Dr. Jeff Beck: Agency and Energy Based Models

January 25, 2026

Machine Learning Street Talk

Podcast Transcript

Revision: 3921d46a

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

Contents

Overview

What makes something truly *intelligent*? Is a rock an agent? Could a perfect simulation of your brain actually *be* you? In this fascinating conversation, Dr. Jeff Beck takes us on a journey through the philosophical and technical foundations of agency, intelligence, and the future of AI.

Jeff doesn't hold back on the big questions. He argues that from a purely mathematical perspective, there's no structural difference between an agent and a rock – both execute policies that map inputs to outputs. The real distinction lies in *sophistication* – how complex are the internal computations? Does the system engage in planning and counterfactual reasoning, or is it just a lookup table that happens to give the right answers?

Key topics explored in this conversation:

The Black Box Problem of Agency – How can we tell if something is truly planning versus just executing a pre-computed response? Jeff explains why this question is nearly impossible to answer from the outside, and why the best we can do is ask which model gives us the simplest explanation.

Energy-Based Models Explained – A masterclass on how EBMs differ from standard neural networks. The key insight: traditional networks only optimize weights, while energy-based models optimize *both* weights and internal states – a subtle but profound distinction that connects to Bayesian inference.

Why Your Brain Might Have Evolved from Your Nose – One of the most surprising moments in the conversation. Jeff proposes that the complex, non-smooth nature of olfactory space may have driven the evolution of our associative cortex and planning abilities.

The JEPA Revolution – A deep dive into Yann LeCun's Joint Embedding Prediction Architecture and why learning in latent space (rather than predicting every pixel) might be the key to more robust AI representations.

AI Safety Without Skynet Fears – Jeff takes a refreshingly grounded stance on AI risk. He's less worried about rogue superintelligences and more concerned about humans becoming "reward function selectors" – couch potatoes who just approve or reject AI outputs. His proposed solution? Use inverse reinforcement learning to derive AI goals from observed human behavior, then make *small* perturbations rather than naive commands like "end world hunger."

Whether you're interested in the philosophy of mind, the technical details of modern machine learning, or just want to understand what makes intelligence *tick*, this conversation delivers insights you won't find anywhere else.

Geometric Deep Learning & Physical Symmetries

Dr. Jeff Beck 00:00:00

Geometric deep learning is a big part of like is a big part of the stack if for no other reason than when we talk about, like, modeling the physical world. That means, like, incorporating the symmetries that exist in the physical world. So it's like we're highly motivated to employ a lot of those methods and techniques.

Dr. Tim Scarfe 00:00:15

But is the world written in code? Or do you mean exploiting the regularities in the code that seem to have some

Dr. Jeff Beck 00:00:21

Exploiting the regularities. No. It's just like, look, we things are it is the world is translation invariant. The world is like rotation. Well, not really because there's gravity. But like in principle, you know, there is a principle axis, but it's certainly rotationally invariant in the x y plane. Yeah. And if you if you wanna have a good model of the world as it actually is, it should incorporate those features. And of course you can discover it, you know, in a brute forcey way, but the mathematician in me really wants to build build the symmetries in. And fortunately, we've got a lot of great tools that were developed over the last several years that can do that.[1]

Defining Agency: From Rocks to Planning

Dr. Tim Scarfe 00:00:56

What's your view on agency?

Dr. Jeff Beck 00:00:58

If I'm being, you know, like an FEP purist, I have to sort of say like, oh, well there's no difference between you know, an agent and an object in a very real way. At least there's nothing structurally distinct between what how we model an agent and how we model an object. It's really just a question of of degrees. Right? An agent is is a really sophisticated object. Right? It has internal states that represent things over very long time scales. You know, it has sophisticated poli-

[1] Free Energy Principle (FEP) — Concept Jeff mentions being an "FEP purist" regarding agents vs objects.

cies that are context dependent, which is basically saying really long time scales again, and things like that.

Dr. Tim Scarfe 00:01:35

Yeah. You know, there's the kind of the philosophical highbrow notion of agency that we introduce notions of intentionality and self causation and things like that. I mean, really no nonsense version of an agency is it's just a thing which acts and performs some kind of computation. And I guess you could almost model anything as an agent.

Dr. Jeff Beck 00:02:01

Yeah. Well, if your definition of an agent is something that executes a policy, then anything is an agent. A rock is an agent. It's an input a policy is an input output relationship. When many people talk about agents, they they're adding a few they're adding a few additional elements that I think have a lot to do with how the policy is computed. Right? So for example, when we think of how the difference between, like, us and, like like, really, like, amoebas, we we often cite things like planning, counterfactual reasoning, goal oriented behavior. Right? We're specifying things that that have that that are specific mean that that are all related to how it is we compute our policies. Right? They're latent variables that represent policies that are compatible with reinforcement learning. And that's the defining characteristic of an agent. But you could very easily just say from an outside perspective, if you can't look at how someone or something is doing the computations, if the only thing you observe is the policy, does that mean that you can never conclude that something's an agent? And I would say no. You'd still like to be able to conclude that this is an agent, Even though the only thing I ever get to measure is its policy.

Dr. Tim Scarfe 00:03:25

But do you think we

Dr. Jeff Beck 00:03:25

should have some notion of the strength of an agent? The strength of an agent. Or how is this like a measure of agency? Is that what you mean? Yes. Or yeah. So I mean, I think you could use, like, notions of, like, transfer entropy and things like that in order to estimate, like, the timetable for which something is incorporating information or the degree to which it's taken into it it exhibits a context dependent behavior and things like that. And that would be a pretty good measure. Now is it normative? No, it's not. But it is a measure, and you could use things like that. But at that point, you're really just talking again about policy sophistication, not does it have a reward function. Like, is it actually executing planning?

