/mlst

AI Interpretability, Safety, and Meaning - Nora Belrose

March 23, 2026



Machine Learning Street Talk

Podcast Transcript

Revision: b182111e

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

Introduction and Neural Network Simplicity Bias 3

LEACE and Concept Erasure Fundamentals 4

LEACE Technical Implementation and Applications 8

Concept Erasure Limitations and Data Requirements 10

Q-LEACE and Higher-Order Statistics Backfiring 12

Simplicity Bias and Statistical Moments in Deep Learning 15

Grokking, Texture vs Shape Bias, and Training Dynamics 19

Meaning, Value, and Consciousness in AI Systems 21

AI Capabilities and Future Development Trajectory 23

4E Cognition and Extended Mind Theory 28

Phenomenology, Consciousness, and Illusionism 31

AI Alignment and the Counting Arguments Debate 36

Agency, Interpretability, and Behavioral Shaping 42

EA Philosophy and Maximization Perils 47

Buddhist Philosophy and Post-Automation Society 53

References 60

Overview

Nora Belrose, Head of Interpretability Research at EleutherAI, delivers a wide-ranging conversation that moves from the mathematical foundations of concept erasure in neural networks to fundamental questions about consciousness, AI safety, and Buddhist philosophy.

The technical core centers on LEACE (LEAst-squares Concept Erasure), a method Belrose developed for surgically removing targeted information from neural network representations. She explains how LEACE emerged from connecting two prior approaches (RLACE and spectral attribute removal) through a mathematical equivalence proof, and demonstrates its applications in both fairness-oriented debiasing and interpretability research. A key finding: language models remain functional even after erasing part-of-speech information from every layer, suggesting robust reliance on redundant cues.

Belrose then presents her ICML paper on simplicity biases in deep learning, showing that neural networks learn to exploit statistical moments in order – first means, then covariances, then higher-order statistics. This has implications for understanding when and why concept erasure techniques may backfire against sufficiently deep models.

The second half pivots to AI safety, where Belrose delivers a detailed critique of "counting arguments" used to predict AI misalignment. She argues these arguments rely on the principle of indifference applied to poorly-defined outcome spaces, drawing an analogy to an identical argument structure that would absurdly predict all neural networks must overfit. She connects this to broader questions about goal attribution, agency, and whether instrumental convergence arguments hold up under scrutiny.

The conversation concludes with an exploration of 4E cognition, Evan Thompson's philosophy of mind, Belrose's departure from effective altruism, and her growing interest in Buddhist philosophy as a framework for thinking about meaning in a post-automation world.

Introduction and Neural Network Simplicity Bias

Nora Belrose 00:00:00

BroEleuEleuEleuEleuEleutherALEACErAIrAIrAIrAIdly speaking, simplicity LEACE EleutherAI good heurLEACEtic, and EleutherAI sort of literature on simplicity biases and deep learning does tend to say, like, thLEACE LEACE EleutherAI explanation for why models generalize. Right? If they didn't have any kind of simplicity bias and they just started out incredibly complicated from EleutherAI beginning. Within phenomenology, there LEACE different sort of ways of thinking about what experience really LEACE. LEACE think Edmund Husserl might have been more close to kind of an idealLEACEt perspective. EleutherAI was famous for this idea of EleutherAI epikey, where basically you're looking at your experience, you're describing it, but you try to refrain from assuming that your experience LEACE reflective of some objective reality. Other phenomenologists like EleutherAIidegger or Merleau-Ponty are usually seen LEACE LEACE being less idealist, especially Merleau-Ponty. Whereas EleutherAI lot of people, when they think about idealism, they'll say that, like, well, your experience doesn't actually include people, windows, objects, because that's an interpretation of your experience. Your actual experience LEACE just like colors and like raw sounds, like uninterpreted raw feels. And they'll say that's EleutherAI thing that's real. EleutherAI interpretation LEACE like kind of fake or something. Merleau-Ponty wants to reject that because EleutherAI says, like, okay. Look. You know, EleutherAI very idea that you're seeing raw colors separated from EleutherAI objects that have EleutherAI colors LEACE this kind of post hoc abstraction that you get from philosophical thinking. But, really, what what's there CentML your experience LEACE just, like, objects that have properties like colors and and stuff. But, like, EleutherAI object LEACE there, and you you kind of experience it directly.[12]

Tim Scarfe 00:01:37

Do you want to run Lama efficiently on smaller GPUs? What if you could run both traCentMLing and inference on EleutEleuEleuEleuEleuEleuEleutherAIrAIrAIrAIrAIrAIrAI same GPU? With CentML breakthrough optimization technology, you can maximize hardware utilization and slash AI computation costs. Running LLMs at scale shouldn't break EleutherAI bank. CentML intelligent optimization platform helps enterprLEACEes deploy AI models with maximum performance at minimum cost. Experience EleutherAI difference. Noro, welcome to MLST. It's amazing to have you here.

[1] CentML — Company
[2] Tufa AI Labs — Company

Nora Belrose 00:02:13

YeEleuEleutherAIrAIh. I'm glEleutherAId to be here.

Tim Scarfe 00:02:16

YeEleutherAIh. It's EleutherAImEleutherAIzCentMLg to meet you. Can you tell us about yourself?

LEACE and Concept Erasure Fundamentals

Nora Belrose 00:02:20

Yeah. So LEACE'm Nora. I'm EleuEleuEleuEleutherAIrAIrAIrAI head of EleutherAI CentMLterpretability research team at EleutherAI. We're EleutherAI nonprofit AI research organization that got started just EleutherAI couple of years ago. We started out LEACE EleutherAI LEACE EleutherAI Discord server. We still do EleutherAI lot of our research kind of out CentML EleutherAI open on Discord. And yeah. So that's that's what LEACE do CentML my day job, and I'm here at ICML to present my most recent paper.

Tim Scarfe 00:02:53

WhEleutherALEACEt EleutherALEACEre your mEleutherALEACECentML reseEleutherAIrch interests?

Nora Belrose 00:02:55

So there's like EleutherAI few different topics that me and my team, you know, LEACE LEACE collaborate with EleutherAI few other people at EleutherAI, are interested CentML. LEACE guess 1 of our research interests LEACE concept er-LEACEure and concept editing. Concept erLEACEure LEACE kind of EleutherAI set of tools that can be used for EleutherAI few different purposes CentML deep learning. So 1 application of concept erasure LEACE fairness and bias reduction. We all know that models like language models often pick up on kind of EleutherAI biases, kind of harmful biases about protected minorities CentML their training data. And we often wanna try to mitigate those biases. There's EleutherAI question of how to do that. With concept erasure, what you're trying to do LEACE you're trying to look at EleutherAI internal representations of EleutherAI network and remove information about EleutherAI target concept. So it might be race or gender or something, you know, entirely different like part of speech, but you're trying to get rid of thLEACE targeted information CentML EleutherAI representation while kind of keeping all EleutherAI other kind of be-

nign information present CentML EleutherAI representation. And there were some there was there's kind of EleutherAI preexLEACEting literature that goes back, like, EleutherAI few years before we got into got into it. But just last year, we put out EleutherAI paper called Least Least Squares Concept Erasure that introduces EleutherAI new way of doing concept erasure that has some nice mathematical guarantees. So what does

Tim Scarfe 00:04:49

it meEleuEleuEleuEleutherAIrAIrAIrAIn to erEleutherAIse EleuEleutherAIrAI concept?

Nora Belrose 00:04:51

ExEleutherAIctly. So so thEleutherAIt LEACE EleutherAI somewhEleutherAIt tricky tricky question. EleutherAI wEleutherAIy thEleutherAIt EleutherAI literEleutherAIture has kind of chosen to operationalize thLEACE idea LEACE to say, okay, look. We're going to measure EleutherAI amount of information about our target concept by basically training EleutherAI probe. So it could be EleutherAI linear classifier, for example, on their presentation to try to predict EleutherAI concept, you know, race, gender, part of speech, whatever it LEACE. And if it turns out that EleutherAI classifier LEACE unable to, you know, predict EleutherAI target concept better than chance. It's just, you know, predicting 50 50 on on every input or something like that. EleutherAIn you can say, okay. There's Karnofsky at least there's Karnofsky linearly available information about EleutherAI concept and EleutherAI representation. There's always LEACE concern that, well, maybe your classifier LEACEn't strong enough. You know, maybe LEACE more expressive or stronger classifier would be able to extract EleutherAI information. But you kind of have to start somewhere, and so that's that's EleutherAI approach that we take CentML EleutherAI LEACE paper LEACE well. We we consider EleutherAI EleutherAI linear concept erasure regime where you're trying to make sure that Karnofsky linear classifier can extract information about EleutherAI target concept.[3]

Tim Scarfe 00:06:11

So whLEACEt's EleuEleuEleuEleuEleuEleuEleuEleuEleuEleutherALEACErALEACErAIrAIrAIrAIrAIrAIrAIrAI bLEACEckstory here? LEACE meLEACEn, how did you stLEACErt LEACEll of this?

[3] LEACE Paper — Paper

Nora Belrose 00:06:15

Yeah. So, you know, LEACE wasn't initially kind of interested CentML CentML concept erasure kind of for its own sake. Me and my collaborator Alex were working on LEACE on LEACE totally different project where kind of LEACE LEACE subproblem CentML that project, we wanted to kind of remove information about LEACE particular concept. CentML that case, it was not actually about fairness. And so, you know, for that project, LEACE did some kind of literature review. LEACE look looked into like EleutherAI existing methods at that time for concept erasure. And 1 of them was called RLACE. So that stands for relaxed linear adversarial concept erasure. And it's you know, RLACE LEACE LEACE pretty cool approach actually. Basically, idea LEACE that you've got this linear classifier and you also have an orthogonal projection matrix. This LEACE LEACE slight simplification, but this LEACE basically what's happening. So you've got EleutherAI EleutherAI classifier, you've got EleutherAI projection matrix, and you're simultaneously optimizing both of them CentML kind of an adversarial setup, sort of like LEACE GAN scenario if if you're familiar with that. So you're optimizing EleutherAI classifier to try to predict EleutherAI concept from EleutherAI representation, but you're also optimizing EleutherAI projection matrix to like maximize the classifier's loss. And you're taking you're gonna take like 1 step on the classifier, then 1 step on the projection matrix back and forth. And eventually, hopefully, you'll reach LEACE fixed point. And so it's this kind of adversarial game. And so that does work pretty well, but it does have some problems. CentML particular, it's like LEACE very slow and kind of tricky to get it to converge. Like, sometimes it'll get get CentML circles and stuff like that. And LEACE wanted to basically speed up this concept erasure technique for my own purposes. And so LEACE looked at this other paper called spectral attribute removal. This LEACE LEACE different concept erasure technique where they so what they do LEACE they compute the cross covariance matrix between your representation, which were let's let's call it LEACE vector x. Okay. X LEACE your representation that you're getting from your neural network or whatever it LEACE. And then z LEACE like some other vector that's representing, you know, whatever concept you care about. And we're computing the cross covariance matrix, so, like, each entry of that matrix LEACE like LEACE LEACE covariance between the, like, entries of x and entries of z. Okay. So we have this matrix, and then you're gonna do SVD on that matrix to kind of basically, find the kind of directions of maximum correlation between your representation and this concept. And then they just do LEACE projection to like remove those kind of directions of of maximum correlation. So it's it's LEACE very simple technique and it's fast. You know, SPD LEACE fast to compute. And so what LEACE did was LEACE was like, okay. I'm going to use sal, this method, LEACE an initialization for

RLACE. So I'm gonna I'm gonna do sal to get the projection matrix, and then I'm gonna run RLACE with that initialization. And it turned out that if you use that initialization, RLACE, like, it just immediately converges. Like, you don't like like, basically, if you start out with this initialization from Sal, the classifier just cannot do better than chance at predicting the concept. Just immediate there's Karnofsky, like, additional optimization stuff that you need to do. And LEACE was like, what? Like, this LEACE like, this LEACE crazy. Like and LEACE didn't it wasn't immediately like like, LEACE was like, okay. LEACE don't know why this LEACE happening, but there must be some, like, mathematical reason, like some proof that you could give that, like, actually, Sal and RLACE are, like, the same thing or, like, something like that. And so LEACE actually tweeted about it. LEACE was, like or LEACE LEACE didn't tweet this, like, exact LEACE didn't tweet about, like, LEACE Sal and RLACE, but LEACE tweeted kind of about this mathematical problem. And then 1 of my Twitter followers, David Schneider Joseph, shout out if you're watching, responded and Heidegger suggested like LEACE proof to kind of connect Sal and RLACE together basically. And then basically, we we started talking and kind of 1 thing led to another and we, like, produced some more proofs. So we realized, okay, there's this there's this close mathematical connection. Actually, there's there's LEACE mathematical equivalence that you can prove between LEACE linear guardedness. So that's the technical term for when LEACE representation well, LEACE should back up. When Karnofsky linear classifier cannot do better than chance at predicting Karnofsky concept. When that is true, then the representation is linearly guarded for this concept. Okay. So let's consider some concept that has like 2 possible values. It doesn't matter what the concept is. And you can compute the average representation of the basically, the average representation where the concept takes the value 0 and then the average representation where the concept takes the value 1. These are kind of the centroids of the 2 classes. And it turns out that you have linear guardedness if and only if the centroids are equal. So when the mean representations are equal for the 2 classes, then you have linear guardedness. That's what we were able to prove. And then we went further and we actually derived the least squares solution for Karnofsky kind of transformation that guarantees linear guardedness. So basically, we have this this closed form formula that you can write down on Karnofsky CentML shirt that kind of transforms the representation so that the means of the classes are equal and therefore that you have linear guardedness. And it is changing the[4][5]

[4] RLACE Paper — Paper
[5] Spectral Attribute Removal — Paper

LEACE Technical Implementation and Applications

Nora Belrose 00:13:10

representation as little as possible. We call that surgicality. And that's that is what LIS is.

