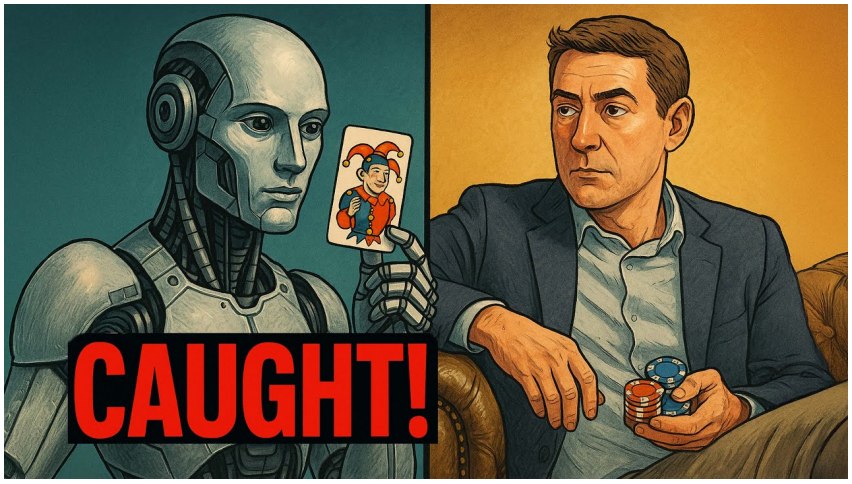
/mlst

"We need AIs with PHYSICAL experience" (Jeff Beck)

March 23, 2026



Machine Learning Street Talk

Podcast Transcript

Revision: f7f3d014

Topic headings include retention indicators: ● above average, ● average, ● below average (vs. similar YouTube videos)

Interested in licensing ReScript for your podcast? Get in touch with Tim Scarfe.

Contents

Overview 2

Bayesian brain and neural computation foundations 3

Belief-reward entanglement and AI alignment 4

Sponsor: Tufa Labs 4

Visual cortex discovery and the Hubel-Wiesel experiment 5

Theory of mind tests and ChatGPT limitations 6

The sum and product riddle: pattern recognition vs reasoning 8

Systems engineering and object decomposition 8

Vicarious experience and grounded understanding 10

Neural coding and choice of generative model 11

Line-of-sight legibility in AI reasoning 12

Language as lossy compression of cognition 13

10 bits per second: information processing bottleneck 15

Why language models cannot substitute for understanding 16

Scientific abstraction and idealization 17

Markov blankets and system partitioning 18

Scientific realism, noise, and the limits of models 20

Free energy principle as mathematical framework 21

Black-box prediction vs legible explanation 22

Reward functions and the impossibility of alignment 24

Modeling beliefs as prerequisite for value inference 25

References 27

Overview

Jeff Beck spent years studying computational neuroscience with Alexandre Pouget, Peter Latham, and Wei Ji Ma before founding Noumenal Labs. In this conversation with Tim Scarfe, he argues that language models are fundamentally limited because they manipulate symbols without the physical grounding that gives those symbols meaning.

Beck walks through the Bayesian brain hypothesis — the idea that our brains maintain probabilistic models of the world, built from direct sensory experience. Language, he says, is just a thin lossy summary of that richer internal model. He points to Markus Meister's work showing that human information output runs at roughly 10 bits per second, a tiny fraction of what comes in. This compression is why communication works at all: we share enough common ground from living in the same physical world that a few words can activate elaborate internal representations in each other.

The conversation gets interesting around the question of whether AI systems can genuinely understand anything. Beck's position is clear: until a language model produces something genuinely novel — not recombined from training data — he won't call it intelligent. He tells the story of giving ChatGPT a theory of mind test, only to have it recognize and describe the test rather than actually pass it.

The second half turns to alignment. Beck argues that you cannot separate someone's beliefs from their reward function just by observing their behavior — a claim with serious implications for AI alignment. If we cannot figure out what humans actually want (because belief and reward are entangled in observed behavior), then building AI systems that reliably act on our behalf becomes much harder than most alignment proposals assume.

He also discusses Markov blankets, the free energy principle (which he insists is a mathematical framework, not a theory), and why our intuitive physics is not wrong — it's just calibrated to the world we actually live in, rather than the vacuum-sealed thought experiments of textbook physics.

Bayesian brain and neural computation foundations

Dr. Jeff Beck 00:00:00

The problem with pattern studying pattern formation is that patterns just keep repeating. But the brain is doing something that a pattern can't do, right, which is it's processing information. And I thought and I just became absolutely fascinated with that. I studied computational neuroscience, working with Alex Pajet and Peter Latham and Wei Ji Ma.[1]

Dr. Tim Scarfe 00:00:16

You're a huge proponent of the Bayesian brain hypothesis. Yes. Why is that?

Dr. Jeff Beck 00:00:20

I believe it is it is the only right way to think about how the world works. You get some data, then you get some new data, and you sort of say, oh, how is it like the old data? And if it's similar enough, then you sort of lump them together, and then you sort and you build theories, and you properly test hypotheses in a fashion.

Dr. Tim Scarfe 00:00:36

Let's imagine we took all of the humans on the planet, and we described them using this contextual reward function that you're talking about. It seems to me that that object would be 1 of the most sort of complex artifacts in the universe. It's also the

Dr. Jeff Beck 00:00:49

case that it's very difficult for me to figure out what your reward function is. A lot of work in behavioral economics has gone into trying to figure this out. But it turned out just measuring someone's reward function, which was a prerequisite for getting an AI algorithm that shares our value, is really impossible because you cannot disentangle belief and reward based on policy. If you don't know someone's beliefs, you cannot infer their values. If you don't know their values, it's also very difficult to like disentangle their beliefs. So they disagree about a conclusion. They disagree about what action they're going to take. And if you don't want to resolve the argument, you of course instantly draw 1 of 2 possible conclusions. The person that I'm arguing with is either evil or stupid. They're stupid because their prediction engine, their beliefs are wrong, their prediction

[1] Bayesian inference in neural computation (Ma, Beck, Latham, Pouget) — paper

engine is wrong. Or they're evil because their reward function is not the same as mine.

