Rationally Speaking #193: Eric Jonas on, “Could a neuroscientist understand a microprocessor?”

Julia Galef: Welcome to Rationally Speaking, the podcast where we explore the borderlands between reason and nonsense. I'm your host, Julia Galef, and I'm here with today's guest, Eric Jonas. Eric did his Ph.D at MIT in Brain and Cognitive Sciences, and he's now a post-doc at UC Berkeley's Center for Computational Imaging.

Eric made a splash last year in the worlds of neuroscience and computer science by co-authoring a paper provocatively titled "Could a Neuroscientist Understand a Microprocessor?" So we're going to be talking about this paper today and what it implies about whether the tools used by neuroscientists are, in fact, as informative as we think they are.

Eric, welcome to the show.

Eric Jonas: Thank you, Julia.

Julia Galef: I was saying on a recent episode of the podcast that I collect this ... I've been compiling this list of papers or studies that have a particularly clever experimental design or approach. And your paper, with Konrad Kording definitely belongs in this list, it was very clever.

So, you guys basically took a number of the common tools that neuroscientists use to study the human brain, and applied those tools to a computer chip, essentially. Like studying the chip as if it were a brain, using the tools that neuroscientists use. Why did you do that? Can you walk us through the rationale behind that study?

Eric Jonas: Great. So I was in industry for a little while, and one of the things that brought me back to academia and wanting to do science again was the Obama administration had this BRAIN initiative. And they said that what we're going to do is put a lot of research money into trying to record every neuron in the brain. And to someone with a neuroscience background, that's very exciting, right? Like, it's traditionally felt like a lot of the limitations in terms of understanding how these biological systems compute are dependent upon the paucity of data that we have.

Julia Galef: Right.

Eric Jonas: But I also have this machine learning background, and I did this machine learning startup and I'm now spending most of my research days doing things that look more like traditional machine learning and data analytics, and so I was very curious about: To what degree are the answers that these kind of advanced statistical and scientific techniques that we're developing capable of giving us insights on these sorts of systems?

And there was a famous paper that ... by a cellular biologist, Yuri Lazebnik, in 2001 when I was an undergraduate a long, long time ago, called "Could a

Biologist Fix a Radio?" And he asked this question: well, biologists are doing all of this reverse engineering of these biological systems, but then at the end of the day they draw a couple of silly block diagrams and say this one influences that one. And were they to try and use a technique like this to understand a radio, they'd be hopeless.

And he really liked this analogy because of course, radios are engineered by people, and it's very clear how they work and we have all this engineering science.

And I've been kicking around this idea of trying to do a similar thought experiment for the past 15 years or so. And when people started getting really excited about the insights this kind of high throughput neural data was going to give us, it seemed like it was a really good time to actually try and make this work.

Part of the impetus, or you know, these things are kind of ... I often feel like many of the things that I do that end up working kind of start as jokes. And so this was partly a joke. Gary Marcus and Adam Marblestone had organized a CVLE workshop at NYU on ideas about cortical computation. So they assemble ... these workshops are these nice things where they assemble a bunch of famous scientists to talk about the state of the art in the field and then they bring along some young blood.

Because I know Gary, he's like, “Come along to this thing and say something provocative.” And I'm like, well, I think the most provocative thing I could say is that we're all kind of screwed. This is hopeless. No matter how much data we think we're going to magically acquire, our analytic techniques are still far from where they need to be. So that was the impetus of the project.

A group, actually based out of San Francisco, called the Visual 6502 Team had reverse engineered the microprocessor that we used, the MOS 6502, partly for retrocomputing history reasons. It's this classic chip, and they wanted a very accurate model of how it worked, and ...

Julia Galef: What was it used for, that chip?

Eric Jonas: That chip was originally used in the Atari 2600, the Commodore 64, a very early version of the chip was used in an early version of the Apple I. So it had this very exciting history, and everyone loves retrocomputing these days.

But the nice thing was that the simulator that they built simulated the processor at the right level. Instead of just doing the things that the processors would do, regardless of how it would do it, or emulating the instructions the way that many of the ... If you play Super Mario World on your Wii, or you download some emulator on your computer, those things are not trying to actually act exactly the way that the original Nintendo acted. They're just trying to make sure they get the same result. The thing on the screen looks the same way, etc.

The simulator that the 6502 team put together actually tries to be accurate down to the transistor level. So every voltage level, every state bit is identical to what you would have gotten in this original piece of hardware that you would have purchased before I was born.

Julia Galef: And that's important because you can intervene in specific parts of the chip, so to speak, in a way that you couldn't using an actual chip.

Eric Jonas: Well, exactly. So there's-

Julia Galef: Or using a more abstract version.

Eric Jonas: Right, on both sides. So on one side, it's much better to have this perfect simulator than the actual physical hardware, because the physical hardware is ... We would have effectively destroyed a lot of them in the process of making this, and you have far less access.

And then it's better to have the physically accurate simulator as opposed to this abstract version, because that's really the level I think we're trying to understand how these sorts of systems compute. We're very interested in this question of how does computation go from things happening at the single wire, single transistor, single computational unit level… up through these more complicated dynamical systems and these complicated behaviors.

Julia Galef: Zooming out a bit more to the rationale behind the experiment: was your reasoning basically, like, we know how this chip works, we built it, we understand how it works down to the transistor level, so we can sort of test the tools of neuroscience on it -- the same way that we could take ... Like let's say we had a ruler, and we weren't sure if the inch markings on the ruler were accurate or not. But we had a piece of wood and we knew exactly how long the wood was, we could measure the wood with the ruler and then thereby see, is this actually an accurate ruler? And that would give us some confidence before we used that ruler to measure other things that we don't know the length of.

Eric Jonas: I think that's fair. The thing for me, the appeal here, is that neuroscientists all ... if you ask someone who self-identifies as a neuroscientist what level of understanding are they seeking with the brain, you'll get many different answers.

To some neuroscientists the answer of understanding is about how the chemicals and neurotransmitters interact at the subcellular level between synapses. On the other hand, you have people who do FMRI studies who are interested in this whole brain activity and how that gives rise to computation. And there's literally hundreds of years of philosophical thought about what does it mean to understand how thinking systems think. And we actually ... This is very nascent, there are lots of interesting questions

that we're even having to ask today with more and more advanced computing systems.

