THE TRANSMITTER | BRAIN INSPIRED

Aran Nayebi discusses a NeuroAI update to the Turing test

And he highlights the need to match neural representations across machines and organisms to build better autonomous agents.

9 APRIL 2025 | by PAUL MIDDLEBROOKS

This transcript has been lightly edited for clarity; it may contain errors due to the transcription process.

[music]

Aran Nayebi
The grand challenge right now in AI, even if you don't necessarily care about the brain in particular, is generalized embodied intelligence and the ability to ideally--

Paul Middlebrooks
Is embodied part of that?

Aran Nayebi
It is the ultimate endpoint. If you took two brains, there's just going to be variability between them within individuals in a given species, fix the same brain area, fix the same stimulus. There's just going to be variability in how they process that. When a model is imperfect as a match to the brain, we need to be able to disentangle whether that's because it's actually a poor match to the brain, truly, or because there is inherent evolutionary variability between brains that we need to account for.

The fundamental difference between AI, this technology that we're building, and prior technology is that now you're starting to build technology that takes in inputs and intentionally produces actions. It's an agent.

[music]

Paul Middlebrooks
This is “Brain Inspired,” powered by *The Transmitter*. Hello, everybody. It is Paul. This is “Brain Inspired.” Oh, we just had our second complexity group discussion meeting. It was so fun. I'm worn out from it, but it was a lot of fun. If you are interested in complexity, a large group of us, I think there are 325 people now on this email list, are going through these foundation papers that I spoke with David Krakauer about a few episodes ago, and we just had our second meeting. I'm learning a lot, so it's awesome. I hope if you're part of it you're enjoying it.

Okay. Welcome to this episode. Aran Nayebi is an assistant professor at my own Carnegie Mellon University. He is in the machine learning department. Aran was around in the early days of using convolutional neural networks to explain how our brains perform object recognition. You'll hear me allude to Dan Yamins who was on one of the first episodes of this podcast. Aran, although he has been through many labs with big names that you've probably heard of, Dan Yamins was one of them, but it was of note to me because it was where I began this podcast. He was beginning to actually do the science.

Anyway, since then, he has had a whirlwind trajectory through different AI architectures and algorithms and how they relate to biological architectures and algorithms. We touch on some of what he has studied in that regard, but he also recently started his own lab at CMU, as I mentioned, and he has plans to integrate much of what he has learned to eventually develop autonomous agents that perform the tasks that we want them to perform in ways that are at least similar to the ways that our brains perform them. We discuss his ongoing plans to "reverse-engineer our intelligence to build useful cognitive architectures of that sort."

We also discuss Aran's suggestion that, at least in the NeuroAI world, the Turing test needs to be updated. The Turing test is this famous benchmark test proposed by Alan Turing somewhat as a thought experiment/joke. Apparently, if a computer can trick a human into thinking that the computer is a human, it passes the Turing test, which means that the computer is thinking. It's been debated whether this is a good test over many years, but it is the thing that people keep coming back to when trying to assess whether an artificial system is a thinking system or at least a good artificial system.

Aran thinks that we need to update this, so the original test was just about the behavior, whether a computer could trick a human. Aran's point is in the NeuroAI world, where we're trying to build models that mimic human behavioral output, human function, or biological function, not necessarily human. It's important to also compare the internal representations between the systems that we build, but also between the species that we're testing them against and between individuals within the populations of those species.

I think I just said internal representations, but that's what he wants to compare, what he calls internal representations, and by that, he means simply the population-level activity in the neural networks or whatever system you're using to build it. Not representation in the sense of mental representations and philosophy of mind or other philosophical notions of the term representation, simply the activity of the populations of units, for example.

Thank you to my Patreon supporters. Thank you for the ongoing support from *The Transmitter*. Hope you guys are well. Hope you enjoy my conversation with Aran.

[transition]

Aran, I literally was just looking up. Dan Yamins was episode number 7 of this podcast, which was like a hundred years ago or something, and now here you are, you matriculated partially through Dan Yamins's lab. You went fast, it seems.

Aran Nayebi
Yes, it felt like time flew. I started in 2016 my PhD with Dan and graduated in 2022, so a couple of years ago, with him and Surya.

Paul Middlebrooks
Surya Ganguli. This is not good, because I think I earned my PhD in 2016, and look, you're way, way ahead of me.

Aran Nayebi
Well, it's been definitely a ride. That's for sure.

Paul Middlebrooks
Has it?

Aran Nayebi
Things have changed so quickly, at least in AI, in those years. I remember when we started, and it almost feels like talking about the old times, but it really wasn't that long ago. When I started my PhD, TensorFlow was not yet out, and we were using Theano, and Keras had just come out, so as a master's student, I was contributing to that a little bit, which is fun.

Paul Middlebrooks
You have a background in math and computer science?