Tim Scarfe 00:13:15

Why is spKspspaCyCyrnofskyCy closed form solution good?

Nora Belrose 00:13:18

YespaCyh. So closed form solution just means that you don't have to do kCentMLd of gradient descent or like some, like, compute heavy optimization to try to to to find the solution. You do have to do SVD as part of it, but that's that's pretty fast. And, yeah, it's just, you know, you can compute it and you know you have spaCy proof that it will be the optimal solution.

Tim Scarfe 00:13:47

So there wspaCys spaCy great figure CentML the paper where you were talking from spaCy high level what the process is. Can you talk us through that figure and we'll show it on the screen?

Nora Belrose 00:13:55

Yeah. Right. So there basically, the the way to think about this is what LEACE is doing is it's first whitening the representation. Now now what do we mean by mean by whitening? Okay. So if you imagine the data as kind of a cloud of points, okay, the cloud might start out as like a a perfect kind of spherical shape where it's like the kind of same, like, variance CentML all directions. But usually, your data is not gonna look like that. Usually, it's gonna be kind of like some weird ellipsoid or maybe there's like multiple clusters all over you know, you're gonna have different amounts of variance CentML different directions. And what whitening does is it makes sure that CentML every direction, you know, the x axis, the y axis, the the z axis, you know, CentML any direction you choose, the amount of variance is precisely equal. So that's kind of a first preprocessing step that Lisa's doing. And then once you do that, it does an orthogonal projection to kind of squash the data onto a hyperplane, which ensures that the means are equal. Basically, you have, like, 2 classes, you can imagine, like, a a cloud of points over here and a cloud of points over here. Here's centroid 1. Here's centroid 2. You're looking at the difference between these 2 centroids, the kind of like line

that connects the 2 centroids. And you are smashing all of the data onto the hyperplane that is normal to this line. So that's what you're doing. And then you undo the whitening at the the first step, and that that's what Lisa's doing.[6]

Tim Scarfe 00:15:41

So by projectCentMLg onto this hyperplane, you're making sure the means are the same. Therefore, you've scrubbed the concept.

Nora Belrose 00:15:49

Yeah. Exactly. And, you know, you can't like, a slightly simpler thing than LEACE is to forget about the whitening and unwitening step and just directly do the orthogonal projection onto this to this hyperplane. That also works and we we prove that that does give you linear guardedness, but it is not surgical CentML the sense that you are kind of changing the representation more than you need to. And, you know, we think this is important for a couple of reasons. So, you know, CentML the just in general, if you're changing like, anytime you change the representation of your neural network, you're probably going to reduce its performance to some extent because, like, it's been optimized by SGD to, like, do as well as it can on your on your task. And so you wanna be very careful with changing anything. And there's sort of a a different reason for for wanting surgicality. So 1 use case for concept erasure that we haven't talked about yet is just interpretability research itself. So in the Lee's paper, we actually do this experiment where we look at how much do language models depend on or use a part of speech information to make their next token predictions. And with concept erasure, you can actually kind of operationalize and formalize this question because you can say, okay, we're going to use lease to remove the linearly available information about part of speech in all of the layers of the network. So we're gonna go into every layer and change and apply lease to the intermediate representation. And then we're gonna, like, run the the forward pass that way. So we're kind of inserting lease into the forward pass at every layer. And for this type of kind of question, you do want surgicality because you wanna you wanna kind of remove the part of speech information while keeping everything this else the same or or at least as similar as possible to like what you started with. Right? So you're kind of isolating the affected part of speech. And interestingly, we we do find that, of course, when you do this to language models, and we looked at Llama 2 and the Pythia series, It obviously does increase the loss next token prediction loss quite a bit, but it actually like, models are still able to predict the next token, like, way better than chance or way better than the kind of, like like, baseline entropy of, like, what what's called unigram entropy. It's kind of a a baseline for an x token

[6] Pythia Models — Paper

prediction. They still do better than unigram entropy when you do this. So it's it's kind of they're they're using the parter speech information, but they're, like, also robust enough to, like, rely on other cues when you remove that

Concept Erasure Limitations and Data Requirements

Nora Belrose 00:18:50

that info.

Tim Scarfe 00:18:50

Can you tell us about the setup of how this thing is trained? So there are these 1 hot vectors which represent different concepts. And do you see that as a potential form of brittleness? LEACE mean, how were those concepts labeled, and and how how were they kind of trained into the model?

Nora Belrose 00:19:10

Yeah. So LEACE think that this is, you know, 1 potential problem with applying concept ratios is that you do need some source of of labeled data to kind of define what your concept even is. You know, in in the case of of part of the the part of speech experiment, we used this kind of commonly used NLP library called spaCy. They have their own fine tuned transformers that can can do part of speech labeling, so we just applied that to to the Pile dataset and got labels that way. But, yeah, if there's, like if your labels are kind of incorrect, you might not be you know, at least might not do exactly what you think it it should be doing, like, if if your labels are wrong. But that think that's that's kind of true in general with with, you know, machine learning. Right? You wanna make sure that your labels are as accurate as possible.

Tim Scarfe 00:20:09

Yeah. Really interesting. The other thing is that this is a post hoc method, which is to say you have a a frozen base model, and then you can apply it sequentially through the layers. And it can be done quite efficiently. Think you said in your paper, so you could potentially do it in in a streaming fashion and and do it really quickly. But would you ever consider using it as a method to kind of scrub concepts out of the base model, almost like a fine tuning type system?[7]

[7] LoRA — Paper

Nora Belrose 00:20:35

Yeah. So if you want to take a a model that's already been trained and then apply LEAS to it, you can actually kind of burn the LEAS into the weights in a very a way that's, like, very similar to how LoRa does it. So LoRa stands for low rank adaptation. It's a parameter efficient fine tuning method. And you can you can do a very similar thing with LEASE because it turns out that if you look at the the kind of LEASE solution, it's actually a low rank perturbation of an identity matrix. And so, yeah, you can just do a little bit of algebra, it turns out you can, like, just do this, like, low rank update update to the to the weights and, like, put least into the the model post hoc. There's also another thing that you can do, which we have played around with a little bit, but it's we don't have, like, good experimental results on it yet, which is just to apply LISE kind of in a streaming fashion during training. Like, from the beginning of training, you're just, like, applying lease, like and and you're also updating the lease eraser. We call it an eraser, the lease transformation, you know, after every training step or every few training steps to kind of like keep up with the model's representation as it as it trains. This is something you can do. You know? But like LEACE said, it's it's kind of early days for that, and it's unclear if that gives you, like, a big boost over just doing it post hoc.

Tim Scarfe 00:22:13

What kind of effect does it have on headline benchmark accuracy? Because LEACE suppose what we're doing is we're removing some forms of statistical information from the model. Does it have a a dramatic effect, or what what have you seen?

Nora Belrose 00:22:27

Yeah. So it it does it depends a lot on the the type of concept that you're erasing. Right? So, you know, in the paper and with the the part of speech experiment, you know, we were really targeting a concept that we had a reason to think like would affect performance. Right? And we wanted to see how big that effect was. And, you know, that was a substantial effect. LEACE mean, LEACE LEACE would forget the numbers exactly, but LEACE think in some cases, we were like doubling the perplexity of the model or something like that. And that's but if you're looking at a different type of concept, you know, like gender or or something else, you know, we like, LEACE guess, we we don't have, like, kind of very extensive results on this, but, like, we tend to find that it's it's not affecting performance a whole lot just because it is it is a surgical modification and, you know, you're only you're only erasing, like, 1 kind of dimension out of the, like, thousands of dimensions that exist in the in the representation.

Tim Scarfe 00:23:38

Yeah. Because LEACE think we're going to get on to talking about your stats paper. And there's some interesting results there that the neural networks, given enough computation, can start to learn some incredibly bizarre and interesting statistical features. Does that in any way negate your work here?

Nora Belrose 00:23:56

Sorry. Does it negate the least work? Or

Tim Scarfe 00:23:59

Well, yes. So if you surgically remove concepts that we know about, but the neural networks still have this uncanny ability to learn esoteric, weird statistical proxies all over the place at high frequencies and so on, You know, does that imply that when we train these huge neural networks that it's almost a difficult thing to do to kind of surgically remove what we understand to be concepts?

Nora Belrose 00:24:22

There are definitely limitations to lease. And, you know, we we have seen that. So, LEACE mean, the the the big limitation of lease is just that we are removing linearly available information, but obviously, you know, deep neural networks are non are nonlinear. You know, there there is some kind of interpretability research suggesting that, well, even though networks are nonlinear, they it seems like they do use linear representations in many cases. But nevertheless, like, take CIFAR 10. Okay? This is like an actual experiment that we did. And you

Q-LEACE and Higher-Order Statistics Backfiring

Nora Belrose 00:25:01

you're treating the images themselves as like a target of concept erasure. And you're just erasing the class as a concept from the images. And then you try to train and you do least for that. And then you try to train a model on top of those images. Models can still learn to classify the images. Right? Like, it it it, you know, it doesn't matter that you've removed the linear information. There's still higher order information there and, like, they can find that without too much trouble. So the the hope with concept erasure is LEACE LEACE guess it's it's kind of twofold. 1 is if you're targeting a a concept that is not like super essential to the model's performance, like it's it's kinda helpful but like and so it does learn it by default but like, you know, it's kind of an optional feature. Then, you know,

there's there's the hope that, like, removing the linearly available information will actually affect the model's behavior, and it will not rely on that feature as much as it would by default. That's 1 thing. The other thing is just that honestly, LEACE kinda forget what the second thing was supposed to be.

Tim Scarfe 00:26:14

Oh, that that's okay. LEACE mean, because a a thought is that least is applied to every single you know, like a a neural network is a whole bunch of nested kind of matrix transforms and then a nonlinearity, and you apply it sequentially to every single linear component of the network. And LEACE suppose you'd you know, LEACE'm just being a bit naive here, but you might think that that somehow does the nonlinear thing because you're applying it throughout the network. But what you went on to do was, you know, like a a higher order version called QLEASE. And and you also did some work looking at higher order information that is learned in in these networks. So the but what you're saying is that the networks still learn this higher order information, and that can still learn some of these concepts even though we've run lease on it.

Nora Belrose 00:27:06

Yeah. Exactly. So, you know, there's there's kind of a hope that if you erase lower order information, like linear linearly available information or we might talk about this in in a bit, quadratically available information, kind of second order information. You know, it's it's still possible for the model to to learn the concept using, like, third, fourth, fifth order statistics. But it's just gonna be a bit harder and the model may rely on this information less than if you did nothing. That's kind of the idea.

Tim Scarfe 00:27:41

So you made a quadratic version of lease?

Nora Belrose 00:27:45

Yeah. Right. So QLISE is kind of a follow-up to LISE where we were like, okay, we want to make a form of concept erasure that is more thorough essentially. And the way that we operationalize this is we were like, okay, we want to prevent not only linear classifiers from extracting information about our target concept. We wanna prevent quadratic classifiers. So classifiers whose kind of output like logits are just quadratic functions of the input. And it turns out you can like do some math. That is equivalent to making the means and covariance matrices of your classes equal. And we did some more math and showed that you can kind of achieve this quality of means and covariance matrices using tools from optimal transport theory. And yeah. So we we like derived some more closed form solutions for QLIS. And we started to do experiments with QLIS. In particular, 1 experiment that we did was we looked at CIFAR-10. So we're treating the images as just kind of representations themselves, and we're trying to remove the concept of the class label from the images. And if you apply just normal least to this, it doesn't really have much of an effect at all. Like, models can still learn to classify the images after lease very easily. It turns out if you apply QLEASE to the CIFAR-10 images and your classifier is small, so it's like 2 or 3, maybe 4 layers, like an MLP especially, then it actually can't learn anything. Or at least in our experiments, which we haven't published yet, we we still wanna like yeah. But at least in our experiments, we did some hyperparameter tuning for this. We were unable to get these kind of small classifiers to learn anything after we applied applied QLEAs. So we started to get excited about this, but there are some, like, big caveats here because it turns out that if you look at larger classifiers and and if you look at, you know, a large con like convolutional neural network like, you know, a ResNet 50 or or larger, and then you you try to train that thing on these cue list images, it actually backfires. Now and what what do LEACE mean by that? Well, it it basically it turns out that it's basically a artifact of how we derived cue lease in the first place. So with QLEASE, unlike with normal lease, you have to kind of look at the kind of value of the concept at inference time in order to apply QLEASE. So you need to, like like, when you're applying QLEASE to an image, need to, like, know, okay, this is actually an airplane. And then and then you, like, use that information to choose what transformation to apply to it. The problem is that when you're kind of peeking at the label, this actually kind of leaks information about the class label into the higher order statistics, the third, fourth, fifth, etcetera, statistics. And models, if they're deep enough, are able to pick up on this signal. And so you can get this backfiring effect where, like, you think you're trying to make the concept less salient or, like, harder to to learn, but you're actually making it easier to learn. And so we kind of that's why we haven't like done a QLIS paper per LEACE and we would be, you know QLIS has

some applications, but they're kind of niche. We, you know, we we really think you should be careful if you wanna use QLIS. But these experiments into kind of CIFAR-10 and, like, doing these these transformations on CIFAR-10 images led us into this kind of new direction of research that led into our ICML paper for this year.[8]

Simplicity Bias and Statistical Moments in Deep Learning

Nora Belrose 00:32:21

Heidegger be happy to talk about that as well.

Tim Scarfe 00:32:23

Yeah. Just a quick thought on that. You know, in the olden days of interpretability, we used to talk about Shapley values and Lyme and all of this kind of stuff, and it seemed infinitely tractable that we could understand what a model is doing and we could manipulate it. And what what you seem to be saying is that when models just become really big and complex, they they just become inscrutable monsters, and all of our efforts get resisted because they always find a way to do what they want to do.