But the nice thing about computer science and about the chip is that we can kind of throw away all those philosophical questions. We can stop asking those questions of what does it really mean to “understand” how the system works, because we feel like we have total understanding. We understand how it works every way, from the physics at the silicon level of the transistor all the way up through to the game play, and ultimately the evolutionary purpose of the processor and the video game console, which was to get parents to buy it for their kids at Christmastime.

We don't really have that level of understanding for a lot of these biological systems. In fact, if you go to a computational neuroscience conference, there's a lot of argument about, is method X actually providing understanding, or is method Y providing understanding, and there's no ground truth there.

So with the processor we can say look, it gave us this answer, algorithm X on the processor gave us this answer -- but is that answer correct based upon what we know? And then does it really have the same kind of, I guess, quality. Does it feel like the kind of understanding that we *know* one would have to have, to say something interesting about how the processor really works?

Not just does the answer it gives us correlate to something that we know to be true, but rather, do we feel that that's a combination of necessary and sufficient to describe its functionality to a level that we're happy with?

Julia Galef: I see. So you were interested not just in testing the accuracy of the tools being used by neuroscientists, like ... are the inferences we're making from these experiments correct, but also to help us interrogate what we mean by understanding a system.

Eric Jonas: I think so. For me, the question of whether or not method X gives the “correct,” in some sense, answer on system Y, is much less interesting than saying: is the answer that method X gives us actually advancing the kind of understanding we want for these sorts of computational systems?

And that's really hard to do in biology. It's really hard to do when you have something as complicated as a mouse or even a little worm, to say well, now I know that this neuron is related to this behavior. And neuroscientists or Ph. D's actually have an entire class on using weasel words like "suggests" and "strongly indicates" and "is putatively correlated with," and we become professionals in kind of talking this way.

Julia Galef: I just sprinkle the word "may" everywhere. That's my universal solution.

Eric Jonas: Exactly. Okay, great, you get an extra master's degree. No, this becomes a real challenge, and the next thing is: Again, with something as concrete as this processor, as this 40-year-old processor that every undergraduate in computer science could basically build starting with sand, we can obviate a lot of that. We can say yes, we know that this signal is correlated with that signal, but we also know that that doesn't really tell us anything about what's actually going on here. And so maybe if this method is in some sense accurate, the view that it's giving us is functionally useless.

Julia Galef: Got it. Let's talk a little more about the relationship between a chip and a brain. Clearly a microprocessor and a brain are much more similar than many other things, but they're not identical. And so it's not self-evident that we should expect the methods that we're using to try to investigate the brain to give us meaningful and accurate answers about a chip.

There's an analogy between neurons and transistors, and you have the wiring of the chip is like the connectome of the brain. But then there are some disanalogies too, like we have this sort of sharp distinction between software and hardware on a microprocessor and we don't quite have the same thing in the brain? I don't know, maybe that's a controversial statement, you're making a face.

Basically, instead of trying to explain the similarities and differences myself, I'll just ask you: how did you think about whether a chip would be similar enough, and similar enough in the relevant ways, to a brain, to make this experiment meaningful?

Eric Jonas: Great. I think the number one thing that when I talk to my neuroscientist friends who haven't read the paper, I'm often like: come on, dude, it's my paper, why didn't you read my paper? But then after we get past that awkward interaction, the first kind of visceral reaction is: yes, but brains obviously aren't chips.

And it's true there are these tremendous differences, and you highlighted some of them. The one that I always think is ... there are two actually that I think are most interesting. One is the structure of connectivity in the brain is just radically different than what we see on the processor. In the brain, neurons have thousands of inputs and outputs ... sorry, neurons have thousands of inputs and then they have a single output which then touches many different potential targets.

Whereas on chips, every transistor has three wires, two inputs and an output, or one input two outputs depending on how you structure this, but it's a much simpler system with very different dynamics, or very different structural things. They connectome just looks radically different.

But on top of that, your brain appears to be very stochastic. No single neuron appears to matter, the patterns of activity always look subtly different in a way that we call a lot of that noise even though it probably isn't. Whereas the

chip, if you start the chip up in one state and let it run, it will do exactly the same thing over and over again.

Julia Galef: It's deterministic.

Eric Jonas: It's very deterministic. And so these are very, very different structures. On the other hand, as you said, I would argue that especially for the purposes of these analyses, they're much more similar than we would think, in that the chip much like the brain shows temporal activity at multiple time scales. One of the things that not very many systems in nature show is behavior that's stereotyped at a different time scale. So you can imagine that if you look at an animal, if you look at one of us, we breathe at ... whatever, roughly 60 times per minute. When we walk, we have stereotyped walks.

But we also have stereotyped behaviors, like I move my hands in a certain way. We have all this fun language structure that we use, and then this even extends out to temporal behaviors like we go to bed at night, we wake up in the morning. All these different hierarchical levels of temporal behavior that we also see exhibited in the chip. So we see that the chip does things where it on a very short time scale tries to move the character one pixel, and then on a larger time scale it tries to change the entire frame periodically. And then at a larger time scale too, even higher, there's this score that's continually increasing or decreasing, depending upon-

Julia Galef: In the game.

Eric Jonas: In the game. It's kind of this notion that there's the temporal structure we expect to see there we believe is similar to the structure of activity in a biological system, in some sort of brain-like system. But on top of that, I think it's important to note that the techniques that we're using for neuroscience, we all call them neuroscience techniques, but we actually all stole them from electrical engineers over the past 50 to 100 years.

Julia Galef: Oh, interesting.

Eric Jonas: Everything like these dimensionality reduction methods, these spectro analysis methods to look at the frequency content -- these are all things that we poached from the El. E. departments over the past 30 to 50 years. So they're not really even neuroscience techniques, they're just general purpose techniques that neuroscientists happen to be using to try and reverse engineer these systems.

In fact, I think that might be the most powerful argument against why we have challenges ahead, for me, which is that in fact we are using a lot of techniques that are cribbed from the understanding of much simpler systems. When physicists and electrical engineers were developing these mathematical techniques, they were trying to use them to describe radio or small electronic circuits or these sorts of much simpler systems. And we use

them in neuroscience, but then they spit back some answers and we're like oh yeah, that means that oscillation is important in the brain.