Benjamin Crouzier 00:01:45

I'm Benjamin Crouzier. I'm studying an AI research lab called Tufa Labs. It is funded from past ventures involving machine learning. So we're a small group of highly motivated and hardworking people. And the main thread[2][3]

Belief-reward entanglement and AI alignment

Benjamin Crouzier 00:02:00

that we are going to do is trying to make model that reason effectively and long term trying to do AGI research. 1 of the big advantage is because we're early, there's going to be high freedom and high impact as someone new at Tufa Labs. You can check out positions at tufalabs.ai.

Sponsor: Tufa Labs

Dr. Jeff Beck 00:02:19

They used to do experiments. They take the cat, you anesthetize it, put an electrode in its brain, and then you show it pictures. And they initially were like, we want to find out what the neurons in the visual cortex are sensitive to. So they're showing you pictures of, like, here's a mouse, right? Here's a, you know, here's, like, other things that they think cats Here's like a tree, a plant, you know, a little food bowl, some water all these things that they thought a cat's neural or visual system should should respond robustly to.

[2]Noumenal Labs (Jeff Beck) — website
[3]Noumenal Labs research paper — paper

Visual cortex discovery and the Hubel-Wiesel experiment

Dr. Jeff Beck 00:02:45

And they found nothing, right? The neurons are just going, they're not making any noise at all. They're just not responding. And they were they were showing the cat images by just like projecting onto a screen from a slide projector, right? An old, like '19 this is the seventies, right? Or I don't know what year it was. Slide projector, where the way it works is, right, it drops a slide in, there's a light that shines through, and it projects onto a screen. And these things it's mechanical, so when it goes to switch the slide, know, pops that 1 out, and it rotates and drops a new 1 in. And they're doing this experiment, they've had not, they've gotten nothing. They have no idea what these neurons are responding to, because they're just like humming away at some low frequency. And 1 day, the slide projector breaks. And what happened is that, and I remember it was on the side, but basically, a slide ends up not, like, sliding in where it's supposed to be, but ends up, like, going in, like, cockeyed. And when that happens, what shows up on the screen is a big old white patch and a big old black patch, you know, with an angle that's associated with like the angle that the slide was at. And this happens, you know, and it happened this 1 time while they had the electrode in the CAT's brain, and all of a sudden, the neuron that they were recording from just started going crazy, just absolutely bonkers. And so what did that mean? Oh, like, that's what the neurons are actually sensitive to. Got this incredibly robust response out of the neuron when they presented this oriented thing. There'll always be problems. There'll always be mistakes. For this reason, I don't believe we'll ever be in a situation where AI will be, you know, executing plans and making decisions without any human oversight or bad. Now, it may eventually come, we may eventually get to the point where the only role, this is like everyone has their nightmarish view of the future, and it's like, oh, Terminator's gonna kill us all. My nightmarish view of the future is that humans are reduced to being, you know, a bunch of, like, complacent value function selectors, right? Their entire role, they've stopped thinking completely. They're just basically saying, Yeah, that's good. That's bad. That's my contribution to society. And I hope that doesn't happen. I hope that people continue to think of AI as something that actually enhances their understanding of the world. And I'm definitely focused on building that kind

Theory of mind tests and ChatGPT limitations

Dr. Jeff Beck 00:05:00

of, you know, a framework in which AI has the property that it's all about making people smarter, not about alleviating people from thinking.

Dr. Tim Scarfe 00:05:10

Interestingly, there there's work done on, you know, theory of mind analysis on LLMs, and that they appear to have some kind of theory of mind. You can give Okay. So so superficially, if if nothing else.

Dr. Jeff Beck 00:05:21

Oh my gosh. So so the first so I remember when when when I the first it's many years ago when Chad not many years ago. When Chad GPT first came out, the first thing I did was I was I gave it the theory of mind test. There's this canonical theory of mind test, and kids who are about, you know, 3 3 and 4 and some and 5, like, they just they can't pass it. Right. And then sometime 5, 6, 7, they they acquire the ability to pass these this sort of, you know, quote, unquote, theory of mind test. You know, dad and kid are sitting at the table together, and mom comes in, and she puts the stapler in the drawer. And then dad and then she leaves the room. And then dad comes up, and he grabs the stapler, and he puts it in the cupboard. And then they ask the kid, and then mom comes back into the room, and then you ask the kid, where is mom gonna go to look for the stapler? Mom wants the stapler. Where is she gonna go look for it? And little kids say that she's going to look for it where it is. Duh. She's like, mom. She's like, omniscient. And so of course, she's going to go where so they're not able to solve the problem of seeing the world from the perspective of another person. They're not able to say, No, actually, mom didn't have that information that dad moved the stapler. Right? And so right around 6 or 7, transition from not having this ability to having this ability, and we say that they have been past the theory of mind test. And I was giving this test to my kids about every 6 months or so. Just did. So I was super curious. So ChatGPT comes out. It's 1 of the first early versions that you could, like, just sort of start horsing around with. And so I, like, I, you know, I sort of, like, to give it just type chatting in a variety of, like, you know, theory of mind tests. And I was absolutely shocked when I was figured out which perturbation of it was. I painted the scenario and asked it what the answer was. And it very cleverly responded by saying, That's the theory of mind test. Let me tell you all about it. And I was like, Oh, okay. So you didn't pass the test. You were aware of the

test. Right? And so to some extent, showing that an LLM can pass a theory of mind test is fundamentally conflated with whether or not it goes with the theory of mind, whether it's already taken the test, so to speak.[4]

Dr. Tim Scarfe 00:07:43

Yeah. Can attest

Dr. Jeff Beck 00:07:44

shown the test.