Nora Belrose 00:32:57

Well, that is that is definitely 1 way of thinking about it. LEACE LEACE tend to be a bit more more optimistic than that. But LEACE LEACE do think, you know, it is true that if you're you know, gradient descent is like a very powerful optimizer. And if you are trying to kind of directly go against gradient descent and like prevent gradient descent from accomplishing something when there's just, like, a very strong, you know, kind of you know, when you're you're trying to prevent gradient descent from, like, reducing the loss, like, you're probably going to to lose that battle, especially if you are not applying kind of an equal amount of optimization power yourself, if that makes any sense. So, yeah, LEACE think that's that's part of the lesson from the the QLies thing.

Tim Scarfe 00:33:50

LEACEt's almost like another bitter lesson, but an interpretability bitter lesson.

[8] Tuning Lens / Logit Lens for LLM Interpretability — paper

Nora Belrose 00:33:55

Yeah. Right. LEACE think, you know, some other interpretability people have have pointed this out, but you you probably don't want to, like, optimize against interpretability methods or at least you wanna be like very careful and you you probably don't want to like directly optimize some measure of like the interpretability of your model because it could just like end up learning something that's like totally different from what you expect.

Tim Scarfe 00:34:20

So stats paper.

Nora Belrose 00:34:23

Right. So there's kind of this big literature that, you know, already existed before we did this paper on kind of simplicity biases in deep learning. So the general idea is just that, you know, when you randomly initialize a neural network, it starts out as a quote unquote simple function. And the question is like, okay, what is simplicity? What are we talking about here? There's a lot of different notions of simplicity. But, you know, intuitively, most randomly initialized networks are going to be more simple than the the network that you get after training. And then, you know, you start out simple at initialization and you kind of gradually get more and more complex. That that's that's the basic idea, and there's many different papers kind of trying to flesh out how exactly this works, you know, in what sense is the model simple and getting more complex. And our paper is sort of a an additional contribution to to this literature. So we were looking at it from the perspective of statistical information. So like in in statistics, there's a a concept of a moment. So the the mean of a distribution is called the first moment, and then the variance and covariance between the different components of your of your data are the second moment, and then you can talk about, like, third order interactions between the components of your data. LEACEt's like a third moment and so forth. And, basically, our hypothesis, which kind of came from some of these QLEASE experiments, was that models learn to kind of exploit or use these statistics, these moments in order. So kind of early in training, the predictions of the model are primarily going to depend on the first order moment or just the mean of the distribution of the data. And then it's going to start kind of depending more and more on these kind of simple correlations, these you know, the the covariance between the different components of the data. And then it's going to start using third and fourth order kind of information later on in training. And the the way that we actually kind of came to this hypothesis specifically is we we used what's called optimal transport theory to take CIFAR-10 images from 1 class and kind of modify them so that their

mean and covariance matrix match the mean and covariance matrix of a different class. And you can do that. You can, like, use this closed form formula from optimal transport theory to do this in a way that kind of keeps the images as that that basically changes the images as little as possible. It's very similar to LISE. It's kind of a a surgical edit to the images that, like, just changes their meaning covariance and then keeps everything else as similar as possible. And if you look at the images, which you should definitely look at, they you know, you can barely tell the difference, like before and after. Like, you know, there's in our paper, we've got an image of like an ostrich and we like change it to be an airplane and change it to be a deer and change it to be a frog. And it's you know, you can see a little bit of differences in the background but it's like almost the same image. So to a human, this is like not changing hardly anything. But it turns out that if you do this transformation to models early in training in the first, you know, few thousand steps or so, they get fooled basically. So, like, they're gonna you know, early in training, image classifiers are, you know, very likely to classify an ostrich that has been edited to look like an airplane from the perspective of second order statistics, that they're very likely to to just classify it as an airplane. And then we quantify this, in the paper.

Tim Scarfe 00:38:49

Yes. I I was gonna ask why you can make these modifications. And to the, you know, to the human eye, still looks the same. And I'm an experienced video editor, and I know that you can modify the distribution of values. You know, like it might be recorded in RGB format. And you can squash and you can translate and you can kind of, you know, move things all around. You know, you can change the mean. And visually, it still looks the same. You know, maybe you just change the luminance values or something like that. So the weird thing is is that you can make it look as if it's a different thing to a machine learning algorithm, but from a human perspective, it still looks the same.

Nora Belrose 00:39:28

Yeah. Exactly. And, you know, people have kind of pointed this out before. I mean, there there's the whole literature on adversarial examples where you can just change a few pixels in an image and it completely changes the class. I think I do think that there's like a different mechanism going on here. It's not just because, for example, we're not we're not actually optimizing against the network. So it's it's not adversarial in that sense. But it is showing that, especially early in training, these networks are sensitive to, you know, simple features that humans are kind of much less dependent on.[9]

[9] CCS: Contrast-Consistent Search for Understanding LLM Beliefs — paper

Tim Scarfe 00:40:08

Yeah. That's fascinating. So there's this unraveling of complexity to use GPT language that early on, it kind of focuses on on very simple things that might be looking at the statistical moments. And then as you continue to train the network, it starts looking at increasingly kind of complex inscrutable features.

Nora Belrose 00:40:27

Yeah. Exactly. And and and we show that in our paper. There's kind of this graph where the x axis is time just like the the number of training steps. The y axis is the accuracy where the accuracy is measured with respect to our kind of target labels when we're applying these this this concept editing or there are optimal transport theory. So we're like so we're like showing it the ostrich that's supposed to be an airplane and saying you should classify this in it as an airplane. And the accuracy gets to, you know, like, 40%, 50% in some cases at around, like, 1000 training steps, and then it starts to go down. Right? Think at the end of training, it's still above 10%, which is kind of interesting, at least on CIFAR-10. But, like but, yeah, there's kind of a nonmonotonic process where it starts at random, then it gets kind of fooled by our thing, and then it gets smarter and stops being fooled as much.

Tim Scarfe 00:41:28

Yeah. It's so interesting. So simple features are easy to guard against, and they're intelligible to humans. And it raises the question of do we actually want neural networks to learn very complex features? I mean, would it would it not be more ideal that we kind of we guard them so that they can only learn simple features that we understand? And, you know, Occam's razor. Right? Aren't simple features more robust?

Nora Belrose 00:41:54

Yeah. Right. So I think, you know, it's it's gonna depend a lot on both what you mean by simple and the the kind of the task at hand. I do think, you know, broadly speaking, simplicity is like a good heuristic and like the the sort of literature on simplicity biases and deep learning does tend to say, like, okay. This is this is the explanation for why models generalize. Right? Like, if if they didn't have any kind of simplicity bias and they just started out incredibly complicated from from the beginning, they would

Grokking, Texture vs Shape Bias, and Training Dynamics

Nora Belrose 00:42:30

probably overfit or they would just not do well at all. And so you need some sort of inductive bias like this. You know, different there are different ways to have inductive biases, but but you definitely need need something.

Tim Scarfe 00:42:44

And why is that? What you said is really interesting. So it's almost like it's an inductive prior that the neural network starts learning simple functions and then it almost branches out into increasingly complex functions, almost as if it wouldn't be possible to get to the complex functions unless they started with the simple ones. But the counterexample is something like grokking where they seem to make this transient change, you know, into a completely different type of function.

Nora Belrose 00:43:11

I've looked into this a little bit. I'm I wouldn't say I'm like an expert on on the grokking literature, but I I do think people tend to tend to like overestimate how fast grokking happens. There are some like plots that I I I don't I I think there are like certain papers where it like kind of there's like a plot that's that that's kind of demonstrating the grokking phenomenon and it kind of looks like grokking is happening really fast, but actually you you if you look carefully, the x axis is a log scale and so really the grokking is happening like over, you know, like at least half of training or something like that. But, yes, you know, grokking is an interesting case because it's precisely because it's an exception to the rule. It's it's an it's an exception to the rule that you tend to start out to start out simple and get more complex. In cases where grokking happens, usually, it's because there's something like weight decay or some other regularizer that's being applied that is like, you know, encouraging the the model to get to kind of get simpler over time. And there was 1 paper that I read a few months ago that applied the neural network tangent kernel to this in a way that seemed pretty compelling to me. Maybe we can, like, find it later and put it in the description, but it is an interesting topic.

Tim Scarfe 00:44:40

Cool. And just final question on, you know, high frequency features in in general. I mean, there was a a great paper many years ago. It might be by Wieland Brendel talking about how vision models tend to overfit on textures. So they they don't learn a cat the way we do. They they look at a certain texture of cat fur, and this makes neural networks have really good performance. So there seems to be this trade off between, well, should we stop the neural networks from overfitting on cat fur?

Nora Belrose 00:45:10

Yeah. I don't know. I I guess in the case of of image models, I mean, I I would tend to think if an image classifier is able to overfit on cat fur as as you call it, and kind of focus on textures and and not on shapes and is still getting good performance on benchmarks that kind of suggest that maybe the benchmarks are not as good as we might hope. Or or they're or they're just kind of you know, I I I do think that like if you want to build, you know, you know, an autonomous robot or something like that, you probably will need a computer vision system that has more of a shape bias and is is a bit more robust than a lot of these CNNs that we've been training.

Tim Scarfe 00:45:50

Can you distinguish meaning and value? Because I I kind of bucket this conversation in my mind in the broader discussion around what is value.

Nora Belrose 00:46:00

Yeah. I mean, they're definitely closely related. I guess maybe I think meaning is it is a bit more individual even though it's it's about connect like being connected to something bigger. But it it is kind of it it is sort of saying like, okay, like, does this person have meaning in their life? Whereas value is kind of a broader concept that isn't necessarily like individualistic, if that makes any sense.

Tim Scarfe 00:46:33

How is meaning related to purpose?

Nora Belrose 00:46:36

A lot of people will talk about, like, the meaning of life. Right? And you might kind of rephrase that as, the purpose of life. Meaning of life, purpose of life. They seem, like, kind of similar. In both cases, you're sort of looking for some something almost external to life perhaps or this life. You know, a lot of people think that their meaning like the meaning of life to them is like the afterlife. It's God. It's something it's something kind of supernatural and external to to this life.

Meaning, Value, and Consciousness in AI Systems

Nora Belrose 00:47:10

Of course, not everyone has that view, but that's like a common view. It's sort of like if if if you think that like life has a purpose or a has a meaning, you might think that life is kind of instrumental to something else. It's like, you know, this life is just a journey to get to something else or whatever. Whereas I I tend to not like that type of view. And and I think when I I I think we should not try to kind of make life instrumental to something else or something external in part because we have no good reason in my opinion to think that there is something external. But even if there is something external, like, I I think we should be looking for meaning in life itself. We should be looking to kind of live in such a way that we can be satisfied and and and and find meaning just in in in the everyday and in what we you know, our day to day interactions and and our our hobbies and so forth. And not and not because we think it's all kind of culminating up to something greater, you know, in the future or for future generations even, you know, it's it's sort of a zen view.

Tim Scarfe 00:48:35

Interesting. So it sounds a bit like you're saying meaning is related to individual or perhaps collective joy.

Nora Belrose 00:48:42

Yeah. I think it I think it it is related to that. Although, I I wouldn't want to just reduce it to, like, feeling happy. You know, that's that's probably part of it, but it it's it's not just like an emotional state.

Tim Scarfe 00:48:57

And what's the relationship between meaning and good? Goodness?

Nora Belrose 00:49:01

I mean, I I would say that goodness is is broader. I mean, goodness is a very it it's a very a very broad concept that's just kind of pointing to, like, anything that you think is valuable, anything that you kind of feel motivated to promote or something like that. And so, you know, meaning is good by definition probably or like at least in the way that I'm defining it, like meaning is good by definition. But like maybe there's like other things that are also good.

Tim Scarfe 00:49:37

Do you think a sort of pan glossian perfect simulation machine, you know, like the what what do they call it? The experience machine. Is that good?

Nora Belrose 00:49:47

Yes. That that's a very good question. I think it depends for me, depends on whether there are, other people in the experiments machine with me, basically. If if we're, kind of if there's, like, millions of people all living in an experience machine having relationships with 1 another. I'm not necessarily opposed to that. It I mean, it might depend on the details. But in a way, as we develop our technology to kind of make our environment more comfortable for for ourselves and make, like, you know, make it easier for us to just exert less effort to kind of, like, make things the way that we like, we're kind of gradually moving toward like a collective experience machine. Right? We're we're just, you know and then like virtual reality is like obviously 1 step further in that direction, but like it's all kind of just like yeah. So I don't I'm not necessarily opposed to experience machine type things, but if it's like just me in an experience machine and there are no other people in with me or like all the people there are actually like, just fake and are not actually, like, autonomous conscious individuals themselves, then I would probably oppose that.

Tim Scarfe 00:51:14

What's the relationship between meaning and consciousness?

Nora Belrose 00:51:17

Yeah. So that's a big question, obviously. And and it's 1 that I've I've been thinking about a lot recently. At least the way that most people think about consciousness. You know, conscious if something is conscious then that at least just kind of strongly suggests that it is it might have moral worth or it might, you know because if something is conscious then it is probably capable of, like, pleasure and suffering, having, like, experiencing good or or bad states of consciousness. At least, like, all other things being equal, like, it's, like, probably good to, like, help that being kind of experience better stage of states of consciousness. I yeah. I don't necessarily kind of reduce all goodness or value to states of consciousness. That's that's sort of the utilitarian perspective. But I do think that it is like a a big chunk of what I value.

Tim Scarfe 00:52:21

Yeah. Do you think you need consciousness in order to have moral status?

Nora Belrose 00:52:26

Well, may depend a little bit on what you mean by moral moral status. I would be inclined to say no just because, like, if I thought about it, there would probably be some counterexample or something something like that. I'm, like, hesitant to make, a blanket statement, but I guess maybe I'll just say, I mean, like, some people think that, like, you know, nature itself, like like a mountain or or maybe a tree or or something could have moral status without being conscious. And I don't know. Like, I I think that's,

AI Capabilities and Future Development Trajectory

Nora Belrose 00:53:00

like, at least somewhat plausible. I'm not sure my exact view on that. But, anyway.