Julia Galef: Wait, but if these techniques were originally developed for use in electrical engineering, then shouldn't we think ... shouldn't your a priori assumption have been they *will* give us meaningful and useful results when applied to a chip? Why then is that useful to test?

Eric Jonas: Well, right. But the interesting part about the chip is, remember, the chip is actually doing this computation. So while these tests were developed to understand small simple circuits and simple systems like radio transmission, they weren't really designed to understand how computation works. And in fact, that's this whole new thing. We really don't know how to do that even at the chip level. There was a follow-on paper to ours where some electrical engineers tried to reverse engineer a microprocessor using some new ideas from formal rule verification stuff. Basically, they were using new methods to try and reverse engineer a chip, because they were like: what an interesting question. I wonder if new technique X could even just work on the chip?

And I think that the question for us as neuroscientists is that when we're applying these techniques and we're claiming that they give us answers, are we really just doing a kind of quantitative phenomenology?

Julia Galef: What do you mean by that?

Eric Jonas: Have we really just gotten good at describing a system that's very complicated, and quantifying its behavior, in a way that we can't really trace back to either causative mechanisms or understanding? Are we very much still in this notion where you ... Ernst Rutherford had this quote where he said, "All science is physics or stamp collecting." To what degree are we still in the stamp collecting phase of neuroscience? This to me suggests that we're much more there than I would have thought we were originally.

Julia Galef: Okay, so let's take an example of a neuroscience ... well, the thing that I would have called a neuroscience technique before you corrected me ... and how you applied it to the microprocessor, and what your results looked like?

Eric Jonas: Great. Oh, you don't have a particular ...

Julia Galef: I mean, I could ask about one if you don't have one.

Eric Jonas: No, no, no. Great. One of the most classic things people have been doing in neuroscience for a long time is what I call lesion studies, where you have some biological system and you go in and you break some part of it and see what happens.

The most classic example of this I think is this patient HM. HM was a patient who suffered from severe epilepsy in the '50s, and one of the ways that you

treat epilepsy back then, and still even to this day, is you localize where the seizure starts and you remove that part of the brain.

For HM, what they did is they went in they removed this medial temporal lobe structure, because they believed that was where his seizures were originating from. And when he came out of the surgery, he had lost the ability to form new memories. And for neuroscientists this was this incredibly exciting discovery that both memory could be split up into a system that was responsible for the acquisition of new memories and a system that was responsible for the retrieval of existing or stored memories, but also that this particular brain structure was important for that.

And these sorts of lesion studies have informed a tremendous amount of neuroscience over the years. In fact, we now know that the part of the brain that was primarily removed, the hippocampus, is vital for the formation of new memories and their interaction with things like dreaming and these sorts of dreamlike states and ... Actually, I think it's the coolest part of the brain. But the problem there is that we do a lot of lesion studies in neuroscience and then say oh, this part of the brain, I removed X and saw behavioral deficit Y, therefore-

Julia Galef: X is responsible for Y.

Eric Jonas: X is responsible. Or X is ... We're of course always very careful with our language professionally. We say X is obviously “involved” somehow in something that gives rise to Y. Of course, by the time it hits the cover of Wired or whatever, it's like X is the whatever neuron.

Julia Galef: We found the Y region, yeah.

Eric Jonas: Exactly. Of course, this is extra hard in neuroscience, because as you said brains are not like computers. In fact, there are lots of parts of your brain that if you removed them and give yourself some time, you'll completely restore function. There's lots of cases of people with what would seem to be otherwise completely ... extremely traumatic brain damage recovering near full functionality, because your brain is very plastic. It has the ability to kind of adapt and change, both at a hardware as well as a software level, to continue that analogy. And of course, your computer's not like that.

We thought well, these lesion studies, though, we now have the ability to do what are very, very precise effectively lesion studies. We developed technology in the early 2000's called optogenetics. And what optogenetics lets us do is take certain types of neurons in the brain and make them sensitive to laser light. And what this then lets you do is you can say: hey look, here's an animal running around, and I can just for a brief period of time turn off these neurons and see what happens.

And this gets at some of the more traditional challenges with lesion study effects, where if I go in and I remove some part of the brain, maybe at the

time the trauma resulted in the animal not being good at this task. So optogenetics is this very powerful technique that people are using to start teasing apart how these systems happen.

But you still end up reading a lot of things in the press or even in the literature that boil down to "brain region X is responsible for behavior Y." And even if that's true, even if in some sense this particular cortical column in the medial prefrontal cortex is the thing that results in you clicking on Facebook ads or whatever ... that's research I'm trying to get Mark to fund ... even if you could show that, that's not really ... I feel like that doesn't really give you the kind of understanding that we're hoping for. And so part of what we did with the processor was we said okay, we have three games that we focus on the processor playing, Space Invaders, Donkey Kong and Pitfall.

And we can go through with the processor and we can run exactly these types of experiments for every single transistor ... that's like 3,500 transistors. I can just break that transistor individually and then see if the game can be played. And so I have three games, and I can run for 3,500 transistors, so I can basically run 10,000 experiments in an afternoon. It's every graduate student's dream.

But we can then look on the other side and say: which transistors were necessary for the playing of Donkey Kong? And when we do this, we go through and we find that about half the transistors actually are necessary for any game at all. If you break that, then just no game is played. And half the transistors if you get rid of them, it doesn't appear to have any impact on the game at all.

There's just this very small set, let's say 10% or so, that are ... less than that, 3% or so ... that are kind of video game specific. So there's this group of transistors that if you break them, you only lose the ability to play Donkey Kong. And if you were a neuroscientist you'd say, "Yes! These are the Donkey Kong transistors. This is the one that results in Mario having this aggression type impulse to fight with this ape."

And of course that's not true at all. And the reason is ... the electrical engineering reason is that there are parts of the chip that are potentially only selectively engaged in this particular video game. So maybe a particular video game has some counter that only ever counts up to, let's say, 63. And so it only ever uses those bottom six bits, and so if you break that seventh bit, you don't really realize ... it doesn't really matter. That video game doesn't care. Whereas the other one, maybe it has some counter that counts higher at some point in time.

That feels much more like the level of understanding we're going after, and so simply being able to point to a chunk of stuff, be it circuit, be it transistors, or be it neurons, and say when this is damaged you lose some functionality, that's not really the level of understanding we're going for.