Dr. Tim Scarfe 00:04:06

Yeah. I mean, there's certainly intuitively agents to me seem to be kind of causally disconnected. Because they're planning into the future, they are not impulse response machines. They're not just part of the massive things going on around them. They are just obviously disconnected from the locality.

Dr. Jeff Beck 00:04:25

So the trick is that, Okay, so I've got this agent. And I know exactly what it does, Right? It takes into account information. Rolls out future internally, it rolls out a whole bunch of future consequences of various different actions or plans that it could take. It selects the best 1, and then it executes it. So all of those variables that occurred inside, from the outside perspective, it just looked like a function transformation. Unless I'm somehow going in and recording and somehow demonstrating the fact that the manner in which it is calculating its policy involved doing those rollouts, I wouldn't be able to show that it's actually doing those rollouts. I would just be able to conclude it has a really sophisticated policy. So can you conclude that something isn't is is so the question is how do you identify something is actually doing planning? And I think that's a really hard question as opposed to having an incredibly sophisticated policy.

The Black Box Problem & Counterfactuals

Dr. Tim Scarfe 00:05:25

I think my intuition is if it feels to me that a function, a simple input output mapping can't be an agent. And in a way, is related to what we were talking about with grounding. It seems that when things are physically embedded in the world, then they're more likely to be agents. This functionalist idea that's just a bit of computer code running on a machine, it kind of feels like that can't be an agent.

Dr. Jeff Beck 00:05:48

It does. So suppose I coded it up so it was doing all of that planning. It's like gets its inputs to some crazy, like massive Monte Carlo tree search, picks the best policy possible, and then executes it. Now you don't observe any of that. Right? Because you know what's going on, you could say, oh, well, it's it's clearly like executing, you know, this is it's doing planning and counterfactual reasoning. It's going on, like, look, there it is. Because you coded it, so you know it's doing it. But if you're looking at it from the outside, right, it you know, if you don't know what's happening inside, it's going you know, all you have access to is, oh, here's the action that it that it that it did given this long series of inputs. And so it's it's really hard to identify what, you know, something as an agent per se from the outside. You kinda have to know what's going on inside. This, by the way, is why I don't think that, like, you know, can you know, these sort of prediction based approaches to, AI are you know, you could sort of say, well, it it's not really doing anything even remotely agentic unless it's executing and doing planning counterfactual reasoning. So, like, your chess program is is like, oh, clearly, it's doing some planning and counterfactual reasoning because you know it's doing it. But but it but you could like write I could describe the exact same set of behaviors just with the policy function.[2]

[2] Monte Carlo Tree Search — Concept Mentioned as the internal planning mechanism that could be hidden inside a black box.

Dr. Tim Scarfe 00:07:10

I think the counterfactual thing is an important feature here because we could take something which was conscious or something which had agency, and we could just take a trace of the actual path which was found. And now we've just got this is reductio ad absurdum. But now we've just got a computational trace. And that thing clearly has now lost whatever agency or consciousness it had. So there's something about considering all of the possibilities.

Dr. Jeff Beck 00:07:33

Yeah. Yeah. I think so in my mind, that is the fundamental feature of an agent. Like if you can show that it's engaged in planning counterfactual reasoning, then it's definitely an agent. My argument is just simply that that's hard to do unless you crack it open and see what's going on inside. Now you could take a a pragmatic view and say, well, if the simplest computational model of the behavior, model it as if it was doing planning and counterfactual reasoning, then you can draw an implicit conclusion that, oh, yes, well, I may as well say it's an agent. And that's kind of the approach that I've taken. So like 1 of the things that comes out of the physics discovery algorithm is that you apply it to agents and what do you get? Well, you get a model. Now bear in mind, I called them all objects before, and I didn't change anything to make it special to an actual agent. Right? But what I do have the ability to do because of the model is I can look at the internal states associated with that object that I want to call an agent, and look at how sophisticated it is, right? And that degree of sophistication is what allows me to say, oh, well, I'm going to go ahead and say that and I like the whole idea. It's a great idea. Let's have a metric, right? And I'm sure it would be something that would effectively be transfer entropy or something like that. But we have this metric on, well, how sophisticated were the internal states that were necessary in order to generate this output? And if it's above some threshold, we'll call it an agent. I don't like thresholds. But we just sort of say a degree of agency, a degree of sophistication.

Simulated Agency vs. Physical Reality

Dr. Tim Scarfe 00:09:00

And coming back to Dennett's intentional stance. So this is that there is a level of representation which serves as a useful explanation, even though it's not actually the microscopic causal graph. And maybe we can agree that no agent can possibly be the cause of its own actions. But when there is a degree of planning sophistication macroscopically it's as if it's the cause of its own actions.

Dr. Jeff Beck 00:09:25

Yes. And that's why this as if phrase comes up a lot. Right? I mean, it's it's important to remember that, like, no matter how clever your model is, and no matter how clever your approach is, and how clever the words are that you use to describe it, a lot of this stuff is is is as if. Right? This is this is the best model. Right? It's not that it's not this is why, like, I I I repeat this over and over again, grind it into the students. Right? Is that that, you know, science is about, like, prediction and data compression and nothing else. And the same thing is going on here, right? You'll never, just looking at behavior, you'll never know for sure in any meaningful way whether or not it's just doing a function transformation or whether it's engaged in planning and counterfactual reasoning. But if your best model of it, if you sort of say, well, tried to model as a function transformation, but goddamn it had a lot of parameters. Right? But then I tried to model it as something that was just doing Monte Carlo tree search on the inside and giving the answer, and that had like, you know, 40 parameters. And it's that's the model I'm gonna go with and now I'm gonna call it an agent.

Dr. Tim Scarfe 00:10:27

If we had a physical agent in the real world that was doing all of this planning and so on, would that have some kind of primacy to a computer simulation of agents that were doing all of this

Dr. Jeff Beck 00:10:38

Oh is like if I uploaded my brain onto a computer and didn't connect it to the world would it still be thinking even though it's like doing all of those things? Is that the idea here or am I That that works.