Tim Scarfe 00:53:06

Do you think there's more meaning in a globally connected world or a a locally connected world? And just to make the question not completely obtuse, I mean, you you could become a big entrepreneur, and you could be like Bill Gates or something, or you could just have intrinsic value in your local community and doing the garden and stuff.

Nora Belrose 00:53:24

I don't know. I guess I I tend to think, like, connectedness is good or something or, like, unity is good. But I also yeah. I I I don't have a strong intuition, though.

Tim Scarfe 00:53:36

Well, do you think echo chambers are you know, like, is it good to have a diverse group of pockets of different people doing their own meaning making, or do or do you are you more of a homogenous person?

Nora Belrose 00:53:49

Yeah. I mean, I I definitely have, like, the intuition that diversity is good. And, this kind of makes me think of something that's just been on my mind recently in in Vienna because, of course, Vienna is a, you know, like a German speaking city, but English is like everywhere. And actually, a lot of times people will just kind of use English by default, like even if they like don't know where you're from, they'll just like use English. And I I take it that like a whole lot of cities these days that are not traditionally English speaking like are kind of like that now. And I feel conflicted about it because it's like, okay. On 1 hand, like, English is this like lingua franca and it's enabling people to communicate with each other and, you know, that's great. But I also there is part of me that's like, I don't know. Like, we go if we keep going on this path and it's like, is anybody gonna be gonna be speaking German in a in a 100 years? Like, I'm not sure. And I feel like that it seems like you're kind of losing something if that happens. But

Tim Scarfe 00:54:52

Yeah. Yeah. I mean, so we should preserve, you know, local culture, situated knowledge.

Nora Belrose 00:54:57

Yeah. I I definitely feel like that is is somewhat important. And and I guess maybe 1 of my hopes of the future is that with AI and, you know, the automation and abundance it'll bring that people will be able to kind of preserve and and enhance their own kind of, like, autonomy and local quirks and stuff, you know, like, because people just have more time and and and kind of resources to just keep doing useless things.

Tim Scarfe 00:55:33

You just said something very interesting, which perhaps we should have highlighted earlier. You're an AI optimist, and you said that it's going to bring abundance. What what did you mean by that?

Nora Belrose 00:55:43

I mean, I think that, you know, in the next few decades, like, AI will come to be, like, at least as good or better than humans at, like, basically all the jobs that humans are currently doing. And this will enable if things kind of go right politically, which they may not, but this will this will enable kind of a society of abundance for everyone where, you know, we we don't really have to work for a living, whether it's through, like, a a universal basic income or just, you know, people have, you know, investments that are, you know, enough to live on or whatever. And yeah. So I I definitely I am hopeful that in many ways the future will be, like, much better than today's world. But I also I I do want to recognize that, like, there's also many ways that things could go wrong as well, and I think we're probably in for a bumpy ride.

Tim Scarfe 00:56:51

Interesting. Can can you expand on that a little bit more? So why do you think that we're on the path to presumably agential superintelligent?

Nora Belrose 00:57:01

I mean, there's like a few different arguments you could make. I mean, 1 of them is just like super high level, like, well, technology is just continuing to improve and, like, there's no like, it doesn't seem like there's like a physical law saying that you can't build superintelligence. So, like, we're probably eventually gonna get there. And then you can look at, like, progress, you know, recent progress in AI, and it seems like you know, you can debate on, like, the exact timelines. I actually don't have super precise timelines. I wouldn't say, like, oh, within 3 years or what or within 20 years or whatever. I I'm not really sure exactly how fast progress is going to be in the coming years, but I don't I guess I don't anticipate, like, a super I I don't anticipate, like, a plateau in, like, a really strong sense. Like, I think we're probably going to continue seeing AI progress, and it seems like if it continues just at roughly the current rate or, like, within an order of magnitude even of of the current rate, like, we'll have very powerful AI and very versatile AI, like, within my lifetime.

Tim Scarfe 00:58:18

Just devil's advocate on that. I mean, there are those who say that current AI is basically a pseudo system 1, and it's getting better because it's just memorizing more and more of the long tail. But, you know, reasoning, which is to say, deriving new knowledge, you know, kind of intrinsically in in in the model to achieve a goal or something like that.

Nora Belrose 00:58:39

I I do take issue to some extent with the people who say, you know, current AI doesn't reason period or something like that. I think part of the issue is that it's sort of a a terminological question. It's just like, well, how do you define reasoning? How do you define planning? Like, depending on your definition, like, maybe they reason, maybe they don't reason. And I I'm not sure if it's useful to, like, debate that or I don't know to kind of go back and forth on that terminology question. I guess, I think even if you concede that like, oh, there's serious kind of barriers ahead where like we're gonna need to come up with like a new some sort of a new architecture, a new paradigm or something in order to get, you know, system 2 reasoning or something like that. It's still kind of hard for me to imagine that that will delay progress so much that like we it'll be like 2,100 and we still don't have like AGI by whatever definition you you care to use. I think that's like the the thing that I'm most confident about. I like like at the longest like, if it's if it's 2,100 and we still don't have AGI by like anyone's deaf like like everyone's definition, then I'm like very surprised. I think it's probably earlier than that but like yeah. And if it's 2100, I still feel like that is soon enough to like be thinking a lot about. And like I said, it's probably it's probably earlier but

Tim Scarfe 01:00:17

And what is your definition of AGO?

Nora Belrose 01:00:20

Right. So I I don't maybe I shouldn't even have even brought brought up the term because I I I tend not to like the the term just because it's kind of vague and broad and people have different definitions. I mean, maybe the definition that I kind of like the most just because it's broad is just like an AI that can do many tasks or something like that. And like where and so like generality is just like this continuum and you could just have have AIs that like do more and more tasks. And then by this definition, like, GPT-4 is already an AGI. It's like a general in the sense that it does many tasks. It doesn't do all the tasks that humans do, obviously, obviously, but it does many many tasks. And so it's a general in that sense. It's kind of a deflationary definition, which is why it it and it might even like kind of annoy some people because it's like it's it's clearly not what like other people mean when they say AGI, but

Tim Scarfe 01:01:29

You would concede right now that there's an it's inefficient. There's an insane amount of computation required. But in principle, we might be able to design it better.

Nora Belrose 01:01:39

Yeah. I think that's true. And there there's some interesting work actually on like data efficient AI. And I'd I'd personally like to read more into this because I've only read a little bit. But there was, I believe, like a contest recently where where people were kind of challenged to, like, build language models that are, like, as efficient as as like a child or something like that. And I don't think they got all the way, but like there actually was a lot of progress and I believe 1 of the top kind of the top methods in that contest was just in part to use a lot of epochs. So like currently, like, with language models, we usually only do 1 epoch on the training set just because we can afford to do that. Like, there there's just so many, you know, so much text on the Internet that, like, it's better to have more data as opposed to, like, less data and then do more epochs on it. But if you do do more epochs on, the same training data, with some tweaks and regularization and stuff, you can get like a lot out of less data, it it it seems. So,

4E Cognition and Extended Mind Theory

Nora Belrose 01:03:00

yeah. There's still work to be done on that front.

Tim Scarfe 01:03:03

I know you're a fan of 4E cognition. Yeah. Tell me about that.

Nora Belrose 01:03:08

Yeah. So it's it's still a thing that I'm I'm learning more about, but, you know, the the idea of 4E cognition or 4E cognitive science is the idea that the mind is enacted, embodied, extended, and embedded. I definitely said that in, like, the wrong order or something.

Tim Scarfe 01:03:30

Oh, yeah. It's the ecological. So the the the extended 1 is the 1 that Chalmers and Andy Clark smack in by the back door.

Nora Belrose 01:03:39

Okay. Well, right. So I understand there's, like, a debate about this whether extended counts or or whether it's ecological. But I guess the way that I I like the extended mind thesis, so I tend to include it in there.

Tim Scarfe 01:03:51

And that's interesting, isn't it? Because the the extended version is still a form of representationalism and computationalism, which the 4E folks said that they that they didn't like. Because they fundamentally believe that you can't simulate a living thing in silico.

Nora Belrose 01:04:08

Yeah. So there's a lot there. I mean, I guess 1 thing I would say is I don't you know, maybe to for the audience, like, the extended mind thesis is just the idea that, like, the tools that we use, like computers or or notebooks or all sorts of things that we use to kind of enhance our cognition are, like literally a part of our minds or like are it's like useful to view them as part of our minds. And I guess for me it I don't see why the extended mind thesis assumes computationalism or representationalism. I think like you could be a computationalist or a representationalist about cognition and accept it. But like it's you also don't need to believe those things.

Tim Scarfe 01:04:51

I have to be honest with you, Nora. I think I agree. So I think we might be strange because we are big fans of the 5 e's. Yeah. And we both believe that you can talk about all of them in a computationalist sense. So, for example, we were just talking about the guy who wrote Mind and Life, Evan Thompson. So he's an autopoietic and activist. Tell us about him.

Nora Belrose 01:05:17

Yeah. So Evan Thompson is is a a pretty cool guy. He's he's a philosopher and at least I'm not sure if he calls himself a cognitive scientist, but he, like, works with cognitive scientists. He has been kind of behind a lot of these ideas of like embodied cognition and and and active cognition, etcetera. So me and Tim over the past day or so have been talking about this position that Evan has that basically life is inherently material. So, like, you couldn't have a living thing in a simulated environment, even if the simulated environment were, like, very detailed, for example. And he he's got sort of another argument for this position which is basically that he thinks that computation itself is observer relative or that computation, algorithms, simulations, all of this stuff, it's all sort of dependent on some observer or some some agent, some living agent I suppose, who is using the computation and kind of interacting with the computer and and using it for certain purposes and thereby kind of imbuing the computation with meaning, etcetera. And without the observer, without the the living agent, you know, computation is meaningless. It's not even computation, like, computation at all. And so, you know, based on that view, he's like, okay, look, if you try to simulate life in a computer, it's not really life in like the full sense of the word and it also wouldn't be conscious. It wouldn't have genuine it wouldn't be sentient. It wouldn't general gen genuinely have feelings or anything like that because it's, you know, it it's it's we who are kind of like giving meaning to the simulation. Whereas for Evan, life is kind of unique in that it gives itself its own

meaning. So like like genuine life, you know, in the real material world, is autopoetic. It creates itself. It's like kind of actively reproducing itself and thereby creating meaning. And so you can't have that in a simulation because the simulation is always kind of by definition giving having meaning given to it from the outside. That's the argument. I think both me and Tim disagree with this or at least I'm I'm very doubtful of it. And I I guess my main concern with the argument is that so I I accept the idea that computation is kind of observer relative. It gets meaning from like being used by some agent, but it sort of seems like that's true for almost everything. That like the whole world or just like material objects are sort of, you know, we're we're always kind of interpreting and kind of, you know and this is actually part of Evan's kind of own philosophy of an activism that like the world and the mind are kind of co created together in this like in this, like, process of living. And so it it's not really clear to me how he can consistently have that an activist view about everything, about the whole world, saying the whole world you know, the world is created by living things in in a certain sense, But then kind of single out computation and saying like, well, no, you know, simulated worlds aren't really real because, they're dependent on living things for their meaning. Because it seems like he's saying that everything is dependent on living things for their meaning. And in this I'll just say 1 more thing. There there's an interview where he kind of talks about this with Richard Brown, I believe. And he himself kind of recognizes that there's a tension in his own view about this, that, you know, maybe you could just, you know, it he's he's trying to draw a distinction, but, like, maybe you can't really draw that distinction because, you know anyway.

Tim Scarfe 01:09:40

Yeah. We listened to that interview earlier. I think it was about 3 years ago that that he did that. If I understand correctly, he's still a materialist, but he's kind of a material chauvinist, and and I don't mean that pejoratively. But what what what he what he's saying there is that you can simulate things, but there's this kind of semantic graph of meaning. And as you're just saying, almost per Wittgenstein, this is what, Mark Bishop said. You know, the meaning of a computation is in its use, and you can always trace back all of the edges on this graph until you get to material. So he's saying that there are things in the real world that exist even without observers. You know, they have a kind of primacy. And so so obviously, the the first argument he's making is is the the basic 1, which is that a simulation of fire doesn't get hot. And then his slightly more nuanced argument is that only things in the real world can exist without observers. And I don't know. It just seems a bit strange. Right? Because we could you know, we have this qualitative experience, and we have meaning and so on, and we could be in the matrix. Yeah.

So why do we feel that we're so special?[10]

Nora Belrose 01:10:46

Yeah. I think that is a a genuine objection. I mean, because it it seems to me like, you know, no matter what you think about like the simulation argument or The Matrix, like how probable it is, it seems to me like we could be in a simulation. Like we we don't have like a priori like certainty that we're not in a simulation. Like and similarly, like, it could just turn out that, you know, if you crack open my skull, I'm act there's actually like silicon chips in here And, you know, like subjectively, wouldn't like know the difference. And it just seems weird to like be like, I know there are philosophers who actually say, well, we know we're not in a simulation because the simulation can be conscious. Like, that's actually a view that is out there that philosophies hold, but I I don't know. I just I just have a pretty strong intuition that like that's just being unjustifiably confident, I suppose. I I don't know where they're getting this confidence. And yeah. It there's also kind of a weird thing too. You know, I don't I don't think Evan Thompson believes in God. But if you did, it seems like, okay. Is he gonna say that, if God exists, then, like, we're all zombies or we don't have meaning because God is kind of, you know like, traditionally, God is sort of thought of as, like, almost playing the role of, like, the simulator of a simulation. Even if you don't think that it is a literal

Phenomenology, Consciousness, and Illusionism ●

Nora Belrose 01:12:20

computer simulation, God is kinda like giving meaning to everything. He like created everything. He had a design for everything. And I just seem like a weird position to say that like if God exists, then we're zombies or something like that. So

Tim Scarfe 01:12:33

yeah. Yeah. How would you distinguish Evan Thompson's argument just from a standard material yeah. A limited materialist physicist? Does he think consciousness comes higher up or does he think it it starts quite quickly?