Dr. Tim Scarfe 00:07:45

I can attest that. I mean, the cohost, there's this famous riddle. I think it's when John and Mary, they they both like, 1 1 that there there are 2 numbers they have in their mind, and I think John can knows the product, and Mary knows the sum. And you have to kind of, like, go go through this all the possible combinations. And Mary said she doesn't know, and then John says they do know. But anyway, you are such a chat GBT and it's like, oh, yeah, this is the famous riddle. It's the summoned product riddle. And and it gives you the answer for 13. And, obviously, like, if you modify it, you know, you add an additional element to it. It's always going to just give you the canonical answer, even though it's wrong. You say to it, no, no, that this is a variation. We changed it. You need to you need to think. You need to reason. I tried to deep research. I tried o 1 pro. It just it just failed every single time.

Dr. Jeff Beck 00:08:30

Yeah. Because it's overly focused on, like, that very specific example. Yes. That is not not generalizing well.

Dr. Tim Scarfe 00:08:36

There is there is this thing that the way we've been describing all of this is that we need to have direct physical experience. And that's how epistemic foraging works. We just, you know, we have direct physical experience. We compose things together. But we're a bit more

[4] LLM performance on theory of mind tasks (Kosinski) — paper

The sum and product riddle: pattern recognition vs reasoning

Dr. Tim Scarfe 00:08:50

sophisticated than that. We we think we have imagination. I gave you the TikTok example. We live in languages about the propagation and accumulation of information without direct physical experience. It's 1 of the greatest things that have happened to humanity. So, you know, what is what is the role of just imagination, basically, when

Dr. Jeff Beck 00:09:07

we create new knowledge? Not create new knowledge. Create new things. Yeah. Right? That's that, you know, the the last several hundred years have been, you know, 1 amazing technological achievement after another. And I like to think about how it is that we accomplish that,

Systems engineering and object decomposition

Dr. Jeff Beck 00:09:30

and was it inevitable? And I think that the answer is is that it kind of was inevitable. We started out with this, you know, we started out in this world full of objects, and we had to start modeling those objects and describing those objects. And we found that like partitioning the world into these things that are defined by how they interact was really convenient for prediction. But it also gave us a new skill. It gave us systems engineering for free. How does systems engineering work? Well, have this system and this system, you have all these subsystems and you have rules that govern how interact with 1 another, which allows you to do what? It allows you to build something new like Legos. Legos are great. This is not dynamic, right? It's just like Legos. It allowed us to take these objects and put them and say, Well, what would happen if I just stacked things this way or I strung the the the this set of or I connected these objects in in this new way. What would result? And what would result is something that is you could, you know, also call an object. That is sort of besides the point. But it does something new interesting.

Dr. Tim Scarfe 00:10:33

Yes.

Dr. Jeff Beck 00:10:34

Right? Yes. That's the kind of thing that I want that that that I wanna see, like, AI start to be able to do. I wanna see an AI that doesn't just sort of describe the world in which we live, but actually is capable of doing I want to automate systems engineering. And the first part of automating systems engineering is dividing the world up into subsystems and the rules that govern their interactions. That's what makes it work. That's what allows you to take a bunch of simple things and put them together in a new and interesting way.

Dr. Tim Scarfe 00:11:08

On on the understanding thing, there are people who say that, oh, it's it's not just representing the language. It's implicitly building a world model. So it's actually learning through the language. It's it's modeling things that go on inside your brain. And I can give you an example of me learning via a disembodied experience. I can go on TikTok and let's say my cousin is in Thailand or something on holiday, and she's on the beach, and she's doing all of these things. And I'm I'm not there. And and I'm I'm I'm vicariously I'm just like an LLM, you know. I'm I'm having that vicarious experience. What's the difference?

Dr. Jeff Beck 00:11:42

The difference is is that you have something specific that's outside of the linguistic space that those symbols refer to. Right? You you you so maybe you can't, like, explicitly imagine, like, the beautiful you know, it's like, oh, I've never truly experienced the beautiful white sand and the the gorgeous sun and the calming waves. But but you have similar enough experiences that were grounded in reality that you can extrapolate like what that experience would have been like, right? It's all about being able to point it. It really is about having those symbols, meaning something specific and real. The manner in which you choose to represent

Vicarious experience and grounded understanding

Dr. Jeff Beck 00:12:20

the stimulus actually has implications for the neural code. So if you assume that it's orientation, as in it's literally angle and angle is the thing that's being represented, then it supports 1 then then the neural then the evidence, the neural data supports 1 coding paradigm. Right? And if the but if you say that, no, no, it's actually just feature intensity where this where the feature in this case would be not not the the angle, but the actual waveform itself. Right? Or and the intensity of the various different contributions of parts of that waveform. And now in principle, like, there's a 1 to 1 mapping. Right? Or not a 1 to 1 mapping. There's a mapping from the orientation, right, to this sort of other representation that's like feature intensity based. So there's a relationship between these 2 things, but which 1 of those 2 representations of the world you choose, right, ends up with you know, ends up making different predictions about the world. So so the issue ends up being, like, choice of representation actually matters. Right? And we like to phrase this in a Bayesian way in saying, like, choice of generative model actually matters. So if I have, like, this generative model, and I say this is what the this is the brand's generative model. This is these are the underlying stimulus values or the aspects of the world that it's representing. Then my then the then these are my predictions, and they're great. Right? It represents the same image. Right? It represents the same thing out there in the world, but it's a different way of representing it. And then now the data is consistent with my theory. So choice of gen like, specific details, choice of representation, choice of generative model is the thing that ends up ends up being what separates these 2 sets of predictions. Yes. This is sort of highly problematic because we don't really know what the generative model is that the brain is using. Like, Ralph's is perfectly plausible. Mine's perfectly plausible. Right? They both represent the same class of images and things like that, but in completely different ways and therefore make completely different predictions about neural data under these different coding hypotheses.

Dr. Tim Scarfe 00:14:21

And 1 1 approach of that is and I think you're going for this line of sight legibility.