Dr. Tim Scarfe 00:10:49

So yeah let's say a high fidelity computer simulation of Jeff. Would would would Jeff be an agent? No. Oh. I wasn't expecting to say that.

Dr. Jeff Beck 00:10:57

Because I'm the agent. And if you upload it, no, I don't know. So if you do a high fidelity computer simulation and you put it in my body, then I think I would have to say it's an agent. Yeah. Right? If it's doing exactly the same I mean, is like the standard. It's doing exactly the same calculations from from a purely like phenomenal audio perspective. It's like it's the same. It's indistinguishable.

Dr. Tim Scarfe 00:11:19

Okay. So agents need to be physical. So I do believe

Dr. Jeff Beck 00:11:21

that an agent needs to be physical. Absolutely. I don't believe, you know, I I believe you can have a model of agency and not have an agent. Right? I, you know, you can put that model in a computer and run it and make predictions as to what an agent would do. You and it might even be a 100% correct, but I still wouldn't call it an agent. But again, this is like getting into philosophy. And like philosophy frustrates the Bayesian because philosophy is not probabilistic. Right? Philosophy is really about drawing clear lines and distinctions. And in my world, those don't really exist. Right? There's everything has an error bar. You know? All of there isn't a clear delineation between, you know, you know, an object and an agent. It's really, you know, in from this modeling perspective, it's really just a question of degrees, And philosophy is terrible at handling questions of degree.

Dr. Tim Scarfe 00:12:12

My friend Keith, he's a big fan of computability theory. And he thinks that an agent is basically a type of computation. And it has access to ambient state and it can take action and there's this kind of like cybernetic loop. And for him, the strength of the agency in the system is the compute type that the thing is doing, right? So if it's a finite state automata, then it's a weak agent. If it's a Turing machine, it's a strong agent.

Dr. Jeff Beck 00:12:41

Yeah, so the degree of sophistication of the compute. Pretty much.

Dr. Tim Scarfe 00:12:45

Yeah. Does that ring true to you?

Dr. Jeff Beck 00:12:46

I mean, if forced me you know, at the point of a gun to put a measure on agency, it'd probably look a lot like that.

Energy-Based Models & Test-Time Training

Dr. Tim Scarfe 00:12:55

Yes. Jeff, let's talk about energy based models. Sure. So Yan Lakun, he had a monograph out, I think in 2006 talking about this. Oh, talking about this for a long time.[3]

Dr. Jeff Beck 00:13:04

Oh, yeah. When you fit your neural network to data, you know, via gradient descent, right, then you have written an energy function in weight space, and you are fall and you're following it to its energetic minimum. You know, the advantage of using an energy based taking an energy based approach as opposed to taking, say, a straight up function approximation approach is that an energy based model comes with something that's kind of like an inductive prior. Right? It basically an energy based model if you're doing function approximation, you're basically saying, there's any mapping from x to y. X is my inputs, y. But any mapping is out there. I just want to figure out what it is. Right? Now in an energy based model, you're you're you're you're effectively placing constraints on what that input output relationship can be. I like thinking about the distinction between an energy based model and a and a traditional sort of feed forward neural network has to do with where your cost function is applied, right? So in a traditional neural network, you take in your inputs, you got your outputs, and the cost function is just a function of the inputs and the outputs. And the only thing that you're optimizing is the weights. In an energy based model, there's another thing that your cost function operates on, and that's something 1 of the internal states of your model. And as a result, like in order to figure out what the best approach is, right, you actually have to do 2 minimizations. 1 that that finds the energetic minimum associated with the the the part of the the cost function that operates on the internal states, like the hidden nodes of your network. Right? And then 1 that is the prediction. That is your like effective prediction error. This is very much consistent with the approach that a Bayesian would take. You a prior probability distribution, which gives you an energy function over every single latent variable in your model. And you are optimizing with respect to all of them. So you take a probabilistic approach. Good examples of this are like a variational autoencoder. A variational autoencoder, I think, is the best example of the most commonly used energy based model out there. Why? Because you have an encoder network. You have a decoder network. And your cost function is based

[3] A Tutorial on Energy-Based Learning (LeCun 2006) — Paper Tim references Yann LeCun's 2006 monograph on Energy Based Models.

on the difference between inputs and outputs. So that's just like a yeah. It's fine. That's still a regular. But it also is how how Gaussian. And it well, it depends on what flavor of VAE. But you also have some some some part of your cost function is a function of the actual rep internal representation. Right? In a traditional VAE, it's it's how Gaussian is. You want that internal representation to be as Gaussian as possible. If it's a VQ VAE, then it's like mixture of Gaussian. But it's still like a cost function that is applied on the internal states as well as on the inputs and outputs.[4]

Dr. Tim Scarfe 00:15:38

Very cool. So a VAE is is a fairly canonical example of an energy based model. Yeah. And what you were saying about the I mean, you know, the whole DL world is obsessed with test time inference at the moment. And in a way that that is a step towards what you're talking about.