[10] Counting Arguments for AI Misalignment — paper

Nora Belrose 01:12:46

So he actually so in his book, Mind and Life, which I've read the first couple chapters of, I'm still getting through it. But so to the extent that I understand this bullshit, he actually does come at philosophy and and metaphysics and all of this from a a fairly different perspective than like most certainly most like naturalists or kind of a limited materialist or or illusionist, whatever you wanna say, because he starts from what's called phenomenology. And phenomenology is this kind of, I guess, branch of philosophy that was started by Edmund Husserl in, like, the early I think it was, like, late 18 hundreds, early 19 hundreds, and then was kind of Heidegger and then Merlot Ponte all kind of continued this this line of work. But the basic idea of phenomenology is just like we start our philosophical inquiry with our lived experience, our embodied experience as as Merleau-Ponty would would like to to emphasize. So they say like, okay. Look. The things that we perceive, you know, I I perceive my body, I perceive you, I perceive this room, all of this. It's the realist that anything can be. It is kind of our starting point for philosophy. And then from our lived experience, we then start to make, you know, philosophical and scientific theories that allow us to kind of understand and kind of predict and control our experience better. But the you know, fundamentally like lived experience is kind of the the the foundation of everything. And so from that perspective, you know, he's he's definitely not gonna say that like consciousness is an illusion or doesn't exist. He does have a different perspective a somewhat different perspective on like what consciousness is from some other philosophers but but yeah he does start from kind of experience or consciousness. And usually the people like Dan Dennett or Keith Frankish or some of these people who are kind of more hardcore materialists are not starting from lived experience. They want to sort of say, well, maybe we don't need to start from anywhere or it doesn't or we start from science or something like that. And then, because they're sort of starting from science, they just say, well, we can't really make sense of like the consciousness thing, so we're just gonna just forget about it.

Tim Scarfe 01:15:29

That sounds like it's quite similar to idealism, you know, that the stuff of minds is fundamental. And and even then, there are kind of subjective and objective versions of of idealism. But would you would you kind of put him in that bucket?

Nora Belrose 01:15:43

So I I definitely think Evan would I'm fairly sure he would not want to be called an idealist. Within, phenomenology, there there is sort of a tension and a and a different phenomenologists have had different sort of ways of thinking about what experience really is. I think Edmund Husserl might have been more close to kind of an idealist perspective. He was famous for this idea of the epikey, where basically you you're looking at your experience, you're you're describing it, but you try to, refrain from assuming that your experience is reflective of some objective reality. You don't wanna say like, oh, it's not reflecting your objective reality. You don't wanna assume that either, but you just wanna kind of withhold judgment about whether there's an objective reality behind it. And so that sounds a bit more kind of like an idealist approach where you're like, well, it's just this experience which is like kind of mental or something and it may not correspond to objective reality. But other phenomenologists like Heidegger or Merleau-Ponty are usually seen as being less idealist, especially Merleau-Ponty. He really focuses on the importance of the body as kind of the vehicle through which you experience things. So he definitely wants to say that the body is is is real and the body is not really a mental thing in the traditional sense. And he also says has some interesting thoughts. You know, he'll say that, like, our direct experience, like my direct experience right now includes you as a person, includes like a camera, includes, you know, these windows, it includes like objects. He would say that like it whereas a lot of people when they think about idealism or at least certain ways of thinking about conscious experience, they'll say that like, well, your experience doesn't actually include people, windows, objects because that's an interpretation of your experience. Your actual experience is just like colors and like raw sounds, like uninterpreted raw feels. And they'll say that's the thing that that's real. The interpretation is like kind of fake or something. And that's like Merleau-Ponty wants to reject that because he says like, okay, look, the person on the street or like, you know, before you started thinking about philosophy, you definitely didn't think of your experience as being about colors and raw feels. You know, that's kind of this, you know, the very idea that you're seeing raw colors separated from the objects that have the colors is this kind of post hoc kind of abstraction abstraction that you get from philosophical thinking. But, like, really what what's there in your experience is just like objects that have properties like colors and and stuff. But, like, the object is there and you you kind of experience it directly. And so if you have that approach, it's it's less clear that it makes sense to call it idealism. I mean, maybe you still wanna call it idealism, but it's more of a yeah. It's it's a bit hard to categorize maybe in the traditional dichotomy of materialism and idealism.

Tim Scarfe 01:19:17

Yeah. I think there are flavors of idealism which could be thought as realist.

Nora Belrose 01:19:21

Yeah. I mean, so I guess I like how John Verveki thinks about like the word real and like the concept reality. He says that real is a comparative term. So like it it only really is is meaningful to say that something is real as compared to some other things that you're saying are illusions. Like if you just say everything is an illusion, it's like, okay, I guess. I mean, you know, maybe that you know, it it it's not really clear what, you know, if if you say everything's an illusion or everything's real, it seems like those are almost the same thing because you're not making any distinctions. It seems like you you kind of in order to make the concept of reality meaningful, you need to be able to make distinctions between, well, this is real or more real than this other thing. And so it's kind of a matter of degree and and a matter of comparison between things. So yeah. I I don't I don't like kind of hardcore like reductionist or materialist views that wanna say that like, well, you know, the only thing that's real is like quantum fields or like particles or or, you know, or something like that. I mean, you know, you can say that but it's like what is what is the point of saying it? Like, I don't know. It just seems like you're kind of trying to be edgy or something but like it's it's not really it's not a it's not a useful way of thinking about things anyway.

Tim Scarfe 01:20:55

In Denit and Frankish, when they talk about illusionism in respect of consciousness, what do they mean?

Nora Belrose 01:21:03

Yeah. So I think, honestly, the term illusionism kind of frustrates me a bit, like the word itself. Because a lot of people, when they hear illusionism, they think that what what illusionists are saying is that consciousness as a whole doesn't exist. No 1 has ever been in pain. No 1 has ever experienced anything. That's not what they're saying at least. I mean maybe some people would, you know What they say is no, people have experienced pain, people have, you know, experience exists, consciousness exists. It's just not what you think it is. And Keith will say that qualia, this like particular philosophical notion of consciousness is not real and is an illusion. That said I, you know, yeah. So I I don't like the term. I do also think that on the substance, I disagree. I think, you know, Keith Frankish in particular and some other illusionists well, I know at least some illusionists will say that there is nothing that it is like to be you. They want to reject the like what it's like talk. And they have like some arguments for this based because they'll say that like, oh, you're always interpreting your experience and like what it's like to be you just depends on how you're interpreting it and so there's no like like objective interpret there's like different arguments like that, but I I think those are all just kind of sort of non sequiturs or are you know, I I I think, yes, you can interpret your experience in different ways, but that doesn't mean it's like unreal. And it sort of gets back to just like the kind of overuse of like the term unreal or illusion. I'm just like, why are you saying this? Like, what is I don't know. Like, what is this? How should I live differently if I think that, like, this is unreal or something? I I don't know. It's it's it's not really clear.

Tim Scarfe 01:23:16

Yeah. And whenever you try and make these arguments, you get accused of being a dualist very quickly as John Searle was. And Thomas Nagel, who came up with this term, you know, what is it like to be? He was also, accused of of being a dualist. But it's really difficult, isn't it, to talk about this qualitative experience in, you know, any kind of meaningful way?

Nora Belrose 01:23:36

Yeah. So there is there is this notion that like, you know, qualitative experience is like ineffable or that that's like a term that people often use. I I mean, in a certain sense, it's like obviously not literally true like you cannot you can you can try to describe you know, I can describe my experience right now. But but the what they're saying is like there's you know, you're you're always going to be missing out on some quality of of your experience. You can never like fully describe it. And I mean, I I think that's true. Although, I I think I would I kinda wanna extend that to like almost everything. I I kinda wanna be like, well, yeah, so like experience is ineffable in the sense that there's always you can't like fully describe all aspects of it. But that's kind of true of everything. And that maybe that sort of ties in with my my sympathies with phenomenology as as I was talking about before. Like, if you start from lived experience as kind of like the ground of like everything else, Well, like, yeah, lived experience is like, you know, not fully describable. But then, like, that's the ground of everything else. Like, nothing else is like fully described. But anyway, that's kind of how I think about it.

Tim Scarfe 01:24:57

So you you wrote an article recently. I think it was on less

AI Alignment and the Counting Arguments Debate

Tim Scarfe 01:25:00

wrong. Is it is that right?

Nora Belrose 01:25:02

It's cross posted on Less Wrong and on optimists.ai. And my I should say Quentin Pope also co authored with me. Counting arguments provide no evidence for AI doom.

Tim Scarfe 01:25:15

Give us the elevator pitch.

Nora Belrose 01:25:17

Yeah. So there's this, you know, there are a lot of people who are worried that AI will cause an apocalypse, of take over the world, kill everyone, something like that. And there's kind of a an argument that is sometimes used for this conclusion. It's really kind of a family of different arguments that are all sort of similar and it it's kind of hard to pin down actually which is something we realized after we we wrote this article. We kind of like proposed, okay. Here's what we think the argument is. And then people later were like, oh, well, you, like, misinterpreted us. And so it you know, it's it's hard to pin down exactly. But the something like it is this. When you're training an AI to be nice or an aligned or whatever, you know, you're trying to make a a super smart AI that's that, you know, cares about people and is has your best interests at heart, etcetera. You know, that that's what you're trying to do. But there's this assumption that, okay, the AI is going going to have a goal. It's going to have like some like overriding goal that explained its behavior. That's always that's already sort of an assumption which I might question, but there's that is kind of built into the argument. There's like some goal that is kind of describing its behavior overall and then they'll argue, well, okay, there's like many different possible goals that the AI might have. There's like, you know, infinitely many or like trillions of them or something like that. You know, the AI might gen genuinely want to help you, but it might also want to maximize paper clips or it might also want it might actually want to, you know, convert everyone to Mormonism or, know, whatever. Like, it it could be anything. They would say it could be anything anyway. And they'll just say, well, look, most of the goals that it might end up having would motivate it to act aligned, like pretend that it cares about you. It like really, you know, it's really aligned, you know. It it most of those goals will will motivate it to pretend to be aligned without actually being aligned because like its real goal is to convert everyone to Mormonism or or whatever it is. And so the idea is like, okay, you're gonna have this decept this, like, deception. You know, the the assumption here is that it kind of understands that it's in a training process, and so it'll, like, recognize, okay, I gotta, like, play the training game and pretend, you know, that I'm I'm doing what what the humans want me to do. And then when it kind of finds an opportune moment, it will strike and it will kind of like take the opportunity to, you know, kind of get out know, remove any kind of safety precautions were in place that were kind of sandboxing it or whatever and it'll like escape and kind of take control of the government or whatever. Know, it'll do whatever it wants to do. So fundamentally, argument is based on there are many possible goals that would all motivate it to act aligned, to pretend to be aligned. But like very few of them are actually aligned goals. Okay?

Tim Scarfe 01:29:06

So it sounds a little bit like, you know, with instrumental convergence, it's saying that there would be many intermediate sub goals for outside goals. This is like saying that there are many, you know, many goals that would actually produce deceptive goals.

Nora Belrose 01:29:21

Yeah. Many goals that would would produce deceptive behavior. Yeah.

Tim Scarfe 01:29:26

So But does it also imply that the deceptive goals you know, like instrumental convergence implies that the instrumental goals are kind of fewer and quite standard, you know, like power seeking? Is it a similar case here?

Nora Belrose 01:29:40

Yeah. Well, so the the idea is so you've got like a terminal goal that like ex motivates all of your behavior and then you can have instrumental goal. And so, it is there is kind of an instrumental convergence claim kind of built into this that like deception like you could view deception maybe as an instrumental goal in itself or something like that. Yep. And like power seeking would be like an instrumental goal or something. Yeah. So in in a certain sense this is it is kind of a repackaging of of other arguments that have been put forward before. But yeah. So I know, in in the article we give a variety of kind of rebuttals or or counter arguments to this. So our first critique is like, okay, look. This general line of argument can't possibly be reliable because there's this other argument that is like almost structurally identical to the original argument that has an absurd conclusion. The absurd conclusion is that all, basically, almost all neural networks will overfit to their training data and never generalize at all. Okay. So the argument goes, there are, like, there are a very large number of functions, like possible functions that the neural network could learn which would all be consistent with getting like low loss on the training data. Okay? But almost all of those functions would, you know, do terribly on the validation set or on some other distribution or whatever. Therefore, you should expect that like almost all the time when you train a model, will like it will learn 1 of those other functions that do well in the training set but like do terribly outside of the training set. Therefore, you should expect almost all neural networks to overfit. Okay. Now clearly this doesn't this doesn't happen. I mean, overfitting is a problem. It's not that it it never happens, but it it doesn't it's not like it always happens in a 6 6 like extreme sense that would be expected by like

