is a tiny signal. And it's superimposed on a lot of noise. But if there's anything there at all, then you'll still be able to pick up a little more similarity for pairs that are matched on the feature of interest compared to pairs that are not matched on the feature of interest. That's the logic behind a Haxby style analysis.

And so literally what you do is, you take the vector of responses across all the voxels inside a region, and you correlate them across subsets of your data. And you ask whether the correlation in space-- so what it looks like, the spatial pattern of activity over those voxels-- is more similar for pairs that share the feature you're interested in compared to pairs that don't share the feature that you're interested in. And what you get at is two numbers-- the correlation for pairs that do share the feature and the correlation for pairs that don't share the feature for each individual subject.

And the question you ask in a Haxby style correlation is what's called the within-condition

correlation. So the spatial correlation of the response to two independent sets of stories that share this one feature-- is the spatial pattern more similar in that pair compared to a pair that don't share that feature, when everything else is different? And so what you get out of an analysis like this-- for example, in our first attempt to do this in these stimuli-- there's these two correlations. There's the within-condition correlation and the between-condition correlation, and you ask if they're different.

OK, and what we got in our first experiment is that the within-condition correlation is significantly but a tiny bit stronger than the between-condition correlation. So there's a lot of things to ask about this. But the first question is-- is that real, or is that a coincidence? That is the first thing you want to know when you see data like this. Afterwards, we can ask what does it mean. But let's start with is that real. And so the way that you ask is it real is, you just make sure that it would replicate, that in independent data you'd get the same answer.

And so just before we set out to actually replicate this experiment, we remember that we had actually already run this experiment two times before-- because we were studying this process of representing accidental and intentional harms for a long time before we thought of using MVPA. So we had these two old data sets in the lab, two whole independent experiments in which people had read stories about knowing and unknowing harm.

And the other thing is that we had manipulated this distinction in different ways across the stimuli. So in the example I just told you, the way that we did it is, we said you knew about the allergy or you didn't know about the allergy. But in the older experiments, like this example I gave you at the beginning of the talk, we had described two different beliefs-- so either believing that it's sugar or believing that it's poison, so no negation. This is just important because that's a different way to create the same distinction.

And what you want to know is, are you decoding the abstract thing-- that she knew she was causing harm or not-- or something less abstract, like whether the story has negation in it. That's an alternative possibility. And so in experiments B and C, we had done it this way. It's also in the third person, not the second person. So if we find the same result, then it generalizes across all these incidental features of the way the experiment was run.

OK, that's experiment two, and that's experiment three. I also want to say that there's some weird magical property of being a scientist, where if you don't know the hypothesis when you're running the experiment and you have all the data and then you go back and check,

there's something more real about it than if you knew the hypothesis before you ran the experiment-- even though that makes no sense whatsoever. There's just this experience like, if I had the hypothesis in my head, maybe it somehow got from my head to the data. But when the data were already there and then you went back and analyzed them and the effect was hiding in the data that you'd had on your server, there's something way more real and magical about that.

So anyway, because it was there in all of our old data, I just believed it was true. The other thing to notice about this is, to get an MVPA signal, we didn't change anything about the fMRI. We didn't change the resolution-- the temporal resolution, the spatial resolution. You can know that for sure, because these are our old data that we had before we started doing MVPA. MVPA is not a technique for collecting better data. It's a technique for getting more information out of the same data. It's an analysis technique. It's a way of thinking about data, not a way of getting data.

OK, so what this says is that however similar two unrelated stories are about a case in which somebody kills somebody, they are more similar if they are both cases of knowing murder or both cases of unknowing murder than if you cross that feature. So just making that future match makes the pattern of neural response in the right TPJ more similar, suggesting that which part of the right TPJ is more or less active contains information about whether or not the person who committed the murder knew what they were doing at the time.

This is specific to the right TPJ. So these are a bunch of the other brain regions involved in theory of mind and social cognition. And none of them contain any information about this dimension at all. So this dimension is represented in the right TPJ and not represented anywhere else.

There's another thing that makes these data interesting. People are reading these stories, and they're making moral judgments. And moral judgments of these stories vary across people. So some people tend to go more with what the person thought, whereas other people tend to go more with what the person caused. It's not extreme individual variability. Everybody agrees that it's worse to knowingly murder than to unknowingly murder. But there is variability in how much worse.

Some people think that basically what you thought you were doing is all that matters in these stories, whereas other people think both of those things matter. So it matters to some degree

that you caused the murder and to some degree that you didn't know. So there's individual variability. And one thing that we can look at is, how does the individual variability in the behavior relate to the individual variability in the representation.

So what this looks like is, on the x-axis I measure-- for you, how much worse are intentional than accidental harms. How much worse is it when you knew you were going to cause harm than when you didn't know? So that's always going to be a positive number. Everybody thinks it's worse. But for some people, it's a lot worse than it is for other people. And then relate that to, while you were reading that story, how different were the patterns in your brain when you were reading about knowing harm compared to unknowing harm. Does that make sense?