Julia Galef: When I think about what would we need to understand to really understand the connection between the activity in the chip and the activity that we see on the screen, when we're playing the game… My answer would be: we'd need the program, the code that the chip is running. And, as I was sort of alluding to earlier, it doesn't seem like there's an analog for the brain. So I'm wondering --what kinds of tests *would* we need to do in order to understand how the brain works?

Eric Jonas: Great. Again, I'm going to split those into two different questions.

Julia Galef: Yeah, please. At least.

Eric Jonas: Okay. On the hardware versus software side, we certainly recognize for a chip that's as simple as the 6502, there's not a lot of what we in neuroscience would call "functional specialization". In neuroscience we think that we have very strong ... I would call them priors, but the results are 100 years of data ... so we have a strong posterior belief that the back of your brain is responsible for vision. The optic nerves project there through the LGN, and we know that V1 does all this early stage visual stuff. And so we believe that that's the part of the brain that really handles that task. Similarly, there's auditory cortex, and there are different nuclei in your thalamus that relay different signals.

We have all these strong notions that here's a task and it's done by this particular unit. Kind of like in your car. The radiator does one thing ... And people generally, when they think of computers, they don't think of them like that. You think that we have this kind of Turing Machine-esque mental model where it's this general purpose computing machine. But that's not entirely true. Even in the processor that we were looking at, it has an arithmetic logic unit that has certain circuits that just add numbers together, or certain circuits that just detect whether or not you're in a loop, or certain circuits which just get data from external memory.

So there is that level of functional localization. That doesn't necessarily mean that the software ... It doesn't necessarily mean it's like this homogenous blob of "computronium" or whatever. But moving out, more and more contemporary processors have even greater degrees of functional specialization. So if you look at something like the processor in your phone, it actually has a dedicated part of it, a small chip inside the processor itself, a small area of silicon, that just does video decoding. And it has a different one that just handles processing stuff that's coming from the camera and doing that early stage imagery processing.

And at that point, this notion of hardware and software, this clean dichotomy that we're used to starts breaking down a little bit. But still it's the case, I would argue, that all of these systems that we built do have a much stronger distinction between hardware and software that we're used to thinking about in the brain. But the fact is, we don't necessarily know what the hardware/software distinction is in areas like the cortex.

The cortex is really amazing, because it's organized into these cortical columns, so one way of thinking of the surface of your brain, when you see a picture or a diagram of a human brain it's all kind of folded up. It has all these folds. If you unfold it, you get this large area that's this laminar structure, so if you can imagine unfolding it onto a baking sheet. And when you look at that baking sheet, it's actually made like one of those seven layer bars, although it only has six layers. We're kind of ...

Julia Galef: Oh man, I love seven layer bars. You're ruining them for me.

Eric Jonas: No, I'm making them awesome. What are you talking about? Now you can think about computation.

Julia Galef: Excuse me, not all of us dream about eating brains in our spare time.

Eric Jonas: This isn't the zombie podcast? I took a wrong turn someplace. No, so across all the cortex you have this six-layer cortical structure where there's different layers of cells that have different patterns of interconnectivity, but are all kind of ... Each layer in the cortex is very similar ... I'm sorry, across all cortex, within a layer you tend to see similar patterns of activity. But then if you look down, if you look down on the brain, those layers are actually then organized into these little cortical columns. It kind of looks like a honeycomb type structure. And in there, there's dense connections between all of the cells inside a column.

And the similarity of those columns across different types of sensory cortex, across different types of motor cortex, or all this prefrontal cortex that does the thinking part of the thinking ... I guess my air quotes don't translate terribly well on a podcast.

Julia Galef: I think I heard them in there. The thinking part of the thinking.

Eric Jonas: But the interesting part there is that it may be the case that actually the physical structure of that hardware is very, very, very similar, and in fact it's the patterns of activity across those areas that give rise to different functionality. And this is strongly, especially when we start thinking about these higher order cognitive functions, executive control, emotion, these sorts of things, we know that damage to one part of the brain can be very rapidly compensated by other parts of the brain. Again suggesting that this is not simply purely a hardware type phenomenon.

We can watch functionality be regained faster than would be ... One might naively think it would be the result of the hardware reassembling itself. So, it's true that there is this distinction. I think that ... I think it's dangerous to kind of push the analogies too far. But especially for kind of critical computation, we still really don't know very much about the sorts of, in some sense, software, the kind of the patterns of activity we see running there.

Julia Galef: Or, okay, maybe a somewhat different way to ask my question would be: let's say we couldn't get access to the program that was running Donkey Kong on that chip. Would there be any way, with some technology, to reverse engineer it? Just by messing with the chip, and trying different inputs and getting different outputs.

Eric Jonas: I think so. I mean, the one way ... Imagine if you had, and this is very much a technique that neuroscientists take, which is -- imagine if I could just kind of cut some part of the chip out, like that outer unit that I was talking about earlier, the ALU. If I could remove that ALU and control its inputs and its outputs, I could probably go through and eventually say, "Look, when I put in this binary pattern, I get out that binary pattern." And go through and maybe I have to try an exponentially large number of these binary patterns, but especially for the case of an 8 bit pattern, it's not actually that large of a space. I could probably figure out, this little block here is doing addition, or what I recognize, as someone who took way too many math classes, as addition.

And that is a thing that neuroscientists use. So, one of the preps that a lot of neuroscientists use is what's called “slice”, where you basically take some slice of the brain out, and you put it in artificial cervical spinal fluid, and then you connect wires to different parts, and you put in patterns, or you do things to it, and you watch the output. As you can imagine, this is a very challenging experimental technique where graduate students are learning how to do it are the real heroes.

But the challenge with slice, of course, is that -- one, you do actually learn a lot, right? You can actually learn a tremendous number of things about how different neurons interact, and these sorts of things. But the challenge there, on the neuroscience side, the problem is that these systems are so densely interconnected that it's almost always impossible to accurately, physically isolate the relevant part. Because it's probably getting inputs from far away different parts of the brain. And so the challenge we have on the neuroscience side, is that it's very difficult to cut out a chunk and properly control inputs and outputs, right? That kind of technology is dramatically beyond anything we can imagine having today.