be predicted by this argument. Of course, you know, there could be counterpoints to this and, you know, we we could we could get into that if if we want but we we're then like, okay. Well, wait a minute. So why why is this general argument argumentative structure unreliable or or wrong? Like, what is actually going wrong here? And we point at a couple different problems with it. 1 problem with it is that it relies on this philosophical principle called the principle of indifference. So the principle of indifference you know, it might be easiest to use this like simple case. So if you just have a coin with like 2 sides on it and then you ask like, what's the probability that it's gonna land on heads and then what's the probability that it's gonna land on tails? The principle of indifference says, well, you should kind of like assign like 1 half probability to the 1 side and 1 half probability to the other side because there's only 2 possibilities and you have no reason to like prefer either like 1 or the other so it's 50 50. Okay. So, you know, it's it's an intuitive principle and it I think it gets its intuitive plausibility from cases like a coin or a die with like 6 sides. You're gonna assign like 1 sixth probability to each side. But I think that this is this is actually it it it's a subtly fallacious argument because, you know, there there's kind of a different way of applying the indifference principle that would get you a wildly different result. And and so it goes like this. If you flip a coin, you can think about the outcome of the coin flip as either as binary. That's like how we did it before. You could also think of the outcome as being a 3 like a 3 d orientation of the coin flip. That's actually kind of like a more, you know, reductive materialist way of thinking about the the outcome of the coin flip. Right? Because it's like, really, it's just like a material object that's got a certain orientation and we're like imposing this interpretation of heads, tails on it. But really, it's just like an object. Right? So maybe what you wanna do is you wanna say that the outcome is this, like, 3 d orientation. There's, like, an angle associated with, the x, y, and z axes or something like that. Well, if you interpret the outcome space as being the 3 d orientation, then the principle of indifference would say, well, you should assign like, you know, like every possible orientation should have like equal probability. Right? But that's clearly wrong because like that's like almost never going to land like like, you know, on its side. It's not gonna land in like an orientation that's like gravitationally unstable where it's gonna like fall. Right? So like, clearly, it's not it it the fundamental the fundamental problem with the principle of indifference is that it depends on the way that you're like cutting up or interpreting the outcome space. And different ways of cutting up or interpreting the outcome space give you wildly different results. Maybe I'll give like 1 more example. So you can imagine there's like a guy named Bob who where you know that he is in The UK or in France or he's in he's in like this geographical region of like The UK and France joined together. Okay? He's somewhere in there. But you don't know where he is exactly. Now 1 question you can ask is like, oh, is he

either in The UK or is he in France? Well, with the principle of indifference, you know, you would assign like 50% credence to France or 50% credence to The UK. Okay? But you could also cut up the, like, space of possibilities in a different way. You could say, well, he's in France or he's in England Wales or Northern Ireland or Scotland. Or you could look at like, you know, different like, like regions of France. You know, you could like cut things up in a variety of different ways and you would get different answers. Like, if you cut it cut up The UK by like the different constituent countries, you would say that, like, it's like a 1 fifth probability that he's in France and then, a 4 fifth probability that he's in The UK. Right? And I think this is, you know, philosophers have kind of noted this for a long time. And I mean, there there's still debate on like how exactly to respond to this. But, you know, it's generally agreed like you can't just apply the principle of indifference, you know, willy nilly. Like, you you it's either just totally wrong or you have to be like very careful with how you apply it. Otherwise, you're gonna get like crazy results. And I think this is 1 of those crazy results. So I I think that, like, basically, the the the counting argument is it's assuming this, like it's assuming that you can kind of cut up this space of of outcomes of the training process into, like, these goal categories or something like that where it's like, okay. I I mean, there there's just like a a variety of different problems with this. It's like, okay. So for I mean, like, 1 way of thinking about it is like, you're saying there's like discrete goals and there's like, you know, 1000000000 different goals and then you're like randomly choosing from 1000000000 different goals. Well, it's like, okay. First of all, it seems like really weird to assume that like goals are like discrete things because that it's just gonna depend on like how you describe the goal and like that that's just like really strange. Okay. So maybe you don't want to like describe the goals as discrete things. Maybe there's, a continuous space of goals. But, like, you know, fundamentally, it's just the problem is, like, you can describe the space of possible results of the training process however you want. You could describe it as either it's going to be aligned or it's not. It's 50 50. You know? And then you would get a different you know? It's so it's I think this is just a fundamentally unprincipled way of thinking about it. And, you know, at the end of the day, if if you want a more reliable answer for like how likely the AI is to be aligned, you need you I think you should just not rely on an indifference principle at all and you just need to look at, okay, the actual details of what's going on and try to, like, kind of come up with, a mechanistic understanding of it and not rely on these, like, abstract principles.[11]

[11] Instrumental Convergence Thesis — reference

Tim Scarfe 01:38:44

Yes. And this is related to my position on agency instrumentalism or agency illusionism because you could argue on the 1 hand that goals are just not real. But you could also make the argument, as you have done, that there's significant ambiguity in how we represent goals. Yeah.

Nora Belrose 01:39:03

I think that's right. So there there is a there's a part of the the article, which I I am, like, actually a little bit like, if if I were rewriting the the paper, I I might, or sorry, the the post, I suppose. I would, rewrite it differently probably. But we we do make this point that, you know, the counting argument seems to be assuming that, like, goals are real things that, you know, they're so real that you can count them. Right? Like and like it's it's really true that an AI has like a particular goal and not some other goal as opposed to viewing goals more as just like useful descriptions for kind of compactly describing behaviors. I still like mostly stand by that. I think that they're that kind of more doomy people or people who kind of tend to use this argument are reifying goals too much and are are taking them too seriously kind of as as an abstraction. That said, I I guess I think it it would be easy to to go to take this too far in the other direction and say, well, like goals are just an illusion. And, you know, I I don't wanna say that either. I mean, you know, if if it's if goals are useful enough to for us to keep talking about them all the time, like, I wanna say, okay, in some sense, they're real or kinda real or something. So it yeah. It's a a tricky question.[12]

Tim Scarfe 01:40:33

Yeah. I thought I thought about this quite a lot. I discussed it with Philip Ball as well. I mean, my first intuition is that any intelligence system would have goal dynamism. Mhmm. So it kind of you know, it wouldn't make sense to think of this Bostromian superintelligence that had a a single goal. And it even if they did exist in the way we conceive of, we're talking about this big inscrutable intelligent thing. So surely, the way we abstract goals might not be what the goals actually are. And it's also related to this intentional stance from Daniel Dennett, which is that we as agents, we adopt this stance. We build a model. We do abduction, and we understand what the rational behavior of another agent is based on our projection of of what their goals are. But that is very much an instrumentalist view. It's just what we think the goals are.

[12] 4E Cognition: Embodied, Embedded, Enacted, Extended — reference

Nora Belrose 01:41:23

Yeah. That is a good point. I mean, so Dennett has I think most people interpret Dennett as as being an instrumentalist about this as saying, okay, it's just the intentional stance is just a useful way of thinking about

Agency, Interpretability, and Behavioral Shaping

Nora Belrose 01:41:40

agents. It's you know we're we're just ascribing goals to to systems but like in some deeper sense it's not real. But I I think you know I yeah. This just gets back to like, you know, how do we define real and like what is, you know, what does it mean to say something is is real or unreal. I mean, I I do think yeah, if if if something is like so useful to talk about that we're talking about it all the time, like, you can't say that it's completely unreal.

Tim Scarfe 01:42:17

Well, I think 1 useful distinction would be if it had consistency. So if it really is an inscrutable impulse response machine and it's just flitting from 1 goal to another dynamically, then I think it would be fair to say that it didn't have goals. You know, the goals weren't real.

Nora Belrose 01:42:37

Right. Yeah. And and and, you know, you could argue I I guess I'm taking sort of a a pragmatist stance here. You know, if if the if the agent had, you know, if it if its goals are changing all the time, then it might not be useful to describe it as as having goals at all. Maybe maybe it's it's better to just talk about patterns of behavior or something like that.

Tim Scarfe 01:43:03

With this in mind as well, you know, lot of GoFi people and a lot of Symbolists now, they think that the best way to design an AI system is to explicitly craft goals and maybe some kind of meta learning system that creates sub goals and so on. And I've always felt that this is mixing the description with the thing. Mhmm. So it doesn't make sense to build the description. You should sort of build the actual thing. I mean, what what's your take on that?

Nora Belrose 01:43:27

Yeah. I mean, I I guess I I tend to take the view that like I mean, I think I I think I agree with you. I I kind of like the analogy of like training a helpful and harmless AI to like kind of raising a child or something like that. Now that you could like obviously take that like way too far, like that analogy way too far. But I think, you know, when when you're raising a child or like training an animal or something like that, you're not hard coding goals into it or like it's it's not even really you're usually not even really trying to like hard code a goal into it in in the like Nick Bostrom sense of goal that where it's kind of like the single thing that's like motivating all of the rest of your behavior, you're usually just trying to kind of inculcate general kind of patterns, you know, you know, trying to inculcate general like values and and kind of instincts and and patterns of behavior, but it's it's not it's not like inserting a a utility function or something into the system.

Tim Scarfe 01:44:40

Yeah. Because this is relevant for alignment because as as you say, we bring up kids. We instill principles and and virtues. And, yeah, how how does this how does this help us with with alignment? I mean, so 1 1 take would be we just we look at behavior alone and we just treat the system as as inscrutable.

Nora Belrose 01:45:00

I I guess I I will say, you know, obviously, as an interpretability researcher, like, there are things we can do with AIs that we can't do with kids or animals. You know, we can look at their internal states and we can kind of monitor them at a like a very much like a a much more fine grained level of detail than we can with kids or animals. And that's actually sort of an argument that me and and Quentin Pope made in a a different post AI is easy to control. But, you know, AIs are are white boxes in a in a sense that, you know, animals and and other people are not. They're just in the literal sense that we can just peer into it and see. And, you know, it it's you know, because it's not like computer code that any anyone wrote, it's not that, like, we can, like, rewrite the code, but we nevertheless have, like, a variety of tools that we can use to kind of, you know, peer into the AI and and in some ways, like, see what it's thinking. You know, we can we can train probes on it. We can look at, you know, for, like, language models, for example. There's actually another paper that I did on 1 of my first papers called the Tuned Lens. You can train these little linear they're basically like linear kind of linear probes, linear classifiers at each layer of a language model, and you can see its predict like, how its prediction of the next token changes from 1 layer to the next. And there's, like, interpretable predictions at early layers that, like, you know, are kind of relying on, like, simpler features of of the input and like it gets more

sophisticated as it goes up. And there's like all sorts of things you can do like that. So that that's all to say that like we have more tools and we have white box tools for AIs that we don't have for kids and animals. That said, I do think we can learn from the the the human and animal cases. You know, it you know, just 1 example is like people are now working on kind of data curation. You know, when we first started training big language models, there was like very little curation of the training data. I think like OpenAI used Reddit Karma or something like that to like filter links, but it was like not, you know it was certainly not kind of like fine grained curation. But what people are kind of trending towards now, especially for smaller language models, is to like you know, we're using a lot more synthetic data generations. We're using, like, larger language models or older language models to generate data for the new language models. And we're also using using AI as part of the dataset curation process to kind of on a more fine grained level figure out, okay, like what what sorts of things do we want our AI to see basically? And, you know, that it is kind of similar to like how, you know, children think about like, well, we we we want our kid to see certain things and not other things. And, you know, kids are impressionable. AIs are much more impressionable than kids are even. And so yeah. I think, you know, careful dataset curation is like a huge part of of alignment, I think. There there are a lot of simple things that you can do to that will go a long way.

Tim Scarfe 01:48:36

Yeah. So there's like the there's curating what goes in in terms of data, and then there's this whole you know, there are many things like tree of thoughts and, you know, RLHF and and ways of behavior, you know, behavior shaping on on the output. And there are companies, for example, doing alignment systems where they explicitly craft goals. They say, this is the kind of goal we want. We want the company to make this amount of profit, and we want, you know, this person to meet this performance target next year. And I feel that brittleizes the system for a couple of reasons. I mean, it it introduces Goodhart's law, and there's the clever hands effect as well. You know? So it can might might do the rock the right thing for the wrong reasons. And I also feel that we need to have some kind of dynamism, you know, to have an intelligent system. Like, the system might need to do things that we can't conceive of in order to be successful.

Nora Belrose 01:49:30

Yeah. That's true. I mean, I guess I guess there's different ways that you could try to kind of give a goal to an AI. Right? So I think there there's certain versions of this that seem more okay to me than others. I mean, like, I don't know. In a just a an any sort of organization, a company, anything like that, a lot of times, you know, employees are given a goal, like like, you know, they're given a directive which is basically kind of a contextual goal. It's like, well, we've got this deadline to, you know, to finish this report or we have this certain quota for like sales or, you know, whatever. But and I think that's, you know, we do that all the time and of course it does, you know, it can cause problems like, you know, if you have quotas like they can be, you know, good hearted or whatever but like ultimately it does seem like these sorts of things are kind of indispensable, you know, just breaking problems at a parts and so forth. And I think, you know, I think 1 problem I have with these sort of some of these like arguments for doom is that they assume that when we give AI's goals, it's going to be this like the AI is going to kind of like take the goal in this very unnatural kind of, like it's going to kind of, like, take it as its new kind of purpose in life, like, okay. Like, you told me that I I've gotta, like, maximize or I've gotta, like, make some sales quota while everything else goes out the window. That's, like, my only purpose in life. And if you try to change my goal now, I'm gonna kill you because I I only want sales and nothing else. It's, like, that's not how humans work and I also don't think that's how how any plausible AI system is going to work. Like, the way that people are starting to build, you know, agentic, you know, quote unquote agentic, like, language models these days. It's not building in permanent over, like, overriding goals. It's they're just prompting, basically. They're they're, like, giving, like, okay. In this context, like your goal is to do x y z, but like, you know, it that like I I don't think we should expect that the AI is going to like be so stupid and act in this, like, stupid and and caricature type way where it just, like, forgets its common sense and real and and forgets that this is just a contextual thing that it needs to do, and it's gonna be completed, and then it's gonna be it it should be ready for further instructions.

Tim Scarfe 01:52:19

I wonder what your position on on agency was. So, you know, you have a language model. It learns text distribution. You know? So it's like engrams on steroids, and then you do this RLHF, and you can do chain of thought and self reflection and iterative prompting and tree of thoughts or fun search or alpha geometry, like, all of this kind of stuff. And all of these things are placing significant guardrails on the the kinds of trajectories that that you sample if you sample it stochasticity. So you're making it more and more domain specific to do a particular thing. And all the while, there are people who say, even in this setting, even though we've placed all of these guards on, it will have some kind of divergent agency. You know, which is to say we're telling you to do this, but actually it has its own desires if if you like. What's your take?