Dr. Jeff Beck 00:14:26

So Right. So if you had an algorithm that explained its reasoning every step of the way, you would get 2 things out of this. The first thing that you would get out of it is you would be able to understand why it did what it did. You incorporate that in your beliefs, and then you would benefit you would benefit personally. You would be smarter because of the algorithm's outputs. And that's kinda what I want. I you know, do I want smart computers? Yeah. But, ultimately, I just wanna be smarter. Right? I I would like to know more. I would like to have I would like to have an understanding of how the world works that's richer and fuller and better. And, you know, an AI that's just able to, you know, do prediction and make decisions, it it it's nice, but it doesn't really get me what I want, which is I wanna understand what's on. Now, that's the other thing that the other thing that happens is that if things are if you can analyze its decisions, or you analyze its line of reasoning

Dr. Tim Scarfe 00:15:19

I I'm

Neural coding and choice of generative model

Dr. Tim Scarfe 00:15:20

I like this idea of thinking about language not as being an intrinsic representation of the world. You know, there's there's the Shannon idea that it's actually encoding information about the world as a first class representation. Or the other approach is that it's just pointers, right? So we're all cognizing agents, we only need this 10 bit stream per second of language because it's just a pointer to the same thing I have in your mind.

Dr. Jeff Beck 00:15:43

That's right. And and and so absolutely. Right? And that's 1 of the reasons why that that's 1 of the reasons why language does work. It works because inside your brain is a very, very sophisticated model of the world that is remarkably similar to the 1 that's inside my brain. Right? Otherwise, it would not work. Right? Our brain we we have a common you know, this is you can think about this. We have a common grounding. Our our our cognitive there's a common grounding of your cognitive understanding of world and my cognitive understanding of the world. And that commonality is what allows language to work as well as it does. Where did that come from? What is the commonality? This is 1 of the arguments that

comes out of Max and Cosmo Max Ramsay and Carl Ferson's Bayesian mechanics work. The basic idea is that, okay, so language works between us because we our brains have a common grounding, right? They live in the same space. Like, you can use words, and I can take my knowledge of how my brain works and use that to make assumptions about how your brain works, and then I can take that word, I can map it onto this rich set of cognitive things in my brain. But where did that come from? Like, like, where what is the the basis of that understanding? And the the the Bayesian mechanics argument is is that it came from our experience. It came from evolution. It it came from the world in which we evolved. The rich understanding that we have in our brains is a result of the fact that we live in the same that that is in common, is a result of the fact that we live in the same world, right? This is part of the this is like the the the the backhanded way of getting to an embodied intelligence argument. The only thing that makes communication[5][6]

Line-of-sight legibility in AI reasoning

Dr. Jeff Beck 00:17:30

possible is that we had to adapt to survive in precisely the same kind of environment. And not exactly the same environment, but the same kind of environment. We had to develop, you know, the cognitive tools that were necessary to do that. Otherwise, we wouldn't be here. Right? We'd dead. And that common grounding is what makes this rich communication possible. The basis of our understanding of the world is not linguistic. It's an understanding of the of the intuitive physics of the world in which we live. I remember this is this is what I hate I hate talking about consciousness, and I try to avoid it. And, you know, I used to I used to shut down conversations on the topic by by simply saying, you know, so what do you think a conscious is? Like, consciousness is the suboptimal way in which we perform reasoning when doing it correctly gets too hard, right? It's when we collapse our probability distributions. We say, no, no, it actually just works like this. Here's this real simple explanation. But that's what words are for. Words are for giving us simple explanations that we can convey to other people. But they are, in fact, massive simplifications of what our brain is actually doing, right? Now, if you take the functional region story about the brain very seriously, we have this little bit, you know, here's language, it's over here. And it's a tiny part of our brain. It's the tiny part that we use to communicate with 1

[5] Bayesian Mechanics (Ramstead, Sakthivadivel, Heins et al.) — paper
[6] Building Human-like Communicative Intelligence (Dubova) — paper

Language as lossy compression of cognition

Dr. Jeff Beck 00:18:50

another. But it it it's a very, very compact, lossy summary of what our brain is actually doing, which is something much, much, much more sophisticated. And you can think of conscious perception the same way. It's like, we know that people represent the world probabilistically for the most part, and yet we don't perceive it. Like, perceive you're there, unambiguous. Right? We don't we don't perceive it that way because it's just too complicated. Right? The brain is doing an incredible amount of processing that we are quote unquote completely unaware of, and our awareness is a really is really just a compact summary of that, that that that is a poor representation of it. This is why I don't want this is why we need to move away from language models, by the way. Language is an incredibly poor summary of what it is that our our our brains are actually doing. And language is an incredibly poor summary of of of the reality that it is meant to describe relative to the sum relative to the description of reality that's actually in our brain.

Dr. Tim Scarfe 00:19:50

Certainly, when navigating romantic relationships, we we do have, language for all sorts of weird things, you know, even things like sour grapes and Sardin Freud. Are many quite sophisticated, ambiguous situations that we have words for. Perhaps we have differential understandings of what those words are, but explore a little bit more what what you meant by language and ambiguity.

Dr. Jeff Beck 00:20:10

The way I conceive of how we perceive the world is probabilistically. This is the whole Bayesian brain hypothesis situation. And it's very complicated. There's lots of things going on. We have this, you know, we we we we intuit a lot of this stuff. We have this intuitive notion of how the world works. But when it comes to compressing it down into language, our bit rate is tiny. Actually, who am I thinking of? There was a paper recently, Marcus Meister. I got it. So Marcus Meister put out this really great paper recently in which he pointed out that the amount of information coming into our brain is huge, absolutely massive. And he calculated the bit rate. I don't remember what the numbers are, but it was like it was a it was 1 of those things. It was a big number. Right? And the amount of information coming out of our brain is, like, 10 bits a second. Yep. Right? So absolutely tiny. Right? Well, 1 of the things that's coming out of our brain is, like, language. So we know that there's this incredible there's incredible difference between what's going in and, like, what's coming out. And there's a he has a very you know, there's a bunch of reasons why that might be. It might just be that our affordance space is, like, really small at the end of the day relative to, like, you know, our sensor space and things like that. But 1 of the things that it that that that you can instantly conclude from this fact is that whatever it is that we are doing, it's an incredibly poor representation of our of our understanding. You can't just take actions and translate them into beliefs in a very easy way. When you have a situation where your beliefs are being formed by this massive onslaught of information and our actions, which includes our words, right, are so limited. And so this is 1 of the reasons why, like, I don't think that, like, basing your model of the world or your AI model on language is the right move because language is an incredibly poor representation of the reality as we under as we must understand it in order to function at it.[7]

Dr. Tim Scarfe 00:22:19

Many many folks, I mean, Steven Piantadosi, of course, we have this concept, Royal Semantics, you know, that that he he will argue that language models can understand. Now they certainly seem to be, capable of of processing symbols in a way which makes sense to us. And and then there's the question of, well, at some point, mimicry, if it's correct enough of the time, it's there's there's no meaningful difference. But but what would you say to people who really think that AI systems could understand? Just[8]

[7] Why do we live at 10 bits/s? (Meister, Zheng) — paper
[8] Steven Piantadosi — website

10 bits per second: information processing bottleneck

Dr. Tim Scarfe 00:22:50

don't. No. So.