On the chip side, though, I do think there is a role for reductionism in science, right? These certain bits of functional modularity are important in figuring out that, "Well, I think this box does a thing, or this chunk of stuff does a thing. I'm going to control its inputs and control its outputs and try and understand what's going on," is a path forward.

Julia Galef: I'm just trying to understand what stage we're at in neuroscience. Do we know what kinds of studies or interventions, tests we would run in order to get the kind of understanding we want, if we only had the right technology? Or are we more like in a stage equivalent to the ancients before anyone had even come up with the idea of a randomized experiment -- where like the

whole concept of how would we figure that out… we didn't even have the conceptual understanding of what questions to ask or what tests to run.

Eric Jonas: Great, so, I think that I'm going to try and choose my words carefully such that I may someday ...

Julia Galef: Just add "may" every other word.

Eric Jonas: Right, exactly. Well, and not sufficiently alienate all of my colleagues that I'm not hireable. But, no, I think everyone ... I think it's true that everyone in neuroscience has some questions that they would like to ask and they feel like they are technique-limited by. I think it is also the case that there is a real gulf between the questions that some people would like to ask, or the part of the system that they're trying to ask questions about, and what would really constitute understanding for the rest of the community.

So, the interesting thing about neuroscience, is that because the system is so complicated, right, just like the electrical engineering and computer scientists, you have device physicists who build new transistors, right, people at Intel are trying to make the transistors smaller. You have computer architects, who wire them up into actual chips, and you have operating systems programmers, and you have people working at every layer of the extraction, right?

In neuroscience, similarly, we have cellular biophysicists, who study receptor dynamics, and we have cellular biophysicists who study then how the individual neurons work, and we have systems neuroscientists who study small circuits, and we have then all the way up to cognitive neuroscientists who just try and build computational cognitive models of how these systems work.

But the interesting thing about neuroscience is if you ask most neuroscientists why did they pick the level that they want to understand, they often think that it's the interesting level. And people of the lower level are studying useless details, and people above them are full of it. And the question is, to what degree are the questions that people are trying to ask at a particular level dependent upon the answers from a lower level or a higher level, right? And in neuroscience, we often don't have good computational models.

So I come from ... my PhD advisor was Josh Tenenbaum, in the computational science lab at MIT, and there was this real tradition of trying to quantify human behavior in a way that we could test specific models about what the computation was that the human was performing.

So, if you see people tend to be very good at seeing three patterns, three examples of an object and figuring out whether a fourth one is a member of that class or not, right? We have all these very quick abilities to do this sort

of induction, and often it's very benign. Often it's clear that we have priors that we're bringing to bear. And you can model all of that computationally.

And Josh's research program, and I think that the larger of goal of that kind of research program, is to understand ... I guess one way of looking at it is the questions they're trying to answer are, "What are these systems even actually doing," right? What is human cognition actually trying to achieve, and what are the functions that it's trying to optimize? Because the argument there is: before we really understand what that's doing, everything else is kind of suspect, right? If we don't really know what this system is trying to do ...

David Marr was this scientist who tragically died early, but in the early 70's, he wrote this book on vision, and he argued that there are, for systems like this that compute, there are three levels that you can imagine understanding how they compute.

One is this computational level, right, what is the task that the system is trying to solve? And do you have a good understanding, can you write down in some sense the code, or an example of the code or the goals for the system to understand it?

And then below that, there's this algorithmic level. Where, okay, let's say that maybe the goal is running, right, getting from point A to point B -- what is the actual algorithm that the system would use to do that? And there might be many different algorithms that would support this particular task.

And then below there is this implementation or hardware level, where you say, "Well, okay, let's say I had a given algorithm, how would I implement that?" Would I implement it with these types or neurons, or those types of neurons, or maybe this chip, or maybe in some other sort of substrate. And it's a real attempt to start drawing these boundaries such that we can ask these sorts of questions.

But, I think a lot of neuroscience is still kind of struggling with these top level questions, right? People who do certain types of vision science, for example, I think have a good mental model of the exact questions that they're trying to ask about. How to stimulus come in through the retina, through the LGN individual cortex to trigger this neuron or that neuron?

Julia Galef: Right, and we have made a fair amount of progress in that part of neuroscience, haven't we?

Eric Jonas: We have, I mean, vision science is ... On one hand, vision science is incredibly mature.

Julia Galef: Yeah.

Eric Jonas: On the other hand, vision science still can't really tell us how to build a computer vision system.

Julia Galef: Right, so -- that's sort of the ultimate test of understanding, is could we actually build a brain?

Eric Jonas: I really buy into this idea, exactly, can we synthesize this in hardware and software? Could we build one?

And in spite of all the recent success with deep learning and artificial neural networks, they actually work very differently from the way these biological systems work, and are still kind of substantially underpowered, right? When you see papers, or when the New York Times talks about how computer vision is now better than humans'… What computer vision's really better at is: in a scene with a single cat or a single dog, telling the difference better than humans. Right? As soon as you get into these real world situations where you have multiple objects and whatever, we're still very far from human or animal performance.

But, yeah, I think synthesis is really the goal there. And, I mean, synthesis is also ... It's a little bit of an unfair goal, because it's like, "Okay, great, we still don't have a good synthesis level model for the liver," right? Like, let's take a much simpler system in terms of computational behavior, right? It's still very hard to build an artificial liver, right?

And not just because we couldn't get it to be small enough or something similar. But because the metabolic behavior that the liver has are actually kind of these ... all those enzymes matter in different ways, right, as a function of different biological responses. But I would still say that we understand a lot about the liver in a way that we don't about the brain.

Julia Galef: Okay, I'm going to try a third way -- not that your answers haven't been ... No, sorry, your answers have been super helpful and on point! I just ... Because this is complicated, so I need to attack it from a few different angles.

Eric Jonas: Fantastic.

Julia Galef: So, another way to ask this question is, let's say we just had an arbitrarily large amount of computing power, or an arbitrarily large amount of data, or maybe both. Could we just brute force it, basically? Could we just test all possible inputs in all possible combinations until we figure out ...

I don't know, I'm imagining, to take an analogy, let's say you're someone who you just have no intuitive understanding of social dynamics and customs, and so, but you're very smart and observant ...

Eric Jonas: I can identify.