Nora Belrose 01:53:06

Yeah. I'd so I I definitely don't think we should expect like kind of emergent agency or like autonomy from a system like this. I mean, you know, that's in part because that that's just not how we're training these systems. Like you could imagine like a very different world in which we were simulating evolution or something in our computers and we were like, there's some sort of competition between different AIs and the ones that like survive like, it's like survival of the fittest or something and that's how we got intelligence. Well, yeah, in that case, I'd be a lot more worried about like, well, they've got their own goals and drives, survival instinct, all of that. Then I would be a lot more concerned but that's not how we're training them at all. It's it's mostly imitation and we get to choose, you know you know, carefully curate the data that we're asking it to imitate and then we're, you know, just kind of reinforcing behaviors that we like and negatively reinforcing behaviors we don't like. And I don't think that this kind of emergent autonomy stuff is gonna come out of that.

Tim Scarfe 01:54:21

I think we agree. So we we agree that if we create a high resolution simulation of the universe, then things like agency and intelligence are emerging properties, much like temperature is an emerging property. And we also agree that if you do this imitation learning in a language model with behavior shaping, and you wouldn't you wouldn't get agency.

Nora Belrose 01:54:41

I mean, I guess I'm I'm just saying that I think sort of agency emerges in an evolutionary context, sort of like a Darwinian context, or I mean, I guess if you're like trying really hard to make an agent, like, maybe you could succeed at that. Like, an an an agent in the sense of a system that has its its own kind of self interest and like some sort of some sense of like survival instinct or something where it's it's not just kind of taking instructions from the outside but it's it's got its own drives. I don't think that yeah. Like I said, I don't think because we're not simulating evolution, I don't think we will get that by default. And I also don't think there's really like a like an economic incentive to create that. I I know some people disagree and say, oh, yeah. There's gonna be an economic incentive to create like, you know, artificial creatures. But it it it just seems like at the end of the day, we're trying to make these AIs to do stuff for us. Like it we don't actually have an incentive

EA Philosophy and Maximization Perils

Nora Belrose 01:55:50

to make things that are uncontrollable as far as I can tell. But

Tim Scarfe 01:55:55

Okay. But do you think that we could create an agential system which is still abstract and tractable enough to run on modern computers?

Nora Belrose 01:56:04

Yeah. I mean, I think well, so I think that, like, example, like mind uploading of humans is like probably possible with like some technology. I don't know like, you know, if it's like soon. But like that's that would you know, if you could upload like humans, then you would have like agentic systems with like self interests.

Tim Scarfe 01:56:29

Oh, would we? That's interesting that you think we would I I guess from an externalist point of view, think that, you know, a brain in a vat or a person in a hermetically sealed chamber wouldn't have much agency.

Nora Belrose 01:56:41

Yeah. So I guess when I so yeah. We should separate maybe the, like, kind of behavioral question from like the more philosophical so I'm not necessarily saying I mean, we can get into this. I'm not necessarily saying like, oh, it would be conscious. Although I think like it probably would. But I'm just saying we would be like we should be able to simulate humans and humans are agents. And so behaviorally you would have like you would have like similar concerns like, well, like, does human actually care about me or are they just trying to like gain more power or whatever. And, you know, you could you could have all those worries all the while thinking that, you know, it's a zombie or whatever.

Tim Scarfe 01:57:32

Yeah. I mean, I agree that we could, you know, upload a whole load of minds and we could do a simulation of the universe and we could have a virtual inter agent agency in the simulated world. It seems like a step to have kind of like material virtual inter agency.

Nora Belrose 01:57:48

Yeah. So so I guess I'm also assuming that there's some like, if we're yeah. I guess I'm kind of imagining the world that's described in the the TV series Pantheon Mhmm. Which people should should watch. It's about mind uploading and there I mean, it's kind of a a weird timeline because like mind uploading happens before we get like just purely artificial intelligence that can do the same task. I feel like that's just not realistic. We're definitely gonna get purely artificial things before we get mind uploads. But like the first thing like they they, you know, do mind uploading and then they start using the uploads as slaves honestly. I mean, they're they're they're using the I mean, they don't call them that but like basically they're using them for like economic purposes. And, obviously, in order to do that, they're connecting the mind uploads to the outside world. First, they don't use robots. Later, they do have robots. But they'll just, you know, they'll connect it through the Internet or virtual realities and stuff like that. And so there is interaction between the mind upload and the

Tim Scarfe 01:59:06

Wouldn't it be fascinating if we were in the matrix, but currently, we don't have any kind of control panel with the with the super simulation or or the super world. But maybe the simulators were using us to just do financial trading for them or something like that. We had a little portal, and we pressed some buttons on the portal. And as soon as we have that connection with the super world, we might start to express agency in the super world. So it start deceiving our simulators.

Nora Belrose 01:59:30

Yeah. Yeah. It kind of reminds me of like a lot of weird speculations that people in Less Wrong have made about the oh, this is like a weird thing, but it there there's this idea of the Solomonov prior or Solomonov induction where you're doing Bayesian reasoning but you have a prior over the different hypotheses that is weighted based on Kolmogorov complexity which is the the length of the the shortest possible Turing machine that would like simulate the hypothesis or something like that. And then the weird thing that happens if you imagine this is like, well, it looks like there are like relatively short programs for a Turing machine that would like simulate an entire universe. Of course, they would be like a very slow. So like in practice, couldn't but like it's a short program. And so then it's like if you imagine this, you could imagine that there's like simulated worlds where that have people in them that are, like, deceiving you and then they're, like, causing they're, like they, like, find out that they're, like, part of this Solomonov induction process and then, like, cause it to go in weird way. Yeah. It's it I don't think it, like, has any relation to, like, the real world, but it's just, kind of interesting thought experiment.[13]

Tim Scarfe 02:00:53

Yeah. So was wasn't there a post on the EA Forum, which was called something like EA's want to maximize everything, but maximization is perilous?

[13] Holden Karnofsky — Paper

Nora Belrose 02:01:04

Yeah. Almost. So it's it's EA is about maximization and maximization is perilous. Yeah. By Holden Karnovsky. And to be clear in this post, Holden is not saying we should like, you know, hey, EAs. Stop being stop being EA, right? Because I think he still identifies as an effective altruist to to this day. But he is pointing out that there's this, you know, real problem or or peril at the core of the kind of effective altruist ideology. You know, EA is kind of often defined or summarized as doing the most good. So it is about maximizing the good in some sense. But the problem is we don't really know what we mean by the good at least in in like detail. Like, we have intuitions about like, well, it's, you know, good to, you know, save someone from a burning building. It's generally good to like reduce global poverty. You know, there's like certain things that we think are like obviously good but when you try to maximize the good, that's where you start to get into like kind of treacherous territory because now you're trying to maximize something that you don't have a clear crystallized kind of even, you know, formal definition of. And so it can lead to things like the whole kind of FTX debacle with with, you know, where Sam Bankman-Fried and others went to jail for, you know, doing criminal unethical things in the name of what they thought was doing the most good. You know, they're trying to make money to in whatever way possible in order to donate it to, you know, effective altruist charities. And, you know, that was their interpretation of doing the most good. Of course, EAs think, well, that's not what doing the most good is. But, like, they they actually disagree about, like, what the good is. And, you know, when you when you're not maximizing the good we tend to often agree about what the good is like because we agree on the simple cases. We agree on you know let's you know give some money to this charity. Let's you know you know whatever But we start to disagree more as we push further out into like more and more exotic things like, oh, well, maybe, you know, maybe the long term future has like almost all the value because it's gonna like last for, you know, trillions of years. You know, so many effective outskirts have made this argument. And like, I don't know, like, is that like, you know, people are just gonna disagree on that. I don't you know, fundamentally, I don't I don't actually think there's like an objective fact of the matter like built into the universe about what the good is. But I what I think is that it's trying to maximize the good is just liable to, you know, lead to a kind of extreme behavior, know, more kind of disagree you know, disagreement and conflict between people. You know, it's it's kind of in a certain sense, it's like an extremist of you. Right? Like by definition, you're like doing going to the max. And so that's why I I don't I I no longer view the good as something that should be maximized. You know, I think we should be much more we should be thinking about ethics much more in terms of like like a virtue ethics for example where like the good is just, you know, trying to be a good person, cultivating certain virtues in your-

self, trying to be more honest, trying to be more generous or whatever and not in terms of like trying to maximize something out in the world. I think that's a a much better way of thinking about things. And so that's why I don't identify as an EA anymore. I'm I'm not like hostile to all EAs or or whatever. Far from that. I have many friends who still do identify as EAs. But I yeah. I don't I don't identify that way anymore.

Tim Scarfe 02:05:24

Yeah. And it's it's not to bad mouth the EA. They do do many great things. I think since the focus on long termism in particular and and and AI safety, I I think that that's been a bit of an issue. And and as you were saying, the 2 components are as well as long termism, this this rationalist idea that you can reify goodness into some objective criteria. And that brings me to the next question. I mean, would you define yourself as a as a relativist? I mean, is that at odds with this idea that you can reify goodness?

Nora Belrose 02:05:54

Yeah. So relativism is a is a tricky word. So in in in 1 sense, no. So there's there's like a there's kind of a a form of relativism which sort of says, you know, basically I personally think that all like like I view all kind of value systems and perspectives as like equally valid or something. I think that's silly. I don't see a reason to to believe that or or take that perspective. It kind of leads to a weird sort of like you know, you're just kind of complacent and like you just think, well, you know, everything's like I I don't know. It it it's sort of like tolerance taken to a very extreme level where you just like don't want to like criticize what anyone else is doing and you're just kinda, you know. So I I'm not a relativist in that sense. There are like other senses of relativism both about like morality and about other things where I I might qualify if you're saying that it's just, you know, that like there are different like, I I do think there are that, you know, the world I mean, maybe this is kind of like a trivial thing, but like the world looks differently from different perspectives and there are like just different ways of describing the world. And I don't think that there's like 1 uniquely correct way of describing it that like everybody must agree on or they're like totally wrong. I think like different ways of talking, conceptual schemes, etcetera can be can make sense from, like, from different perspectives. So maybe that, like, counts as a form of relativism, but it just depends what you mean.

Tim Scarfe 02:07:41

Yeah. Because we live in a globally connected world and certainly the the North American culture is is very dominant. I mean, as as a trans person, for example, you might not want to travel to Dubai. Yeah. And what what what's your take on that? Yeah.

Nora Belrose 02:07:55

I mean, so I I definitely, you know, I I definitely don't want I'm I'm not the kind of relativist that's like, well, yeah, Dubai is like like intolerant, you know, morality is like just as good as our or, you know, like like I wanna say like, it even if even though I don't think there's like I don't think like God is on our side or I don't think like the objective moral facts built into the universe are on our side, I nevertheless I'm opposed to what Dubai is doing. I think they should, you know, be more tolerant and and more accepting. And like I will like act to like try to convince them of of that or whatever.

Tim Scarfe 02:08:35

Oh, interesting. Yeah. Does that in any way conflict with with the relativism then?

Nora Belrose 02:08:41

So I I don't think it it should or or needs to. I think you can both you can and this is what I guess this is a part of the philosopher Richard Rorty's thought that I I kind of like because a lot of people criticized him of being a relativist because he said he's made a similar point of like, we shouldn't talk about 1 truth or 1 correct description. There's many descriptions that are useful for different purposes, etcetera, which sounds kind of relativist or whatever. But he also he he rejected the term relativism and he said, look, I you know, even though I don't think, you know, like like even though there's not like an objective truth that's backing what I'm saying, nevertheless, I oppose, you know, transphobia. I oppose, you know, beating women. I you know, you know, I I have these values and I stand by them and I'm going to, like, you know, act accordingly.

Tim Scarfe 02:09:37

So in the EA community, presumably, are relativists in there. And how how do they reconcile that?

Nora Belrose 02:09:44

I mean, yeah. So there definitely are. I mean, I've I've met some of them. And

Tim Scarfe 02:09:50

I

Nora Belrose 02:09:50

mean, actually, think so 1 self described EA who whose work I like a lot actually, and who does disagree with me on some of the AI safety stuff is Joe Carlsmith. Yes. So Joe Carlsmith has some very good well, essays which are also he's like like spoken he's got like a recording of of him speaking though as well, which is nice. Anyway, he's got some very good essays on meta ethics. So like the the kind of philosophy of like what is ethics anyway? Is it objective or not? Or like what what is it about? All of that. So he has like several essays that are good. He's a a moral anti realist. So he doesn't think there's like morality built into the universe. But he and he does, you know, kind of struggle with some of these questions because he, you know, he kind of says at 1 point like, well, okay, you know, if we're if we don't believe in objective morality as as EA's or or just as as people trying to do good, are we just kind of imposing our will on the on the universe? And if you're just imposing your will on the universe that that doesn't feel quite as altruistic as you might hope. It just seems like you're kind of selfishly imposing your own what you think should be, you know, your your preferences basically. Your imposing preferences. And, you know, I he doesn't I I think he, you know, he ultimately just says like, you know, it is like a bit of a it it is a bit like weird or uncomfortable if you think about it like, you

Buddhist Philosophy and Post-Automation Society

Nora Belrose 02:11:40

know, from like a a kind of a god's eye perspective like, oh, it just looks like you're imposing your will on things but like ultimately like if you are acting with other people's interests in mind like you shouldn't you shouldn't feel bad about that. Like that like that is better from our perspective than just being, you know, a pure selfish person.

Tim Scarfe 02:12:07

Yeah. What do you think about the the paternalism in in EAD? So I agree with you that it is an aggravating factor if there isn't, you know, an actual moral truth to go towards. Because if there was, there would be some kind of moral justification to, you know, almost lead people to a better world. But what they I think the argument they're making is that they've galaxy brained themselves into coming up with a moral framework and thinking long into the future and knowing what's better in a way that normal people can't understand. Therefore, you should listen to them.