Dr. Jeff Beck 00:15:52

Yeah. You're treating a certain yeah. You're treating some of the weights of your model. Right? I mean, well, yeah. You're treating some of the weights of your model as if they're latent variables. Right? Because when you you when you show a new input, right, you're allowed to change some of the weights without looking at the output. Right? And so what are you doing? Well, you're treating the weights as latents. Now I think that like which makes it a great trick in my opinion. It's like, oh great. I gather they're they're moving in the direction of energy based models. I love it. The only thing I don't like about test time training is the vast majority of the training that is done. So in a traditional energy based model, you always find the minimum with respect to the latent variables, like these extra weights, which in the case of test time training is the subset of weights that you're allowed to change during test time. When you do the training for a traditional energy based model, you're allowed to make those changes, right, throughout the entire course of training. The way that we're often doing test time training these days is we just do regular old neural network learning like we don't do and and then and then and then finally when it comes to when we get to the deployment phase, then we suddenly turn on, right, these these additional latents which are basically some of the weights of the network, and we do additional an additional bit of learning at that point. This seems monument now again, not an expert here, right? But this seems unwise to me. And the reason it seems unwise is because you didn't train the original network with that on. Right? You trained it as in a completely supervised way. Yes. Now I'm sure that people are aware of this and it's been addressed in the literature, but I'm not

[4] Auto-Encoding Variational Bayes (VAE) — Paper Jeff cites VAEs as the canonical example of an energy-based model used today.

personally aware of that. I don't think that's how it's used in practice.

Dr. Tim Scarfe 00:17:29

We should also introduce this term transduction. So my definition of transduction is that you're actually doing search or optimization as a function of the test samples. Like I interviewed Clement Bonnet, had a VAE on ARC, searching latent spaces. And he actually searched through the decoder as a function of the test sample. And because these models, they are maximum likelihood estimators, right? Which means they're always giving you a kind of smoothed out average. And there's so much information in the test sample. Let's just riff on the relationship between energy based models and Bayesian inference. So of course, they have this advantage that you don't need to do this for expensive intractable normalization. Yes. Yes, tell me about that.

Bayesian Inference & Free Energy

Dr. Jeff Beck 00:18:11

My take on it is is that an energy based model and a Bayesian model have a lot in common. Right? In many ways like energy, I mean, well literally in physics, right? Energy is like log energy is log probability. Now, of course, there's the normalization, you know, factor that you don't need to worry about if you're just doing if you're just minimizing energy. And so the difference between, you know, like which is sort of like, you know, in a Bayesian framework, that's like saying, well, know, I'm not actually gonna treat some of these latent variables in a probabilistic way. I'm just gonna do maximum or map estimation on some of my variables and just be okay with that. And that's 1 way to interpret the relationship between an energy based model and a properly Bayesian model. There's there's a happy medium here, though. Right? And the happy medium is you can still treat it as if it's, you know, you know, you don't have to just minimize the energy function. But you can calculate the curvature down there too, do a Laplace approximation, and call yourself a Bayesian again. Right? Yes. There is more computation involved, but we've got a lot of great tricks for making that totally tractable.

Dr. Tim Scarfe 00:19:12

What's the relationship between the free energy and the free energy principle and the energy and energy based models?

Dr. Jeff Beck 00:19:18

Regularization term, I think is the short answer. Right? No. So so the interesting and and and if you're being very, very, very pedantic, the difference between an energy based you know, minimizing energy and minimizing free energy is that free energy has this additional entropy penalty term. Now if you're just doing maximum likelihood estimation, if you're minimizing your energy function with respect to some particular well, just we'll just pretend we're only worried about 1 variable. And I'm just gonna, like, get a point estimate and call it a day. Do, like, you know, some kind of map estimation to get to get that that 1 thing. There's not that big of a difference, right, because you're you're not there is no probability distribution over the latent that allows you to compute that regularization term. But that's the only difference. It's it's are you regularizing or not, is I think the easiest way to think about it.

JEPA, Latent Space, & Non-Contrastive Learning

Dr. Tim Scarfe 00:20:07

So Lakun is a big advocate of JEPA. So the joint embedding prediction architecture is using non contrastive learning where essentially the learning objective is comparing the latents of observed and unobserved parts of the space. This is an architectural design. Well, is Okay.[5]

Dr. Jeff Beck 00:20:24

So what does JEFA stand for? It's joint embedding Prediction. Prediction architecture. There we go. So what's the joint embedding bit about? Well, the joint embedding bit about is is, know, is, well, I'm gonna take my inputs. I'm gonna take my outputs, and I'm gonna embed them in some space. Right? And then I'm gonna learn a prediction between the 2 embeddings. And that's a great idea. It's a great idea because it has some of the flavor of what we would like to get out of our models. Like, we're not interested in predicting every in many situations, I should be very particular about this. In many situations, we're not interested in predicting every single pixel on the image. We want to get maybe something that's a little more gestalt, a little more high level, a more conceptual understanding of what's going on. And so emphasizing the goal of predicting every single pixel, which is what's typically done in generative modeling right now, might lose some of the power, the abstractive power of some of the networks. And so like, let's do it. So so the whole point of JEPPA, as I understand

[5] JEPA (Joint Embedding Prediction Architecture) — Paper Discussion of LeCun's JEPA architecture and non-contrastive learning.

it, I'm sure there are other points, is that is that you're gonna take you're you're gonna you're gonna compress your inputs and compress your outputs, and then do all the learning in this compressed space. Love it. Right? Science is about prediction and data compression. Let's make that compression explicit on the front end and the back end. The downside of this approach is that is it is it it doesn't work out of the box. Right? It because it's very easy to find a compression or an embedding of the inputs and an embedding of the outputs for which prediction is perfect. Which is to basically make both of them 0. And so you have to do some other thing. Other tricks need to be employed in order to make it work.

Dr. Tim Scarfe 00:21:55

Yes. Yes. I remember LeCun was talking about this. So there's the traditional contrast method, which is from it's kind of Hinton's idea apparently of the negative sampling and whatnot. And that's very expensive because you actually have to do lots and lots of sampling. This non contrastive thing Yeah.

Dr. Jeff Beck 00:22:13

This, by the way, what he should have won the Nobel Prize for.

Dr. Tim Scarfe 00:22:15

Right.