Aran Nayebi
Yes, that's right. I did my undergrad in math and symbolic systems, which at Stanford was basically the cognitive science major. I was always interested in the brain. I just didn't know what to do with that interest until later on, basically. Then I did a master's in CS in AI to kind of transition to a slightly more empirical field. By then, I kind of had-- I took--Actually, I had a meeting with, very graciously actually by Bill Newsome. He had mentioned that, "Oh, we need more people with math backgrounds, basically."

Paul Middlebrooks
Back then. [laughs]

Aran Nayebi
Back then, to contribute to neuroscience. He actually referenced the Dayan and Abbott, the theoretical neuroscience book. I was really excited. I was like, "Wow, there's a whole thing," like information theory, and there was a little bit of machine learning there. I was like, "Well, okay, I should get myself familiar with that a lot more than number theory and logic, like theoretical computer science, which is more of my background. That's what led me to do a master's in AI, actually to prepare to do basically theoretical neuroscience, at the time.

Paul Middlebrooks
We could go many different ways here, but-- I mentioned Dan Yamins because those early convolutional neural network models that accounted for brain activity were one of the things that got me into-- an interest in using AI models to study brains. It was one of the early successes, and you were right in the thick of it in the early days. Coming from your background, that mathematical, computer science background, it's so non-biological, right? It's like going towards biology. What I want to ask you is, from perception to embodied agents and come to appreciate how you think of what the brain does and studying the brain. [laughs]

Aran Nayebi
Totally. Like I mentioned, I was super gung-ho about, like basically applying math, very theoretical thinking to questions, whatever they may be to the nervous system. That was me as a junior or senior in college. I didn't know how to do that, but when I saw these books, there was at least some hope. What became kind of quickly clear, there's this book that I started to read called *Spikes.*

Around that time, basically, as a senior, as I was kind of transitioning to do a master's, I was like, "Well, I should actually work in a neuroscience lab to really start to speak the language a bit more of biology" because I've done nothing biological at this point. I was very lucky to have the opportunity to work with Steve Baccus at Stanford. He's a very famous retinal neurophysiologist but also a computational neuroscientist in his own right.

He always encouraged thinking mechanistically, "Don't be led by just the numbers and the fanciness of the technical model." That was really helpful in my own development as a very math and CS-focused student to engage more directly with biology. You mentioned ConvNets. I actually didn't hear about Dan's work until we did a lab meeting in the Baccus Lab basically about that paper, which by then was already like a year, year and a half later. [crosstalk]. Yes.

Paul Middlebrooks
Wait.

Aran Nayebi
Yes.

Paul Middlebrooks
What does it mean to think mechanistically? What did he mean by that?

Aran Nayebi
He meant literally what part of your model is corresponding to biology and what is the question, what's the scientific, or biological really, question that you're trying to answer than just making a more technically fancier model that might predict--

Paul Middlebrooks
What does it do? Like MARS computational-level kind of question?

Aran Nayebi
Even more so. In the retina, you have multiple cell types, so how do these cell types, for example, talk to each other and collectively yield a particular type of response to a particular type of stimulus. There's different types of-- bipolar cells, there's amacrine cells, and then there's ganglion cells, there's horizontal cells at the beginning there, which are more linear in their response profile. At least in the retina, this very clean and clear mapping to a particular computational model and each of those components.

Paul Middlebrooks
What would the opposite of-- I'm sorry, but I want to know where he was coming from and what you took from it. What would be the opposite of that, let's say, in the retina, non-mechanistic thinking or--?

Aran Nayebi
Totally. It's harder to do this in the retina, I think, because there's just so much ground truth, the non-mechanistic approach. You could imagine fitting-- Just to put in the language of today, you could imagine fitting maybe a transformer to that data, to just retinal data, and then being like, "Well, okay, you did a direct data fit, and on held-out response patterns, you do well." That's it.

Paul Middlebrooks
Like massively predictive, basically?

Aran Nayebi
Massively predictive, and then there's no clear-- like you don't check if the internals at all develop anything that maps onto the interneurons that are there in the retina, that sort of thing.

Paul Middlebrooks
Okay. Sorry, I interrupted you. You mentioned *Spikes* also, and I feel like *Spikes* is one of those books that is sort of like *Gödel, Escher, Bach* used to be or something. There's this massive tome by Douglas Hofstadter called *Gödel, Escher, Bach, GEB.* Oh, he's going to grab it from his shelf. There it is.

Aran Nayebi
I have it.

Paul Middlebrooks
Have you read the whole thing?

Aran Nayebi
No, I read parts of it. It was very popular in college, but I--

Paul Middlebrooks
Nobody has read the whole thing.

Aran Nayebi
Nobody has read it, yes, I can't say I have, but I have it still.

Paul Middlebrooks
All right. That's another influence on you. Anyway, I was going to say *Spikes* is one of those books that has been a major influence,

Aran Nayebi
Yes.

Paul Middlebrooks
-and there's the *Spikes* book, but that was 1992? 1999?