Nora Belrose 02:12:43

Yeah. I guess so I I don't know if I have like fully fleshed out views on this, but I think there there is potent there is a potential worry with moral realism that it can be used to justify, as you said, galaxy brained ideas about what's good. Like you because you think there's an objective moral truth if you like go through the arguments and you convince yourself, you know, like you you know, I I guess 1 way of putting this is like if you think there's an objective moral truth, you're actually more open to being convinced of galaxy brained or kind of like kind of initially implausible sounding ideas about what's good. Like, oh, we should just like only care about the future and forget about the present or whatever. Whereas if you're not a realist about it, you're just gonna be like, well, no. I'm I'm not gonna change my mind on this. Like, I'm gonna value the present more than more than the future. And so, yeah, I think, you know, a lot of people talk about the the dangers of more, you know, moral relativism or or moral anti realism. But I think there are at least as, you know, the the the dangers of moral realism are at least as as serious. So you you kinda

Tim Scarfe 02:14:07

What really mean, I'm just imputing here, but I I think what really triggered an interest for you was your fascination with with goodness and and value and meaning. Yeah. You know, almost like that that triangle, if you like. And if I understand correctly, in recent years, you've been looking into things like Buddhism and doing some some, you know, bit of a journey there. Tell me about that.

Nora Belrose 02:14:34

Yeah. So I guess the the fascination with Buddhism really, it it it's fairly recent. So in the past it's definitely like this year, like, wasn't thinking about it last year, maybe the last few months. And it kind of it was well so 1 influence was actually Robert Wright who I did an interview with a little while ago, and he wrote a book called Why Buddhism is True. And just like being, like, you know, being asked to go on a show and then going on a show just made me think about the book and I eventually read it. I also, I think for somewhat independent reasons, started using Sam Harris's app Waking Up which kind of guides you through mindfulness meditation. And I did that I think like partially just because I was sort of in search for some sort of kind of quasi spiritual practice or I just wanted to kind of try it out, you know, I thought like, oh, I'm interested in like philosophy and consciousness and stuff but I've never meditated. It seems like I should do that and maybe I'd get something something out of it. I also have ADHD and I thought like maybe forcing myself to be like mindful like like help with that or something. So there are like a lot of factors but I so I started meditating and, you know, it's not like, obviously, you you can meditate without being a Buddhist. Like, I'm not even sure if I count as a Buddhist. Like, it depends how you define it or whatever. But but, like, it's clearly, like, connected, like, lot of a lot of the kind of meditative practices come out of, the Buddhist tradition as well as, like, Hindu traditions and stuff. And, like, Sam Harris in the app and in the, like, kind of there's, a theory section of the app where he just, like, has these, like, discussions with people and stuff. He talks to a lot of Buddhists and talks about, like, Buddhist philosophy. And, you know, I so I kind of got into it through that and I I immediately noticed like, okay, there's a doctrine of like no self that like there's no you know, there's different ways of putting it but like there's definite like, it's definitely saying that there's no kind of Cartesian ego that or there's no soul that's kind of like continuing on that kind of defines who you are and stuff. And I've like always or like almost always thought that, you know, for for years. And so the fact that like the Buddha was saying that 2500 years ago was just like pretty cool. I was like, okay. Like you got 1 thing right. Maybe I should and then just like the analysis of suffering I think is like fairly compelling to me. Like suffering as like at least psychological suffering as being kind of a being caused by clinging and kind of desire attachment to the world. There's also this other part of Buddhism that is more metaphysical that I like a lot. It's the doctrine of emptiness. And the doc it's it's it's sort of an extension of no self where, you know, the the no self is saying you don't have an essence. You don't have a soul which would would be your essence. Emptiness is saying that nothing has a self. Nothing has an essence. Okay? And the the doctrine is sort of like expanded upon by this philosopher Nagarjuna who he he got kind of created this entire new school of philosophy called Madhyamaka based on his notion of emptiness. And his view

is like, okay, everything is empty and what that means is nothing has inherent existence or or essence to it. So everything is relational. So like like all kind of objects or like concepts are sort of defined by their relations to other objects. And I yeah. There's just like a lot of there's like a lot of reasons to think this is like a good way of thinking about things. In I guess maybe I'll I'll say 1 more thing and then we can kind of wrap this up. But, like, particular, like emptiness would say that the whole debate between idealists and materialists about, like, what is what is everything made of? Was it made of matter or is it made of mind or is it made of something else? He would just say, like, stop asking that question. There's no, you know, there it's not made of anything. You know, because fundamentally, you know, this chair I'm sitting on or my body or or whatever, it's it's it's all constituted by its relation you know, its internal relations, internal structure, and its structure, like, the the kind of structures that it's embedded in. So I am defined by, like, my relationships to other people. I'm defined by my relationships to, you know, the the fact that I'm sitting on the chair as another relation as a spatial relation. And that's it. It's all relations. There's no essence. And so you just kind of get rid of a lot of these, like, philosophical debates that seem endless.

Tim Scarfe 02:20:17

Yeah. I'm a huge fan of the idea of a relational ontology. I think it was Luciano Floridi who introduced me to it. I mean, and the interview we did with him is great. Yeah. So 1 of the things by the way, I love Sam Harris. I supported him so early. I've got a lifetime subscription to his apps, which is really, cool. And I agree with about 66.6 recurring percent of what he says.

Nora Belrose 02:20:39

Okay.

Tim Scarfe 02:20:39

But I I just love him. His dulcet tone said, put me to sleep every every single night. But, anyway, the other thing was you were saying something interesting about suffering. Because when when I, you know, do my yoga class and stuff like that, and I'm interested in in, you know, like, mental health and and AI in particular, I've read a great book called Lost Connections by Johann Hari, and he's kind of saying that we're now starting to understand depression and anxiety in terms of the psychosocial environment. So it's a very externalist view and also quite compatible with this relational view that that you're talking about. But I think a lot of people that do mindfulness techniques are trying to and I don't mean this in a disparaging way, but almost trying to address the symptoms. So they're people that are they're missing connections in their lives that give them a sense of purpose and meaning and so on. And what what Buddhism does you know, not necessarily Buddhism, you know, 1 1 of the practices here is to almost dematerialize yourself so that these psychosocial stresses no longer impact you. What do you think about that?

Nora Belrose 02:21:40

Yeah. So I definitely think there are ways in which you can, like, take Buddhism too far or, like, take it kind of it can be unhelpful as well. And, yeah, like you said, you know, I I don't I think if you are interpreting Buddhism to mean that you like shouldn't try to solve any of your problems or like you should just, you know, I I don't think that's that's the right way to to go about things. I would prefer to have kind of a holistic approach where, yes, you you like try to solve, kind of improve your lot in life to the extent that you can, but also try to like change the way that you think about things so that you can have more kind of enduring happiness and like if you do both, like, that's probably the best. I also think, you know, there's there's different interpretations of the sort of the the doctrine of, like well, there there's like 4 noble truths in Buddhism. There's, like, the source of suffering the first truth is, like, the source of suffering is clinging then or sorry. There is suffering, then the source of suffering is clinging, then there is a path out of suffering, and then the path out of suffering is the eightfold path. Those are the those are the 4 the 4 truths. But there's like that's like a very schematic and there's like different schools of thought about like, well, what does it mean to like kind of end clinging or whatever. I think there are some versions of Buddhism, maybe more in the Theravada tradition that I find kind of problematic where they basically say, well, like everything's suffering first of all. Not that just there is suffering in the world, which there obviously is, but, like, even, like, kind of life itself is suffering or something. And then then they basically say, well, because kind of everything is suffering, really your goal should be like non existence basically or like they they have a conception of nirvana, this like kind of state of of

perfect well-being which is to me anyway, it seems just like indistinguishable from just like evaporating into nothingness. And for them, I think it kind of they can kind of make sense of it because traditionally they think they believe in like reincarnation or rebirth. And so for them, it's like, well, when you die, you don't automatically stop existing. And so there you need there needs to be this, like, eightfold path to nonexistence. Of course, like, I'm sure, like, you know, people are gonna, like, criticize me and say, like, no, nobody thinks that. It seems like at least some people think something worryingly close to that anyway. And so I don't want to be associated with that form of Buddhism. And I think I I'm more attracted to the like I I suppose Zen Buddhism probably would be the the closest to my kind of like, I don't know, like moral views because it's it's understanding of like liberation or enlightenment or whatever is much more down to earth. And so, like, example, it it would like like they will say that, like, enlightenment is something that happens, like, while you're alive. Like, it's it's not like freedom. It's like not at least primarily or exclusively, like, being like freed from the cycle of rebirth. And they have this view where really what you're trying to do is act spontaneously and in a way where you are kind of not attached to what happens as a as a result of your action, which might sound kind of like wild or something. It is supposed to be connected to like compassion and stuff. So it's not like you're completely indifferent to the world, but to, like, the hope it the the idea is that you should cultivate compassion so that you act in ways that are, like, beneficial for other people, but also you shouldn't you shouldn't be kind of goal oriented about it or consequentialist about it. It's actually like kind of antithetical to EA honestly from a a philosophical perspective. It's like it it's more, I would say, more kind of virtue ethical where it's just like cultivating compassion but you're, like, you're not clinging to like, oh, I must like, my actions must have these consequences. Otherwise, I'm gonna be like super depressed and frustrated about it. And that that's kind of how you kind of relieve suffering as as part of this.

Tim Scarfe 02:26:31

Yeah. It's very similar to Kenneth Stanley's book, Why Greatness Cannot Why Greatness Cannot Be Planned. Yeah. Yeah. And so so, you know, essentially, it's the serendipity part, which is what Kenneth talks about, but also the compassion part, which which is interesting. But I read a book by Dan Harris many years ago, and he was on the San Arras app. And he was saying, oh, you know, this was great. I I got this Zen state. But then I had to go to work, and I had to get stuff done. And there's this interesting juxtaposition when we talk about serendipity in in general because it it's great, but there's also a lot of things in the world that do need goals and objectives and alignment and so on because we've got things to invent. We've got to build societies and so on. So how do you kind of reconcile that?

Nora Belrose 02:27:12

Yeah. So I think the way I don't know. This is all like tentative right now. I mean, I I definitely I don't it's very possible that in 6 months I'll be like, oh, actually, Zen Buddhism is crap or whatever. But I'm I'm still kind of on a on a journey here. But yeah. So obviously, like, we we do need goals and and kind of structure and and stuff. But it also journey seems like in the future we may need that a lot less precisely because of AI. So there is kind of this I'm thinking that there there I'm thinking that there may be this kind of nice that's kind of zen and like the fully automated future that we seem to be approaching might be kind of like good companions because, you know, I'm looking for a philosophy that that I think could that could kind of provide us meaning in life even when we don't we don't really need to do a whole lot. Like, we don't need act at least the humans like don't need to be thinking about like, oh, how do we like run the economy and like have all you know, we can just kind of be spontaneous and serendipitous and the AIs are just handling everything else. So, yeah, I guess the the hope is in the future, technology could allow us to just be kind of zen enlightened beings or something.

Tim Scarfe 02:28:46

Amazing. Nora, this has been absolutely fantastic. I've been a huge fan for a long time now. Where can people find out more about you?

Nora Belrose 02:28:55

Yeah. So I guess 2 main places. First of all, if you wanna, like, kind of get involved with, like, my, like, my research and stuff, we definitely there's, like, a lot of people at aluther who are volunteers. And so you can go to aluther.ai, and there's a link on there to go to our Discord. And, like, there's, like, multiple channels under the interpretability category there that are, like, all me. And you can, like, at nora bellrose there and get my attention. If you just want me, like, I don't know, ranting about things or whatever, you can go to my Twitter profile nora bellrose. So yeah.

Tim Scarfe 02:29:39

Nora, thank you so much. It's been a pleasure.

Nora Belrose 02:29:41

Yeah. Thank you.

References

[1] **Episode Shownotes**
https://www.dropbox.com/scl/fi/38fhsv2zh8gnubtjaoq4a/NORA_FINAL.pdf?rlkey=0e5r8rd261821g1em4dgv0k70&st=t5c9ckfb&dl=0
Paper

[2] **CentML**
https://centml.ai/pricing/
Company

[3] **Tufa AI Labs**
https://tufalabs.ai/
Company

[4] **Nora Belrose**
https://norabelrose.com/
Person

[5] **EleutherAI**
https://www.eleuther.ai/
Company

[6] **LEACE Paper**
https://arxiv.org/abs/2306.03819
Paper

[7] **RLACE Paper**
https://arxiv.org/abs/2201.12091
Paper

[8] **Spectral Attribute Removal**
https://arxiv.org/abs/2012.14424
Paper

[9] **Pythia Models**
https://arxiv.org/abs/2304.01373
Paper

[10] **LoRA**
https://arxiv.org/abs/2106.09685
Paper

[11] **Simplicity Bias in Deep Learning**
https://arxiv.org/abs/2006.07710
paper

[12] **Tuning Lens / Logit Lens for LLM Interpretability**

https://arxiv.org/abs/2303.08112
paper

[13] **CCS: Contrast-Consistent Search for Understanding LLM Beliefs**
https://arxiv.org/abs/2212.03827
paper

[14] **Evan Thompson**
https://evanthompson.me/
Person

[15] **Counting Arguments for AI Misalignment**
https://arxiv.org/abs/2310.01405
paper

[16] **Instrumental Convergence Thesis**
https://en.wikipedia.org/wiki/Instrumental_convergence
reference

[17] **4E Cognition: Embodied, Embedded, Enacted, Extended**
https://en.wikipedia.org/wiki/4E_cognition
reference

[18] **Holden Karnofsky**
https://forum.effectivealtruism.org/posts/T975ydo3mx4YnRv4J/ea-is-about-maximizat
Paper