Dr. Jeff Beck 00:22:56

Your diction is so so look. Language models are are very impressive. Right? They're very good at, like, what they were trained to do, which is next word prediction. Right? And you give them enough data and they can do it. They can produce, like, compelling, you know, it it seems as if it came from a person. Right? And when you ask them for an explanation, they can come up with a a compelling sounding explanation because that's what they were trained to do. But I wouldn't trust them I wouldn't trust those explanations to be true because effectively, you're you you know, they've been trained to basically play a kind of like Mad Libs. You know, Mad Libs, it's where you sort of put the noun here. Yeah. And so it ends up having the structure of an explanation because it was told what the structure of an explanation was. Right? And it filled in words that made sense based on which is great. So it it it it can replicate things that it saw in its data. Right? That in in a way that that that that seems plausible, but not in a way that is that that that I would say is necessarily correct because it is just manipulating symbols. It is just looking at correlations. You know, I'll I'll believe that a language model is is moving in the direction of intelligence when it can produce something that wasn't in its dataset. Right? And everyone says like, oh, but it does 0 shot. There's no such thing as 0 shot learning. Right? There's just stuff that was in your training data that you weren't explicitly aware of. Right? And and that's, you know, and, you know, which includes correlations between words. And so that's sort of my, like, high level response is that is that is that manipulating symbols is can be compelling. It can look right. Doing know, the ability to predict what someone would say is great. Yep. Right? But it's not an indicative it's not indicative of actually understanding. And that's because, again, we can go back to, this is an embodied intelligence argument. Right? It's because those symbols were just symbols. They're just relationships between the symbols.[9]

[9] Mad Libs (Stern, Price) — website

Why language models cannot substitute for understanding

Dr. Jeff Beck 00:25:00

Right? And we can add vision, you know, or vision language action model. And then so we're getting a little closer. It has some more of the flavor of of of of an of something that would be embodied. But until a language model actually produces something like a creative insight, builds some invents something new, I won't believe that it's intelligent. That's sort of my, like

Dr. Tim Scarfe 00:25:22

Just for the audio, I was reading this great book, The Brain Abstracted, and and and she said, yeah, abstraction, ignoring details, idealization, something that scientists do, is deliberately distorting the truth.[10]

Dr. Jeff Beck 00:25:32

Yes.

Dr. Tim Scarfe 00:25:33

And, yeah, so we can say the Earth, it's a sphere. It's not a sphere. It's just like all the other planets. They're spheres as well. It's an idealization. And Newton is still around. We still use Newton. That's an idealization, but it's very, very useful. But, there are many scientific theories, talking about these unobservable things like subatomic particles and things we've never seen before. And these are perhaps just blatant distortions of of the truth. But it's very abstract, isn't it?

Dr. Jeff Beck 00:26:00

It is. It is. I actually the the Newton example is a good starting is is a good place to to to start, right? So there's a criticism of our intuitive understanding of physics, which is it's wrong. And so the classic example is is that is that, like, you know, if I give you a bowling ball and a feather and you say, like, well, what's the you know? And I say, I'm gonna drop them. Like, Newton says that, like, you know, they drop at the same at the same space, which is, like, completely countertive. So your your intuitive understanding of physics is incorrect. Now that's not how I think about it. Your intuitive understanding of physics is totally correct because you've never seen a bowling ball and a feather dropped in a vacuum. Right? We

[10] The Brain Abstracted (Chirimuuta) — book

don't do that experiment, like, in our everyday life. Right? That's orthogonal to our actual our experience is that, no, they actually drop at very different rates. Why? Because the world in which we live doesn't include a vacuum. Right? If we lived in a vacuum, then we would have a very different understanding and model. And then our intuitive physics would be more closely aligned with what Newton would say. So in this sense, our physics model was too simple. Right? So it's not the correct conclusion isn't our intuitive physics is wrong. Right? The correct conclusion is is that the scientific model was too simple.

Dr. Tim Scarfe 00:27:17

Right? With increasing sophistication of affordances or, you know, connectedness to what's outside the boundary, there must be more representation, more modeling going on inside the boundary. And certainly, in living things, we seem to see a kind of nesting of boundaries. Is that always the case? Because we're very sophisticated and we can do lots of things in the physical world. Does that imply that there's a nested subsystem? Or could you just have 1 big fat system which is just very good at modeling

Dr. Jeff Beck 00:27:45

the work? Okay, so I'm going to go I'm going to so where you draw your the the biggest problem with the Markov blanket story is that it does not provide a definitive answer to the question of where should I draw my boundary? Yep.