Dr. Jeff Beck 00:22:17

In my opinion. Yes. Because the whole point of the wake sleep algorithm and contrasted divergence was that, oh, it's actually biologically plausible. Right? It was an end run around the need to do back prop. And that's what made it so clever and interesting in my opinion.[6]

Dr. Tim Scarfe 00:22:32

Lakun is a big fan of this non contrastive thing where you work in the latent space. There are many different algorithms that do this. Had a whole load of shows all about non contrastive learning. There's things like VCREG and BYOL and Barlow Twins. And there's an entire thread of research all around that. And in many different ways, what they're trying to do is avoid this motor collapse problem that you're talking about. And they use different forms of regularization. There's an old school way of accomplishing the same thing.[7]

[6] The Wake-Sleep Algorithm — Paper Jeff mentions Hinton's Wake-Sleep algorithm as a biologically plausible alternative to backprop.

[7] Barlow Twins: Self-Supervised Learning — Paper Tim lists Barlow Twins among other non-

Dr. Jeff Beck 00:22:59

And that is to do all of your, it's called pre processing, right? And this is something that a lot of people do. You take your data, in fact we do this all the time with like vision language models, Right? So we wanna do, we wanna use an LLM and we wanna predict images. So what do we do? Well the first thing we have to do is tokenize the image. Right? And so what do we do? We run a VA that we do the preprocessing. And we do it by the preprocessing step is completely independent, right, from the actual algorithm that's gonna be tasked with solving the problem of interest. And that's not something that we necessarily have to stick with, right? It would be very nice if there was a way of like, again, jointly, we're getting right back to JEP again. What we'd like to do is we'd like to choose our preprocessing algorithm in a manner that that, you know, not a priori, not do it first. We'd like to choose the preprocessor that works the best in in this space. Yep. And I think that that's the ultimate motivation for a lot of this work is that it's like, what's the right embedding? 1 of my favorite tricks of course, I pre process the VAs all the time. In fact, time someone hands me a new neural data set, the first thing I do I'm not ashamed to admit, I run PCA on and pass it through a VAE, and then sort of take a look, right? It's the first thing you do with your data, because it gives you a good idea of what the signal to noise ratio is in the data set itself. Yes. And then I yeah. And then what do I do? I subsequently do most of my analysis, right, in that discovered embedding space. And there's I I I I don't see a huge problem with that from a purely pragmatic perspective, But it's certainly cleaner, right, to have a single algorithm and approach and not just be stringing these sort of things together in an ad hoc way. There's, you know, when doing PCA, PCA is a really great example of this. There's a failure mode for principal component analysis, which is actually really common in neural data. Because principal component analysis basically says, well, where's the most variability? Okay, I'm worried about that. And then all the stuff that's not varying very much, I'm just going to throw it away. Just like dimensions in which there's low variability are not important. Well, turns out that in neural data, the dimensions in which there's very little variability are some of the most important dimensions. Yes. And so preprocessing with PCA runs a risk of throwing out the most valuable information in your data set. Yes. And so there's a lot of wisdom in jointly fitting your preprocessing model as well as your inference and prediction model.

contrastive learning methods (BYOL, VICReg).

Dr. Tim Scarfe 00:25:45

I mean, this subject of not throwing things away, JEPR and non contrastive learning, it's part of this bigger field of self supervised learning. And we want to learn representations that maintain fidelity and richness. And Lakun's hypothesis is that when you do something like supervised learning with some particular downstream task in mind, the neural network gets wise. And what it does is it kind of discards all of the long tail stuff that aren't relevant for that particular task. So when you train these models, you're trying to do is sort of maintain enough ambiguity so that it compresses the information, but it also maintains enough fidelity to work broadly for different things.

Dr. Jeff Beck 00:26:24

Yes. And that is a laudable goal, right? And I certainly share it, right? The last thing you wanna do is I mean, you know, fortunately, like, networks are so big, we don't really run the risk of of, like, overfitting so as much as we used to. But the last thing you wanna do is throw is is train your network to toss information that you might need down the road. That said, the vast majority of what, you know, the brain does just like these neural networks is decide what information is currently task irrelevant. But that's all the more reason to do things in a self supervised or unsupervised way. Right? Because you're basically not telling it this is the important. You know, you're not telling it like what's all task relevant and task irrelevant.

Evolution of Intelligence & Modular Brains

Dr. Tim Scarfe 00:27:07

So I interviewed about the version 2 of the ARC challenge. And 1 thing that struck me is I think of intelligence as being multidimensional. So version 1 got saturated. The ARC was actually really amazing because it's the only intelligence benchmark that has survived for 5 years before being defeated. Since the advent of these thinking models, it has been defeated very quickly. But they're working on version 3, and there'll be version 4, there'll be version 5. Will there always just be something left over?[8]

[8] ARC Prize (Abstraction and Reasoning Corpus) — Challenge Tim discusses the ARC challenge as a benchmark for intelligence that resisted saturation.

Dr. Jeff Beck 00:27:38

That sounds like another philosophical. So yes is my answer. There will always be there will always be something left over. In the sense that like, you know, you know, we we we have this this has been the trajectory things have been going for a really long time. Right? It's sort of like, we get algorithms that do amazing new cool things, and then someone comes along and says, yeah, but it can't build me. It it can't pull a rabbit out of a hat. Right? And then and then of course, what does someone do? They oh, they they figure out the new training protocol, a slightly different architecture, or they just train it to pull rabbits out of hats, and then suddenly it can. And then someone proposes a new challenge, and a new challenge, and a new challenge. And it's always this game of like 1 upmanship. So the question becomes, well, what's the point at which there are no more new challenges? I'm not entirely certain we're ever gonna get there. Right? It may very well be the case that we get, you know, these sort of algorithms that are capable of replicating the complete suite of human behaviors, and then someone will come up with some criticism like, yeah, but it's not really doing x, it's just faking it. Right? This is just the direction things go because people really do think they're important.

Dr. Tim Scarfe 00:28:42

Yeah. Do you think that the concept of recursive self improving intelligence is a valid 1?