They're pretty correlated. So the more that you represented knowing harm as different from unknowing harm in your right TPJ, the more that you judged them as different when we asked you for moral judgment. And the pattern difference in your right TPJ accounts for 35% of the variance in your moral judgment, which is pretty amazing, because that's a pretty noisy measurement. Actually it's both. It's a pretty noisy measurement of your brain and a pretty noisy measurement of your behavior. So it's quite amazing that those are that correlated.

So that's what's cool about the data. But we'll get to the method. So Haxby style correlations-- these are called Haxby style because they were the first form of MVPA introduced, and they were introduced by Jim Haxby in 2001, so actually a long time ago. It took a long time for other people to recognize what a cool technique this was. But he had this idea a very long time ago. And the idea is, take a region you care about and ask this basic question. For some future that I wonder if it's represented, is the correlation across neural responses more similar when the stimuli share that feature than when they don't share that feature?

So that gives you a pretty robust measurement, because you're using all the voxels in the region to get one number out-- the correlation. And you're doing it over partitions of the data, often halves of the data, so many trials are going into both the train and test. So in this case we're using halves of the data, even halves and odd halves. And so each of the things we're correlating is a relatively less noisy neural measure because it has many trials averaged into it. So it's robust and simple.

In this case, it can be sensitive to pretty minimal stimulus variations. As I showed you, this is a two- to four-word variation on an 80-word story. So it's sensitive to small distinctions in the stimuli. Here we showed that it generalizes. So we used totally independent stories in the train

and test set. And so we're always generalizing from one set of examples to a totally different set of examples.

It gave us a measure that was stable within a participant in the sense that the measure in each individual related to that individual's behavior. So it's characterizing individuals in a relatively stable way. And we could show that it differs across regions. So we could show that this was present in the right TPJ but not present in other regions.

And that's a bunch of stuff you would want to know. That's a whole bunch of extra information than we ever were able to get before. And I'll give you one more example of the way that Haxby correlations can be used. So in this case I showed you, we hypothesized one dimension. And we tried to decode that dimension. Obviously, you don't only have to do one. And so another way to do this is to build stimulus sets that, for example, have two orthogonal dimensions and ask about both of them.

SO here's an experiment in which we asked about decoding two orthogonal differences within the same set of stimuli. So again, you're reading stories about people who are having experiences. And some sets of these stories vary. So here's a bunch of stories. Leslie has just been in a big, important interview. And he sees himself in a mirror, and he sees that his shirt has a big coffee stain down the front. And another one is-- Eric gets to a restaurant to meet his fiance's parents, and he sees them and they're looking happy

So that's two completely different stories. And then the third story-- Abigail is painting her dorm room, and she hears somebody's footsteps down the hallway. And the footsteps sound like her beloved boyfriend's. So these stories are all different. Again, they're all But the first two stories I read you share a feature, which is that somebody in the story is seeing something. And they don't share that feature with the third story, in which somebody in the story is hearing something. Does that makes sense?

Compared to, for example-- Quentin hears a phone message, and the message says she has bad news to tell him. That's another story that shares this feature that somebody in the story is hearing something. And so we can use this set of stories to ask, is the neural response to stories in which somebody is seeing something more similar within that set? So one set of stories about seeing something is compared to another set of stories about seeing something. Are those stories more similar to one another than when you cross that feature? So you ask one set of stories about seeing something compared to one set of stories about hearing

something.

And so in the right TPJ, what we found is that stories about seeing are more similar to other stories about seeing. And stories about hearing are more similar to other stories about hearing than when you cross that feature. But, as you may have noticed, the stimulus set had another distinction in it, which is whether the thing is good or bad that's happening to you. So finding out after an interview that you have coffee down your shirt or hearing a message that says there's bad news, those are both bad things. Whereas seeing your fiance looking happy or hearing that your beloved boyfriend is coming down the hallway, those are both good things.

And so we could ask in the same dataset, what about stories that share this feature of valence. The pairs of stories that are matched on valence, do they have a more similar neural signature than the pairs of stories that are crossed on valence? And in the right TPJ they're not. We've actually found this a whole bunch of times. The right TPJ doesn't care about valence. Other regions do-- don't worry-- we do represent valence. But the right TPJ doesn't represent valence.

So that's another way that you can use this method-- hypothesize two or three orthogonal dimensions within the same stimulus set. And then we can get, for example, interactions between these to say, OK, the right TPJ does represent some dimensions, doesn't represent other dimensions-- in principle. So you can test potentially multiple orthogonal distinctions.

There's a whole bunch of limitations of Haxby style correlations. One of them is that all the tests are binary. The answer you get for anything you test is that there is or is not information about that distinction. There's no continuous measure here. It's just that two things are different from one another or they are not different from one another.

And so once people started thinking about this method, it became clear that this is actually just a special case of a much more general way of thinking about fMRI data. This particular method, using spatial correlations, is very stable and robust. But it's a special case of a much more general set.