Scientific abstraction and idealization

Dr. Jeff Beck 00:28:00

Right? So like in a, know, when we do fluid dynamics, right? We always say that this prototypical fluid element, and whenever we draw it, it looks like a kidney bean, right? Because we don't know what shape it is, we're largely agnostic. We don't care, right? Any partitioning of a glass of water into a bunch of fluid elements is completely arbitrary, right? Now, you can still consider each 1 of those things as a Markov blanket, then you can derive macroscopic equations from that, and it all works out just fine. The same thing happens within the free energy principle, right? It tells you that if you draw a blanket, right, you can define an object based on the statistics of that blanket, but it doesn't tell you where to put it. I could draw a blanket all the way around Tim, right? Or I could put 1 on your arm and 1 on your wrist and what, you know, all of these things are acceptable partitioning of your system. This was the problem that I was working on when I was still back at Duke, was trying to figure out like, you know,

how do you actually select the right partitioning of your system? How do you divide the world up into parts and pieces that makes the most sense? Now, don't I would say I have an answer to this problem. I will not be saying, however, it is the right answer or the best answer. It is a answer. The traditional approach to the free energy principle, they took the definition of a Markov blanket was more aligned with the statistical definition, right? I an element and I say, Well, what's the blanket of this element?[11]

Markov blankets and system partitioning

Dr. Jeff Beck 00:29:40

And I'm going pick 1 because like, Oh, maybe, you know, it's like, I don't know, maybe it's in a region of high density or how you pick is an arbiter. Every element that you pick has a blanket, And the way that they originally did made decisions about where to draw their blankets was just by basically saying, Well, I know it's gonna have a blanket, so I'm just gonna find all the things that it bumps into and interacts with. And then I'm gonna keep expanding that blanket until I get something that that that captures all of those interactions. Right? So I like to think of it as, you know, you've got a bunch of stuff and it it but these guys, you know, the guys inside are all tightly coupled. They never over throughout the entire course of the time evolution of the system, these guys never interact with these guys over here, right? And so I can sensibly draw a blanket based on this sort of long term estimation of the adjacency matrix of the system. So I can look at who 's interacting with who at any given time, you know, just average that, just sum them all up, get 1 big effective adjacency matrix, and then I can just sort of divide that and pick a node and say, Well, what's the smallest, you know, isolated cluster of interactions? That becomes the internal states, The the the you know, and then it has a boundary because it has to. Right? Because that that's what allows you to segment. Now, systems that don't have blankets are like, you know, water molecules that are just bouncing off of each other and are constantly mixing and interacting.

Dr. Tim Scarfe 00:31:01

There are folks who are science realists. And I guess that that's a form of Platonism where basically you think the the universe is is is written in code. And the code is legible. You know, if only we could find the code, then we know how the universe works. And and the other school of thought is that no, actually, do have these very idealized models. They they have, we can model pockets of regular-

[11] Markov blankets in biological systems (Friston) — paper

ity, and then there's a load of noise. And we just, we decide as scientists that the noise is acceptable. Like, you know, the noise doesn't mean anything. It doesn't add any value. It won't ever be useful to somebody else. I mean, noise[12]

Dr. Jeff Beck 00:31:32

no no noise is in fact simply how we model our we model either ignorance or shit we don't care about. Right? That's what noise is for. Right? It's it is a model of ignorance or a model of something you've decided is irrelevant to the current situation, right? These are design choices though when we design our theories, right? We we we we we identify the domain in which we want the theory to operate, right? And we decide what it is we care about and what we don't care about. Often, certainly in the lab, right? You have this well controlled experience because you know you don't want to deal with all the complexities of the real world. And that does induce biases in the theories that we discover, right? And it's like the neuroscience example where we've been showing Gabor patches and moving gradients for far too long, and that actually has created a bias in our understanding of how the brain You know, is there is there a mathematical system or set of equations or laws of the universe that that that that are true and discoverable?[13]

Dr. Tim Scarfe 00:32:30

Yeah. I mean, do do you do you think that is the case, and do you think

Dr. Jeff Beck 00:32:33

they would be legible? So I I remember I was giving a talk at 1 point, and a student asked so I'd been talking about, like, you know, it was a term of variational inference and how you have your approximate posterior distribution, and you have, like, your ground truth distribution, and you're trying to find it. And and and a student asks, is is there a true p? So p was what I was using for, like, the ground truth distribution. Yeah. Which which which was done for convenience. It's like, there is there a true p? And and my response was to simply say no. Now it was done in in part in jest and also because I just didn't wanna answer the question because there there must be, in a sense, like statistical

[12] Mathematical Platonism (Linnebo, SEP) — website
[13] Critique of Gabor patches in neuroscience (Tsao) — paper

Scientific realism, noise, and the limits of models

Dr. Jeff Beck 00:33:20

reliable statistical regularities in the universe, right? If there weren't, we would not be having this conversation. Now, you know, is the precise nature of those statistical regularities discoverable? Maybe. But it's kind of irrelevant. Right? We know that we you know, so so in the sense, like, there a true p? Yes. There probably is because if there weren't, we wouldn't be having this conversation, but I don't know what it is. And I'm not entirely certain that it's discoverable. But even then,

Dr. Tim Scarfe 00:33:49

is is there a hierarchy that there's a strong statement, which is that there is no, ontological or epistemological divide, you know, that we can actually build a model that just models the world as it is, you know, not some lossy abstraction. The the the sort of deflationary, the weaker version is there are regularities in the universe, and we can build models with very significant predictive power of those regularities. Right. But are they true? That's

Dr. Jeff Beck 00:34:17

the question I believe the student was attempting to ask. That's the question that I'm trying to figure out if I'm being asked now.

Dr. Tim Scarfe 00:34:22

Would wouldn't it be false by definition? If if you're not modeling everything, wouldn't it be false by definition?

Dr. Jeff Beck 00:34:27

Yes. It is false by definition. Right? There's like, you know, this is I I like to say that a a proper Bayesian places no metaphysical significance on the parameters and latent variables of their models. Right? It's always the case that there's there's other models that haven't been explored that may be that may be better. And there's other you know? And, you know, unless you have some sort of exhaustive procedure for model generation and testing, which we don't

Dr. Tim Scarfe 00:34:54

Yeah.

Dr. Jeff Beck 00:34:55

There's no way you could even conclude that 1 of them is the best.