Julia Galef: I'm just saying hypothetically!

Eric Jonas: Right.

Julia Galef: And so you want to be able to function socially. So you just practice a lot and study conversations as you're practicing. And you gradually learn, like, "Oh, okay, if I smile this amount, with this frequency, and if I smile in response to these kinds of statements, that will cause the person to like me and want to see me again."

And obviously that's way too simplistic, it depends on a bunch of other things, like the kind of conversation it is, and your relationship with the person. But if you had enough practice and tried enough different strategies in different combinations with enough people, you could basically become an expert conversationalist. Despite having no *intuitive* understanding of why people like it when you smile in certain situations.

And so, I'm just saying, maybe we could do something like that for the brain, with enough compute, where we just understand how it works because we just through all the compute in the world at it.

Eric Jonas: Well, but again, this gets to this question of understanding, right? So I could… imagine if the technology existed such that I could understand how every neuron in your brain interacts with every other neuron and just clone them, but in software, right? If I could do this kind of Hansonian, create a bunch of ems, right, so the complete synthetic consciousness that we create simply by emulating everything that a human being does, but in software, right?

That's great, now I have a virtual human being in software -- but do I really understand what that system is doing? I can perfectly predict behavior, I can perfectly understand how it will respond to inputs and outputs, but I don't really feel like I, in sense, have the level of understanding, right? Just as, if someone gives me a good physics simulator like Grand Theft Auto, right, that doesn't necessarily mean that I understand physics. Even though I can, in the simulator, push the ball off the ledge, or run over the person as the case of Grand Theft Auto, and have a reasonably, physically accurate copy of what happens.

For neuroscience, I think, the challenges right now are that *both* the techniques aren't there, right, the techniques are very far away from where they would need to be… But also, and I think the thing that you're trying to get out of our paper, is that: even if you can purely observe the system -- and so much of the research effort right now is on trying to record, or observe, a large number of neurons -- the analogy in our processor is observing a large number of transistors. If you don't really have the ability to do that kind of perturbation, right, to perturb the system carefully, then all you really get out of these behavioral outputs, right? The animal's not as good at remembering, the animal turned left instead of right.

And what we're trying to argue is that even if you have the best analytic techniques, it's hard to conceive of analytic techniques that would, from observing that data, give you the level of understanding that we seek, right? And so, kind of the passive, the LHC model, let's say, of ...

Julia Galef: Sorry, Large Hadron Collider?

Eric Jonas: Large Hadron Collider one, they have a big science model of neuroscience where we're going to have these institutes, and they're going to acquire this high throughput data. And then math nerds like me ,who maybe learned social interaction via simple emulation with the other humans, will go through and figure this out. That is unlikely to bear the kind of fruit that we hope without coupling it to experiments, right?

I often describe it as, since the hypothesis space for how this computation happens is so incredibly large, then unless we can perturb the system, right, unless we can shift from primarily an observational paradigm to a one that has an interaction ...

Julia Galef: But you don't think that lesion studies count as perturbation?

Eric Jonas: They certainly do, but the level of granularity is extremely coarse, right?

Julia Galef: Okay, so: very, very fine grained perturbation.

Eric Jonas: So, there might be, there's a universe where you could start teasing apart some of these systems pretty well. And I do think we'll get there. I think we'll probably get there over the next 30-50 years, let's say.

Where if we could do single cell specific simulation, and measurement, and we could also do a reasonable job of removing this in either very simple organisms, or via something that looks like slice, removing the system to do this reverse engineering, we could make some progress.

But, just as if I didn't understand what addition was, and I tried to understand how the arithmetic unit in the processor works, and I put in all possible inputs, and I get out all possible outputs, right -- if I have no concept of addition, I'm going to look at that and say, "Well, these are the ones that go in, and these are the ones go out, great!" And then I'll get my paper in nature, or whatever, and it's fine, but it doesn't really ... until we understand what that computation is, right, that's just going to be numbers. It's just going to be this quantitative phenomenology.

And there's where I think that thinking about with what understanding means and thinking deeply about this computation level becomes really important, right? Even if I have all of the dynamics of my underlying system, if I don't really know what it's doing, I'm basically going to be curve fitting.

Julia Galef: Do you think that neuroscientists now are going in ... So, when you were doing these experiments on the chip, it was basically atheroetical. You were just sort of putting in different inputs, or doing different perturbations, and observing the outputs. But do you think that neuroscientists really are approaching their experiments with a similar degree of atheroeticalness? Or do they have a sense, at least what they think might be a sense, of what "addition is," in this case?

Eric Jonas: It really depends on the system, but no, I think there's ... In fact, I think there's a real push towards being more theory-driven, and people come up with specific theoretical models about how particular systems or sub systems work. And the closer you get to either sensory systems or motor systems, the closer you are to the inputs or the outputs where the neural activity correlates much more strongly with things we can observe, the easier that gets.

Konrad made his career ...

Julia Galef: Your co-author?

Eric Jonas: My co-author, yes -- my co-author, my wedding guest, and science BFF -- made his career on showing that a lot of motor control actions were actually like Bayes-Optimal, right, that organisms were doing the correct job of incorporating prior information to make the next decision. And in fact, there was this rigorous computational framework of Bayes statistics that kind of models how these systems are working. And, no, I think that you see a lot of that.

On the other hand, as you get more towards ... The closer it gets towards things that look more like computation -- so, critical activity, or decision making, or anything like that -- because we don't tend to have good models, what will happen is we'll record a lot of this data, we'll acquire a lot of this data, and then we throw it into some algorithm. And it spits out some numbers. It says, "Well, look, I think that there's this low dimensional structure in the data."

And Surya Ganguli at Stanford gave this nice talk at one of the computational neuroscience conferences a few years ago, where he said, "Well, look, we record this high dimensional activity," right, we record a thousand neurons, or something, "We put it in these algorithms, and we say, 'Okay, show me the two true dimensions that the system is operating on.'"

But we do that -- because the animal, while we're recording this data, is doing a two dimensional reach task. And in fact, the dimensionality of the activity that we seem to pull out when we observe this, is related to intrinsic dimensionality of the behavior. Basically, the animal, especially for areas like motor cortex, the thing that the animal is doing, is very strongly correlated, or has the same kind of structure with the activity that we're seeing within the brain.