Dr. Jeff Beck 00:28:48

Yes. I do think that is So I think that 1 of the most critical missing elements right now is some form of continual learning. Right? At the end of the day, you really want an algorithm that doesn't just learn on the training set and then just gets deployed. You want something that that that runs around in the world and comes across things that it doesn't understand. Right? And then is able to incorporate to build, you know, append its model in some sense. Right? So this is like the this is, you know and there are some approaches to this all based on, like, Bayesian nonparametrics and Dirichlet process priors and stuff like that, where you you sort of see something that's surprising or unique or different, something you didn't expect, and it causes you to say, I need to turn learning on because I gotta figure this out. That is an absolutely critical element that we need to be developing. We are developing that. And it turns out that that's 1 of the nice things about this sort of object centered physics discovery thing is because it's object centered, if it comes across a new situation that it does not understand, it is capable of instantiating a completely brand new object just to explain this new situation. Continually learning agents can acquire new knowledge autonomously.

Dr. Tim Scarfe 00:29:56

And the whole thing just learns more knowledge. But intelligence feels different. It feels like in the system that we've been describing, the intelligence is the way we're implementing the Bayesian updates and actually building the algorithms. Could the systems on their own metaprogram themselves and develop better algorithms or something like that?

Dr. Jeff Beck 00:30:19

That's a very good question. Something that would be closer to true artificial intelligence than what we currently have would be capable of building models on the fly to deal with new situations, to taking things that it knows about, right, and combining them in new and different ways. There are approaches that have some of that aspect to it, like gFlowNets from like Benji and stuff is like is like a great example of something that at least in principle is a generative model of generative models. Right? It's sort of like, oh, like, you know, I might actually need a new node. Like, it's time to create a new latent variable because like like, the current set's just not cutting the mustard anymore. Those are things that that that I think are hallmarks of of true intelligence. I don't wanna ever make the statement, as soon as it's got that, it's truly intelligent. I will never ever ever say that. But I do think that that is a critical component that that needs to be present. Right? Is the ability to generate new models on the fly to deal with novel situations and data. Most of that, you know you know, as well as the ability to combine old models, previous models in new and interesting ways. This is actually how the brain evolved, right? We started out with like, really simple brains, and there were different regions, and they solved sort of different problems. And what eventually happened as we evolved is that these different regions of the brain learned to communicate with each other in new ways, and through that communication acquired new abilities, right, and then eventually evolved into new capabilities and things like that. I often like to point out to the I think olfaction is like the sense that's not studied nearly enough. It's an incredibly old part of the brain. And arguably, right, it's the first part of the brain that evolved the ability to do proper associative processing, right? The odor, unlike visual space, right, where there's translation symmetries and all that sort of stuff and things are smooth, Olfactory space, that does not exist, right? It's really, really, really combinatorial and complicated. And the part of the brain that evolved to solve the olfactory problem arguably is the part that evolved into our frontal cortex. Don't quote me on that. There's a lot of disagreement there. That's just my take. But it certainly has a lot of the features that we associate with associative cortex, right? Wow, I just said, like, 6 uses 3 different uses of the word associate in that sentence. But but I think you see what I mean. Right? It it it

was all about, like, taking old capabilities. Right? Combining combining, you know, simple models and modules to create something that was more complex. And then over time, right, so that was what made the brain work, right? It was all about taking little things that worked and combining them in new and different ways in order to evolve, you know, effectively an emergent, you know, emergent properties, emergent, you know, computational abilities, and an emergent understanding of the world in which we live. And I do think that, like, what what, you know, if when we get to the point where we start really saying, oh, this is actually truly intelligent, it's going to have that feature. It's going to have the ability to have a it's gonna have a modular description of the world, and it's gonna have the ability to to combine those modules in a way that creates a more sophisticated understanding. It's like Legos. Right? I can you know, the the Lego bricks all connect in certain ways, and I can build, all sorts of new and amazing things that were never built before, right, out of them. That's the capability that we have. And that's the essence of like creativity. It's why I refer to systems engineering as like the thing we really want our AI models to be able to do.[9]

Scientific Discovery & Automated Experimentation

Dr. Tim Scarfe 00:34:00

Collective intelligence is a bit different. We have this plasticity, right? We can adapt our behavior day by day. We might see some kind of meta learning or some kind of change in our organization dynamics. Maybe some agents will specialize. And it might be an existence proof of this kind of recursive superintelligence that we're talking about.

Dr. Jeff Beck 00:34:19

Yeah. I think that's absolutely correct, right? So the specialization is great. In fact, I would argue that specialization is how we got all of this, Right? And this was I'm pointing at London in case you there was some confusion there. Right? It was it was really about, you know, the interconnected, highly specialized intelligences that are people and their ability to learn how to to to work together that that that, you know, gave rise to the technological revolution. The brain is the same way. Right? It's in my view. It's highly specialized little modules or agents that are capable of of of of being repurposed, reused, capable of communicating with 1 another in order to solve really complicated problems. But there's always

[9] GFlowNets (Generative Flow Networks) — Paper Jeff mentions Bengio's GFlowNets as an example of models that can generate models.

a benefit to specialization. I don't believe in, like like AGI. AGI seems like a bit of a misnomer to me. What we really want is not artificial general intelligence. We want collective specialized intelligences.

Dr. Tim Scarfe 00:35:17

What about scientific discovery? Do you think that we could what would the world look like when we could discover new drugs, we could discover new knowledge in science?