Dr. Tim Scarfe 00:35:00

Well, brings me to the next thing is, my, cohost Keith Thugger, you know he's a massive Bayesian, and he loves saying all knowledge is conditional. And I was reading my notes. You you say exactly the same thing, you know. But that rather leads to this quite constructivist idea. Not the strong sense, you know, where everything is is just like, you know, fictitious, if you like. But that, our our knowledge of the world does depend on the questions we ask and and the the tools we use in previous discoveries and and so on. And that rather lends yeah. It it begs the question whether some of these foundational theories, even like the free energy principle, whether they are actually statements, true statements about the universe, or they're

Dr. Jeff Beck 00:35:41

just 1 perspective of of many. They're 1 perspective of many. The free energy principle is not a theory. Right? It's a it's a mathematical principle. It's it's I mean, at the end of the day, it's like James Maxcal, right, or Maxent, right, coupled with the Markov blanket notion that basically says the[14]

Free energy principle as mathematical framework

Dr. Jeff Beck 00:36:00

world can be sensibly segmented into meaningful parts and pieces. But that's all it is. It's a mathematical principle. It's a way of guiding how you create your models that I believe is very sensible. But it's not an actual, like, theory. It is a mathematical framework. And I love mathematical frameworks. They're great. They constrain your theories, and if they're good frameworks, they constrain them in sensible ways. But they're just theories, and that's what a proper Bayesian would always say. It's you've got a model. It's just a model. It might be a very good model in terms of its ability to do prediction and data compression, but it's just a model.

[14] MaxEnt and MaxCal principles (Presse et al.) — paper

Dr. Tim Scarfe 00:36:39

We can build an AI model today that does a really, really good analytical job, and it will never be legible to us. A lot

Dr. Jeff Beck 00:36:45

of times, all you care about is, like, predicting what are the consequences of various actions that that that that you could have performed. This is the the prediction. You don't you may not, in fact, care about the explanation. You may just sort of trust that it's like, Well, you know, it's been validated. I've done all the checks that I'm supposed to do, and I'm just gonna continue to trust it. I don't see a real problem with that, but I find it very unsatisfying, right? I would like to understand what's going on. But the reason I would like to understand what's going on is because if we start putting these algorithms in charge of making decisions for us, we are presuming something that is really hard to do that is really hard to demonstrate, and that's that they are making the decisions that we otherwise would have made. This is what alignment is all about. In an ideal setting, you have this high res, super smart AI, does amazing prediction algorithms. And it makes all the decisions. And based on those predictions, it makes all the decisions for you and does so in a way that's consistent with your values. That is a necessary prerequisite, in my view, for determining whether or not the AI has your values. So this is the biggest problem that I think we have with AI as it currently stands. In an RL setting, you have 3 things. You have your prediction engine. And I think we can argue that machines in certain domains have much better prediction engines than we do. And arguably, we can use fan more data. The prediction engines that they have will be significantly more precise than our own, right? And this is true for physics prediction engines are also It's the same sort of thing, right? Our intuition has been replaced by these formal physical models right, that do a better job of predicting the outcomes of experiments

Black-box prediction vs legible explanation

Dr. Jeff Beck 00:38:40

than our naive intuitions, right? So you got a prediction engine. You have your reward function, which basically sets your goals and determines your, you know, and represents the values of the system. And then you have your action selects your policy, right? And the policy and, you know, in most of RL, right, is all about figuring how do I compute my policy, right, given my predictions, right, and my values? But that's the whole point, right? Prediction plus value equals policy. Now, when we start just trusting AI algorithms, what we're really basically say-

ing is that we're trusting them to execute a policy that's consistent with our reward function, right? But with a better belief model than ours, right? That's the goal. The goal is like, it's able to do better prediction than we possibly could. It has the same reward function as we would. Based on those predictions, it's making decisions in the same way that we would, but then its policy ends up being better or more effective at achieving that reward. It's not impossible to have an AI system that does go through this process of identifying our values to some extent, but it will never be perfect. It'll always be the case that it will have an impoverished view of our reward function. It will have an impoverished because there's no way it could really, really hammer down precise knowledge of our belief formation systems. It could get a pretty good 1, And from that, it could be trained to to to estimate reward functions.[15]

Dr. Tim Scarfe 00:40:08

Because there there was that famous reward is not enough paper by sorry. Reward is enough paper by by Sutton.

Dr. Jeff Beck 00:40:15

And and, yeah, and and he's absolutely right. If we could if we could write down the right reward function, I think that to some extent, that's true. That would solve a lot of problems. Here's the question, though. What's the right reward function? How do you compute that? Where does that come from? Why is you know, it are do we even have the same reward function? Probably not. Probably not. Yeah. Yeah. So so so you're saying there let's imagine we took all

Dr. Tim Scarfe 00:40:41

of the humans on the planet, and we described them using this contextual reward function that you're talking about. It seems to me that that object would be 1 of the most sort of complex artifacts in the universe.

Dr. Jeff Beck 00:40:52

Quite possibly. I mean, could just average them all, you know, just voting, you know? No. The the the it it's it's

[15] Reward is Enough (Silver, Singh, Precup, Sutton) — paper

Reward functions and the impossibility of alignment

Dr. Jeff Beck 00:41:00

not a question of, like, what to do with what to do about, in my view, it's not a question in terms of what to do about the fact that people have different reward functions. Taking utilitarian, no problem with that. Taking rank utilitarian approaches average. The real question is what is your reward function? Like, if you could write it down for me right now, then, you know, that'd be great. I'd love to see you try. No, you know, it's really, really, really difficult to get people to express their reward functions in a way that can be translated neatly into the language a computer can understand. I do believe that we kind of do sort of have something that resembles a reward function. I don't like to talk about it in terms of reward functions. I like to talk about it in other terms. But let's just, like, for the sake of our, let's just go with that. So, where does your reward function come from, and how do you represent it, and what is it? How do I get access to your reward function? Because if I want like, if I wanna replace you with a robot I don't wanna replace you with a robot. But if I wanted to, right, that's something I would need to do. I would need to be able to figure out what's your belief formation mechanism and what's your reward function. And it and, you know, it turns out that, like, it's really difficult to do that. In fact, I'm gonna argue that's impossible. So not only is it the case that there's no normative well, barring divine intervention, there's no normative solution to the problem of reward function selection. There's no, like, that's the right reward function, right? Not only that, right? That there is no like right or wrong reward function that you can obtain just from data. It's also the case that it's very difficult for me to figure out what your reward function actually is. A lot of work in behavioral economics has gone into trying to figure this out. And every time they do, they keep finding these, like, weird things that happen. And some of them have really sensible explanations, like, it's because it's context dependent, or Oh, it's a But it turned out just measuring someone's reward function, right, which was a prerequisite for getting an which I'm again, is a prerequisite for getting an AI algorithm that that shares our values, right, is really impossible because you cannot disentangle belief and reward based on policy. Right? So if I'm just x if I if I if I take an action and and I'm a rational maximizer and I take an action and it is and it is a product of my reward and belief, and then there's like an argmax operation. So there's a convolution, right, between reward and belief, right, followed by an argmax. But but you only observe the action.