Now, that's not surprising, but that suggests that naively recording a bunch of data and then throwing it into some algorithm, it's just going to tell us stuff we already know.

And so that was the other side of this, was trying to say, "Hey, look guys, the algorithms that we actually have right now for understanding this data are woefully incomplete," right, and if we just sit around waiting for this big data to get here in ten years, we're all going to wake up and realize that, well, maybe we have this data now, but we have another ten years of algorithm development.

Julia Galef: Is there any way to get the equivalent -- like, you're saying the areas where we've made more progress are the ones where we can sort of observe the outputs, like motor function or vision, and connect that on a pretty granular level to what's happening in the brain. Is there any way to get an equivalent of that for something like computation or decision making?

Eric Jonas: Not that we know of. I mean, all of our ... Well, right, so, if you study something like decision making, you try and reduce what the animal is doing to the simplest kind of decision making task, right, so maybe you have it run to the end of a maze, and then decide to turn left or right, right? Or there's these two, what are called two alternative forced choice tasks, where basically you make the animal pick one thing or another, and if it does it right, then it gets some sort of reward.

And so you can try and do this, right, but the problem is that the dimensionality and the behavior there are so incredibly low, that you're going to get out just a couple of bits of information back.

Julia Galef: Just like, correct or not correct? Left or right?

Eric Jonas: Well, so, right, so you know the animal did the right thing or the wrong thing, and you maybe you watched it get better at doing the right thing over time, but then you when you try to map that back to normal activity, it's a far sparser and less informative signal than tracking where every muscle activation at every joint position in my arm, or being able to control every pixel that's coming into my eye.

Julia Galef: Right. Okay, so, hopefully you can answer this without similarly worrying about annoying everyone in your field, but: Would you say that the data that we're getting back, from the results of the experiments that neuroscientists are currently doing, would you say that it is *necessary* for an eventual understanding of how the brain works, it's just far from *sufficient*? Or would you say that it's not even necessary, that we're just basically barking up the wrong tree, and we should be running different kinds of tests?

Eric Jonas: Well, I mean, the reality is that if I knew the right kind of experiments to run to actually advance our understanding of these systems, I'd just go and do those things and then collect my Nobel Prizes.

Julia Galef: Right, although, you could say, "I don't know the right ones, but I can tell you these aren't it."

Eric Jonas: Well, so, right, and one way of looking at it is -- I have to be careful, I think, of making the same mistakes that the stereotyped neuroscientist makes, where I think that everything that's at a higher level than what I'm interested in is just hand waving, and everything at a lower level, because as someone whose kind of classically city systems level neuro, where we're interested in how circuits give rise to behavior, that's the level I'm most interested in, right?

But, I'm generally skeptical of a lot of the value of fMRI work. I think that the analogy that I make there is it's like trying to understand how your computer works with a thermal imager. Good luck with that.

But, everyone I know who's working on some neuroscience question generally has a fairly refined question that they're attempting to answer, with very precise formulated hypotheses, where they're trying to carefully advance the field, right? I think, though, that when you sum up, when you integrate across all of this, we're still not going to be at the level that we've been promising Congress in terms of neuroscience insight. And it might be the case that these insights come from weird corners.

But I'm very partial to people who study much simpler systems. So people study c-elegans, this tiny little worm that has 302 neurons. We still don't really understand how its behavior works, and every single c-elegans has the same 302 neurons. You know, like, "Guys, if we can't do this ..."

There are people who do incredibly detailed, careful work of the -- this is going to sound weird -- the cluster of neurons in the lobster stomach. So, digestion is a complicated process, right? And so most animals, most vertebrates have a cluster of neurons, or a ganglion of neurons that control that entire process, and that process is very complicated, and some people study what's called the lobster stomatic gastric ganglion, which is this cluster of cells, and it's responsible for all these incredibly complicated motor patterns that has to have to happen.

And so, people like Eve Marders' group at Brandeis have ... This is a nice prep, because you can take out this from a lobster, you can basically do exactly the experiments I was talking about. You can fake inputs, you can fake outputs, and you can start understanding it, and they're at the point where they're really close to understanding how that system works.

Julia Galef: So that they can build ...

Eric Jonas: So that they could build a fake one, right?

Julia Galef: Right, nice.

Eric Jonas: And, the thought is that's a very small simple system, right -- but that kind of work, I think, points in the right direction at least?

I think that there's no place that's more interesting to be asking questions than in a lot of these higher order cognitive systems, and where is the neuron that does X, or the part of the brain that does Y. But, it's hard, especially given current technology, to -- even if I identify that this part of the brain is responsible for me clicking “likes” on Facebook, we don't really have the technology then, to start kind of opening up and experimenting on that box.

And so, that's one of the reasons why I think these lower level systems, even though they might be less sexy -- hopefully people won't kill me for saying that -- but even though it might be harder to convince Congress, let's say, to study these systems, right, periodically, whenever some Republican senator or congressman decides that (they seem to inevitably be Republican) that the NIH is wasting money, they always go and they find someone who's studying something about Drosophila, and they say, "Well, look, this person's spending ..."

Julia Galef: Flies?

Eric Jonas: Yeah, fruit flies, and they say, "Well, they're spending, the NIH's budget, we're spending 300 million dollars a year on fruit flies, this is insane." But, of course, no, fruit flies are this incredibly simple organism with 100,000 neurons, that we understand the genetics incredibly well, and we can start picking apart using genetic techniques. You can breed a million fruit flies in your living in two days.

Julia Galef: That's a horrifying image.

Eric Jonas: My house probably sounds great right now, my poor wife. But you can do all these experiments with these very simple systems, and so even though Senator X or Congressman Y, he or she might not necessarily be excited by that kind of research, it's like, "Guys, this might be the only way we have to make real progress, that isn't just painting pretty pictures," right?

Julia Galef: Like thermal imaging?

Eric Jonas: Like thermal imaging, right. I mean, it's so ... These fMRI scans are so sexy, because they ... do you have pictures of people's brains, and you can say it lights up, and everyone says that's interesting, Malcolm Gladwell writes a book about it, and it's great… but it's not really, especially from a computational side, telling you all that much, right?