Dr. Jeff Beck 00:35:25

Right now, the way that we're doing that is largely focused on summarizing vast troves of data and looking for correlations that are present in it. I think the next major milestone in this trajectory is experimental design. Not just, oh, well, here's some correlations you may not have seen because they're really small, and this is what computers are good at. They're really good at identifying small but highly relevant correlations. And the next step, of course, is constructing a system that tests these hypotheses explicitly, right, and generates the experiments that will identify like, that will fill in the gaps of our knowledge. And all of this, I believe, can in fact be automated in a very sensible way. I I don't, you know, I don't I don't see any, like, major obstacles to automating empirical inquiry other than we probably wanna place some safety constraints when we start letting them work when we start letting the AIs run the labs. Right? Because you never know. It's sort of you always have this AI. Was like, well, you know, the most effective experiment to determine if this is correct is to set off a nuke. And that that would be bad. Yes. Right? So pure empirical inquiry, right, does run risks like that. But I think that that's not not not the biggest issue. I think what we need to do is we just had need to have a nice concise framework for saying like, oh, look. You know, like, I'll give you an example. So we had the we we we had this problem that popped up a while back. A gentleman we were talking to is you've got these you know, you got these robots and the robot sees something it's never seen before. And and and I, you know, so a robot is like running around, it comes across like a beach ball. Never seen a beach ball in its entire life. And what you'd like is you'd like the robot to know how to figure out that it's a beach ball and to figure out what its properties are. And if you tell the robot, like like, if you see something new, just stop. Right? You're kinda then that's that's no good. Right? What you really wanna do is you wanna figure out a relatively noninvasive procedure for the robot to like poke do what a child would do. What does a kid do when they see a beach ball? Right? They run up and they poke it, and they say, oh, right. Yeah. And then it moved, and it it actually learned it actually experiments with its environment for the purposes of identifying the properties of the objects that

exist in it. Mhmm. Now I do think we probably wanna test this out virtually before it's deployed in the real world, because you never know. It might very well be that optimal experiment is to run up and kick it as hard as you possibly can. And we we certainly wanna avoid that. But, like, something along those lines. Something you know, a robot that is able to test the theories that it has about how things work in an online way and learn from those results in an online way is definitely part of the goal.

AI Safety, Enfeeblement & The Future of Work

Dr. Tim Scarfe 00:38:04

Looking forwards, what do you think the future will look like when we have more autonomous AIs among us? A lot of people worry about enfeeblement, loss of control, know, it making us dumb, all of this kind of stuff.

Dr. Jeff Beck 00:38:16

I do worry about AI making us dumb, right? I mean, offloading your thinking onto a machine, which is something that AI allows, is potentially a big problem. I don't really want to have a situation where humans are reduced to, like, value they're just reduced to, like, value function selectors. They're just basically going, oh, no, I don't like that outcome. Like, do this instead. I do want to see a future where where where we have an AI that actually improves our understanding of the world. And simply automating everything runs the risk that you specified. Right? It runs the risk of people becoming couch potatoes that just watch TV and occasionally say, like, yeah, you know, these chips are no good. That seems like a bad outcome to me. I worry less about that, I think, than some because people are remarkably adaptable. Right? I mean, know, they have all these arguments about like, oh, you know, this new technology comes along and it's gonna completely destroy this way of life. And you know, and that's gonna be awful for people. And it is maybe in the short term. I think of like tractors, right? Or just go back. How many hundred years do you have to go back when like 99% of people were involved in agriculture? And now it's like, what, 2? Right? I consider that a solid improvement, right? Because it allowed the rest of us to it allowed us to do a bunch of other things that we find more satisfying, that are more interesting. It allowed us to I can read spend some time reading a book, don't have to labor in the fields all day. That's the future that I sort of see. And that's the future that I hope for, is that is is 1 in which, you know, all of these artificial agents running around and doing things autonomously are there to to free us up to pursue more interesting and more you know, you know, to improve ourselves in in in in in more interesting ways. But at the end of day, it's just another techno you know, at least initially, it'll just be another technology like the tractor. Now, 100 years from now, who knows?

Dr. Tim Scarfe 00:40:18

What will the value of work be if the AIs can do everything and there's nothing left for us to do?

Dr. Jeff Beck 00:40:23

I don't think that will ever be the case that the AIs can do everything. Like I said, the future I worry about is 1 where like it's, you know, the the sole role of people is like sitting around like making sure the AIs aren't aren't going rogue and and and things like that, which I don't consider a good outcome. I would really like to see human improvement. You know, I I I envision a future of, I don't know, this is like cybernetic transhumanism, if I'm gonna go sci fi on this, right, where where, you know, the technology and us evolve together in a way that's beneficial for both. That's the goal. You know, are there these dystopic possibilities where like, oh, well, what are humans in a world where, well what are they? What are what are humans in a world where everything can be done by a robot? Yep. You know, that's that's a good question. And that's and at the end of the day, right, they end up just becoming like reward function selectors. Right? They end up just sort of saying, oh, I don't like this and I do like that, and they're basically, you know, I mean, you end up with a sit it is another nightmare scenario. I don't like talking about these dystopian futures because honestly, I think people are too clever, and I think people are too motivated, and people are too interested in how the world really works, and then people are too interested in actually understanding things that they will never stop. That AI will become a partner, not an adversary or a crutch. And that's that's that's what I think will happen because that's but that that's a statement more about my belief about humans than it is about my belief about the development of AI. You know, am a techno optimist, if you will. Mhmm. Not a not a pessimist. I believe that we will find a way to adapt to an ever changing world as we have done for millions of years, including 1 that includes technology that alleviates most of our labors.