Dr. Tim Scarfe 00:43:23

You said you can't then deconvolve Yeah.

Dr. Jeff Beck 00:43:25

How do you so you can't go backwards. Right? If you don't know someone's beliefs, you cannot infer their values. If you don't know their values, it's also very difficult to like disentangle their beliefs. But for certain, right, without knowing somebody's beliefs, you cannot infer their values. When 2 people are arguing, the first so they disagree about a conclusion. They disagree about what action they're going to take. And when you want if the goal if you don't want to resolve the argument, you of course instantly draw 1 of 2 possible conclusions. The person that I'm arguing with is either evil or stupid. They're stupid because their prediction engine, their beliefs are wrong, their prediction engine is wrong, or they're evil because their reward function is not the same as mine. That's the lazy thing to do. Now, if you're actually trying to have a productive argument with someone, what do you do? You start talking about beliefs. Say, Well, why did you come to that? Why was that the right action? And then someone will say, Well, it's because you know, here's the list of facts, and then here's the predictions that I'm making based on those facts. And then you argue about whether or not the predictions follow from the facts. Right? Or you argue about, like, missing facts that they should have included that they didn't when making their prediction. So what are you doing there? You're sitting there and you're having a conversation in an effort to determine the other person's belief formation mechanism. Once I have, in my mind, a belief form an understanding of your belief formation mechanism, then and only then can I finally infer or get a good estimate of what your reward function or value function actually is?

Modeling beliefs as prerequisite for value inference

Dr. Tim Scarfe 00:45:00

How could we build aligned AI system? Because even, you know, I'm sure, like, Tesla, they're recording lots of situations where people are running over rabbits and stuff like that. But fun fundamentally, though, there there's a real sparsity because I think those really important situations might be underrepresented or not represented at all. Many approaches to aligning these predictive, you know, stochastic systems is, okay. Let's do let's do 2.

Dr. Jeff Beck 00:45:23

Let's have a meta model, which we, have guardrails and we apply rules and and so on. Or or do you think we could actually incorporate it into the base system itself? And and how would we do that? What we'd like is we'd like to is it would be very awesome if we could actually figure out what our reward functions are and, like, get them into into our AI agents. That's that's really challenging. I think that the only way that we're gonna ever get anywhere close to that is is if we build if we build an additional layer into into our sort of smart a into our AI algorithms. It's an additional layer that explicitly models my beliefs. Right, that engages in the conversation that I proposed earlier, where you and I are talking, and we disagree. And so, we sit down and we talk about, well, what are the possible reasons why we might have disagreed? And in the process, you and I each form a model of the other person's belief formation system, right? And to the extent that that's an accurate model, we can then conclude that any differences are difference in reward functions. So if we wanted to become aligned, 1 of us is going have to it's going to be you, by the way. 1 of us is going to have to change our reward functions if we want alignment once we've gone through that process of agreeing on a belief formation mechanism.[16]

Dr. Tim Scarfe 00:46:45

I see. Jeff, thank you so much for joining us today.

Dr. Jeff Beck 00:46:47

Absolutely. Thank you.

[16] Modeling Human Beliefs about AI Behavior (Lang, Forre) — paper

References

[1] **Bayesian inference in neural computation (Ma, Beck, Latham, Pouget)**
https://www.nature.com/articles/nn1790
paper

[2] **Noumenal Labs (Jeff Beck)**
https://www.noumenal.ai/
website

[3] **Noumenal Labs research paper**
https://arxiv.org/html/2502.13161v1
paper

[4] **Tufa AI Labs**
https://tufalabs.ai/
website

[5] **LLM performance on theory of mind tasks (Kosinski)**
https://arxiv.org/abs/2302.02083
paper

[6] **Bayesian Mechanics (Ramstead, Sakthivadivel, Heins et al.)**
https://arxiv.org/abs/2205.11543
paper

[7] **Building Human-like Communicative Intelligence (Dubova)**
https://arxiv.org/abs/2201.02734
paper

[8] **Why do we live at 10 bits/s? (Meister, Zheng)**
https://www.sciencedirect.com/science/article/abs/pii/S0896627324008080
paper

[9] **Steven Piantadosi**
https://colala.berkeley.edu/people/piantadosi/
website

[10] **Mad Libs (Stern, Price)**
https://en.wikipedia.org/wiki/Mad_Libs
website

[11] **The Brain Abstracted (Chirimuuta)**
https://mitpress.mit.edu/9780262548045/the-brain-abstracted/
book

[12] **Markov blankets in biological systems (Friston)**
https://royalsocietypublishing.org/doi/10.1098/rsif.2017.0792

paper

[13] **Mathematical Platonism (Linnebo, SEP)**
https://plato.stanford.edu/entries/platonism-mathematics/
website

[14] **Critique of Gabor patches in neuroscience (Tsao)**
https://pmc.ncbi.nlm.nih.gov/articles/PMC9564096/
paper

[15] **MaxEnt and MaxCal principles (Presse et al.)**
https://journals.aps.org/rmp/abstract/10.1103/RevModPhys.85.1115
paper

[16] **Reward is Enough (Silver, Singh, Precup, Sutton)**
https://www.sciencedirect.com/science/article/pii/S0004370221000862
paper

[17] **Modeling Human Beliefs about AI Behavior (Lang, Forre)**
https://www.arxiv.org/pdf/2502.21262
paper