Now, of course, the people who actually do this work are very careful. And the challenge for, even as neuroscientists, is that a lot of this thing gets filtered through the popular press, even into us. So I'm being incredibly unfair to my fMRI colleagues.

But the part that's interesting about these systems, I think, or I guess one thing we haven't talked about, which I think is worth mentioning, is that nothing I've talked about today, I think, necessarily has clinical implications.

Julia Galef: How so?

Eric Jonas: My hunch is, and I think that the current state of the art in neuropharmacology suggests, that many of the cognitive or brain disease problems that we see with people today are -- while there is a computational component to them, right, while ultimately they result in some sort of computational dysfunction, often that seems to be the result of very low level, or kind of molecular or genetic problems with the system.

So it may be the case that we can actually treat a very large number of diseases without ever having this level of computational understanding, right? We may, we even see that with ... I mean, we all worship at the altar of Scott Alexander, we've all read his blog posts on depression, and understand the state of the art with the research there, but also understand that even though we don't have good mental models for what's actually happening inside people, some of the pharmacological interventions are actually quite astounding.

Julia Galef: Yeah.

Eric Jonas: And we don't even necessarily understand why they work, right?

Julia Galef: Sure. Yeah, I mean my hypothetical person with no social intuition could still be extremely effective socially, despite not having any understanding of why those smiles work the way they do.

Eric Jonas: Exactly, and so many of neuropharmacological interventions that work are often discovered serendipitously, right? You give patients with Parkinson's a drug, and they happen to have less of phenotype X. And you say, "Oh, maybe that's causing X, Y, or Z." Or, we even know that diseases like Parkinson's or Multiple Scoliosis have very firm, cellular bases. These are not necessarily what we would think of as diseases of computation.

And from that perspective, I guess, on the funding side, it gives me hope that the American public will continue to fund us to understand these systems at all of these levels. Because on the high end, you have cool, pretty pictures; on the low end you have, I think, much closer, more direct path to clinical impact.

Julia Galef: Nice. Well, before I let you go, I want to ask you the question I ask all my guests, which is for a book or article or blog that has influenced your thinking in some way, what would your pick be?

Eric Jonas: What would my pick be? So I ... in getting this question ahead of time, I wracked my brain a little bit. Because of course, I have my favorite bloggers,

and heart Megan McArdle forever, but… I think the article that actually impacted me the most is ... So, I grew up in Idaho, which is a great place to grow up, but we were a very small group of nerds in my high school. And so, every month, and I was this teenager in the late 90's, kind of the height of internet one, and we'd get copies of Wired, and we would read Wired cover to cover. And, of course, as a kid in Idaho, you're like, what is this Burning Man thing they're talking about? I don't understand, who are these venture capitalists?

But there was an article in the February 1997 edition of Wired about Julian Simon. Julian Simon's a ... Well, the late Julian Simon's an economist whose most famous for kind of being this anti-Malthusian, right? He made these bets with Paul Ehrlich and the other doomsayers that, in fact, no, things are getting better. And was a strong popularizer of this kind of Solow-esque model of: technological innovation and human ingenuity give rise to wealth. They're what results in us getting past these resource catastrophes.

And that had this incredible impact on me. Because it made me, being someone who always liked building things, and liked science, and these sorts of things, made me realize that, in fact, maybe both we aren't all collectively screwed. But also, that in fact this kind of innovation was a force of good, right, in fact ...

Julia Galef: As opposed to just being intellectually satisfying, fun things?

Eric Jonas: Or a thing that's destroying the world, right? If you read *Silent Spring*, you think, "Oh my god, there were all these chemists who were doing all this work, and in fact it was turned into agent orange and these horrible pesticides and we're all going to die!"

Julia Galef: Right.

Eric Jonas: But, in fact, it seems that as countries get richer, they take environmental controls more seriously. As countries get richer, everyone starts doing better, and in fact, I feel like, even if it were held today, the Simon-Ehrlich bet may have come out differently. Because China's had this tremendous appetite for natural resources, and maybe some of the resources actually aren't cheaper today than they were say 10 or 20 years ago.

It really impacted me that both a combination of technological innovation and capitalism could have this really kind of positive impact. And set me -- my actual high school commencement speech was called, "The Antithesis of Malthus," ...

Julia Galef: Oh, man, that's so great.

Eric Jonas: And I got up there and I told everyone, "We're not screwed, actually!" And, you know, I opened quoting Malthus, and everyone's like, "Woah, he's telling us we're all going to die, this is horrible, why did we ask Eric to do this?" And

I'm like, "No, it's actually all getting better, because we're putting in this kind of effort."

Julia Galef: That's so great.

Eric Jonas: And I know that everyone, right now, it's very in-vogue to take issue with a lot of these tech companies. And I think that's probably deserved, for a whole host of societal reasons, but I think it's ... I hope that it will not along the way turn to us being opposed to technology in general, right? I mean, I think the most impressive human success story of the past say, 50 years, is the fact that China's lifted 280 million people out of rural poverty, right?

Julia Galef: Yeah, people really underestimate the decline in the fraction of the world's population living in extreme poverty.

Eric Jonas: It's amazing!

Julia Galef: It's like, one of the biggest stories of the last 20 years.

Eric Jonas: No, no, of like, of my life, right?

Julia Galef: Yeah.

Eric Jonas: I mean, I don't understand why this is not something we're ... and obviously, there are problems and there are challenges and all of these sorts of things. But I guess the piece about Julian Simon in *Wired*, it was called "The Doomslayer," really made me realize that, in fact, if you scale out your view to what's happening in your neighborhood, or your city, or your state, and in fact, you start looking at these things globally, and looking at them historically, we're doing great. And, in fact, that's because we're really, really, creative monkeys. And I think it gave me a real sense of optimism for both understanding how all of these systems work, and then trying to make them better.

Julia Galef: Well, just like with your commencement speech -- our trajectory is good, right, we started with "things are tough and maybe not as great as they seemed," and ended on a positive note.

Eric Jonas: A positive note?

Julia Galef: Yeah, that was good architecture there.

Eric Jonas: I learned it from observing other humans giving talks ...

Julia Galef: Taking careful notes?

Eric Jonas: And taking careful notes, exactly.

Julia Galef: Eric, it's been a pleasure having you on the show, thanks so much for joining us.

Eric Jonas: Thank you, Julia.

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