Dr. Tim Scarfe 00:42:09

On that, there's an AI literacy thing because AI has moved so quickly now that certainly my parents don't understand anything about it. But by the same token, policymakers don't understand anything about it. And there are people saying AI is going to kill everyone. And there's people making negative arguments. There's people making positive arguments. There's a bit of a fog of war now because there are so many people saying different things about AI. How should they make sense of

Dr. Jeff Beck 00:42:31

all of this? We are now well outside my area of expertise, so I'm just gonna say that before I say anything else. AI is developing very quickly, But I am much more concerned about what people will do with the new technology than I am with what the technology will do all by itself. I don't have this big concern about, I don't really believe that like Skynet's gonna take over, the internet's gonna suddenly become conscious and kill us all. Right? In part because AI is not that advanced, but also because we are telling, we are still in the position where we specify the goals of the system. And that will likely continue for a very long time. And it will always be the case that these systems, you know, will can be, you know, are are subject to review. We will always keep an eye on them. They will always at least initially be be be released in relatively restricted domains and where we're where we're test where where we're keeping a a close eye on what it is that they are and are not doing. So I don't worry too much about, like, the going rogue. I worry a lot more about somebody building you know, it's sort of like a virus, which we already have to deal with. Like, somebody builds, like, some insane virus and, like, takes down the Internet. I'm more worried about malicious human actors than I am malicious AI actors because at the end of the day, all of these algorithms, they simply do what they are told. Right? We train them. We tell them here's your objective function. As long as we are specifying the objective function and we understand the objective function, we're probably going to be okay. I think the safest way to deal with AI concerns is to tell people, hey, look, this AI is just doing what we told it to. We set it up to make really good predictions and to achieve these outcomes. Now is it dangerous to specify these outcomes without being very, very, very careful? Yes, it is. This is the whole, hey, Skynet, end world hunger, and it kills all humans. Is a real possibility, but whose fault was that? The fault was the person who, like, was very, very naively specified their goals. There are, in fact, relatively straightforward ways to specify the the reward function that that don't run that risk nearly as badly. And the best 1 is so are you familiar with maximum entropy inverse reinforcement learning? I like to call it active inference because it's really similar. And so there, what you're doing is you're basically observing someone's policy, and then you're trying to do a maximum entropy model, you're doing maximum entropy model on the reward function itself. At the end of the day, what ends up happening when you do this is, this is why it's like basically just like active inference. You get a reward function, so you have some, you know, organism or whatever and you're trying to do this for it, it's got some stationary distribution over actions and outcomes, right? It's inputs and outputs of the stationary distribution. That becomes your reward function. Like not directly, there's some math involved, but basically your reward function is a function of the steady state distributions over actions and outcomes. So we could do this, right? We could take the current manner in which humans are making decisions, and we could write down, right, what's the stationary, what is the current estimate of the stationary distribution over actions and outcomes? So this would include things like everyone's getting, this number of people are going hungry, this, you know, and and, you know, all of the stats that describe like the inputs and outputs to our policy make, you know, to our policy. Then we could just ask an AI, your reward function is the 1 that results in the same outcome that we currently have. Right? On average. And it would execute it and it would and and to the extent that it works. Right? It it it would it would ultimately result in a in an AI algorithm that just sort of is like mimicking human behavior, right, or at least achieving the same outcome that we were achieving before. Now here's the safe way to like improve the situation. You don't say end world hunger, right, you perturb that distribution over outcomes, right, and just just over outcomes a little bit, and then you evaluate the consequences, right. It's it's all you're doing. You make these little changes in the reward, in an empirically estimated reward function, right, rather than just sort of specifying 1 by hand because that's the dangerous thing.[10]

Dr. Tim Scarfe 00:46:46

Jeff, thank you so much for joining us. Yeah. It's my pleasure. Amazing.

[10] Maximum Entropy Inverse Reinforcement Learning — Paper Jeff proposes MaxEnt IRL as a safer way to align AI by observing stationary distributions.

References

[1] **Free Energy Principle (FEP)**
`https://en.wikipedia.org/wiki/Free_energy_principle`
Concept Jeff mentions being an ”FEP purist” regarding agents vs objects.

[2] **Monte Carlo Tree Search**
`https://en.wikipedia.org/wiki/Monte_Carlo_tree_search`
Concept Mentioned as the internal planning mechanism that could be hidden inside a black box.

[3] **The Intentional Stance**
`https://mitpress.mit.edu/9780262540537/the-intentional-stance/`
Book Tim brings up Daniel Dennett's concept of the ”Intentional Stance” as a useful explanation level.

[4] **A Tutorial on Energy-Based Learning (LeCun 2006)**
`http://yann.lecun.com/exdb/publis/pdf/lecun-06.pdf`
Paper Tim references Yann LeCun's 2006 monograph on Energy Based Models.

[5] **Auto-Encoding Variational Bayes (VAE)**
`https://arxiv.org/abs/1312.6114`
Paper Jeff cites VAEs as the canonical example of an energy-based model used today.

[6] **JEPA (Joint Embedding Prediction Architecture)**
`https://openreview.net/forum?id=BZ5a1r-kVsf`
Paper Discussion of LeCun's JEPA architecture and non-contrastive learning.

[7] **The Wake-Sleep Algorithm**
`https://www.cs.toronto.edu/~hinton/absps/ws.pdf`
Paper Jeff mentions Hinton's Wake-Sleep algorithm as a biologically plausible alternative to backprop.

[8] **Barlow Twins: Self-Supervised Learning**
`https://arxiv.org/abs/2103.03230`
Paper Tim lists Barlow Twins among other non-contrastive learning methods (BYOL, VICReg).

[9] **ARC Prize (Abstraction and Reasoning Corpus)**
`https://arcprize.org/`
Challenge Tim discusses the ARC challenge as a benchmark for intelligence that resisted saturation.

[10] **GFlowNets (Generative Flow Networks)**
`https://arxiv.org/abs/2111.09266`
Paper Jeff mentions Bengio's GFlowNets as an example of models that can gen-

erate models.

[11] **Maximum Entropy Inverse Reinforcement Learning**
`https://www.aaai.org/Papers/AAAI/2008/AAAI08-227.pdf`
Paper Jeff proposes MaxEnt IRL as a safer way to align AI by observing stationary distributions.