Aran Nayebi
I think 1998. My copy says 1996 actually, but MIT Press Paperback Edition 1999.

Paul Middlebrooks
I got it right there at one point.

Aran Nayebi
Yes.

Paul Middlebrooks
But it's so nonbrain-like that book because it's all information theory, et cetera.

Aran Nayebi
Right. We were talking about ConvNets. Around that time, the ImageNet benchmark was a thing, and ConvNets were the main way to do at least neural network-based vision, and were very promising. That's the context at this time. We were thinking, in the lab, Lane McIntosh and Niru Maheswaranathan were working on. They were taking the first deep learning class basically on ConvNets at this time. That was taught by Andrej Karpathy and Fei-Fei Li and Justin Johnson, I believe.

As their class project, they were like, "Well, let's build a convolutional model of the retina. It's a shallow neural network at the end of the day, let's build that." By the time I joined, that would have been during the semester, and then I joined in the summer as a master's student as part of that project. One of the reasons why it felt very motivating to work on that was actually because of *Spikes*. I was drawn to *Spikes* because it was very mathematical, and it was related to the system that I was going to be working with, the retina. I was like, "Wow, there's all this beautiful information theory, and this can be great."

Paul Middlebrooks
Mathematically tractable. Something you can--

Aran Nayebi
Yes. Exactly. It's something you can do and very much spoke to what I was familiar with, so it was a great way to bridge my interest, actually. One thing that really stuck with me in *Spikes* was a passage there that was saying the natural scenes have many, many parameters.

You can prove optimal filter guarantees with things like white noise, what's known as Bussgang's theorem, but with the optimal linear filter, with natural scenes, not only do you not have that guarantee, but unlike a lot of the stimuli that we tend to probe in the retina, which are controlled by one parameter like the intensity of the light and it's a 1D stimulus that varies in time, high intensities can step down to low intensity. There's just like an infinity of parameters that you could imagine could control natural scenes. That hampers our ability to understand the circuit as a result under the condition of natural scenes.

Paul Middlebrooks
Natural, yes. All of a sudden, you're in the real world and you run into a hammer, yes.

Aran Nayebi
That's exactly right. Actually, I think I have it highlighted here. It's like--

Paul Middlebrooks
Oh, come on. Now you're just showing off that you've engaged with it.

[laughter]

Aran Nayebi
Well, as a student, it was just like I had no other reference, and this was the only way I could really start to make a connection with what I already knew, so it's really powerful. Then I just didn't read the rest of the book at that point.

Paul Middlebrooks
Oh, really? [laughs]

Aran Nayebi
Yes, because I was like, "Well, it's not doing-- It is very faithful in saying that these aren't maybe the right set of tools to engage with natural scenes," but ConvNets were that tool.

Paul Middlebrooks
Okay. It's weird though that *Spikes* is kind of the polar opposite of *Gödel, Escher, Bach,* because I think of *Spikes* as dry and clean and *Gödel, Escher, Bach* as meaty and wet somehow, although a lot of people say they got into computational neuroscience because of *Gödel, Escher, Bach.* I guess in that way that they are alike. I don't mean for us to go down the road of comparing and contrasting the two books and styles, but yes.

Aran Nayebi
Absolutely. Early on, I would say, when I was in college, it was all about cognitive science and philosophy of mind, and those are just incredibly deep and interesting topics. I actually took a philosophy mind course with the late Ken Taylor, who was a philosopher of Stanford, the only actual African American philosopher on the faculty, really remarkable man and a very clear orator of the problems and issues about the mind. One of the things that always stuck with me in his class was like, "You and I all have minds, but we don't have access to how they work."

It's not like you're in your head and you're like, "Oh, my visual cortex is relaying information to my prefrontal areas or something, then there's this thalamocortical loop that's now active." You're not doing any of that. You're not even aware of it, and yet we all have minds, and because we're not aware, weirdly, we don't have access to their internal workings, we don't really know-- Like there's all these mysteries as a result, despite all of us having minds. All to say, what pulled me in initially was that obviously studying the brain is the deepest philosophical object. You could study, and it's about the nature of our condition, yet it was so mysterious.

The only way to really engage-- As I mentioned a little bit earlier, I was in the more of the logic and philosophy community initially. One of the things that was very interesting was that there was an annual logic conference that would happen at the Center for Language and Information, CSLI it's called at Stanford, and I would go as an undergrad student. I would go to the graduate logic seminar, and then I'd go to this event.

One of the things that really stuck with me around that time when I was starting to transition to neuroscience was the logicians were saying, "Well, look, we have all these theories in the philosophy of mind about how the mind and brain should work, but unless you do an experiment, there's no way you're going to answer--"

Paul Middlebrooks
The philosophers were saying that?

Aran Nayebi
Actually, a logician, his name was Peter Koellner, he is at Harvard now. He is a set theorist, actually works on those things, but has some interest in philosophy of mind and was saying this, which really stuck with me. That's why then kind of went away from the *Gödel, Escher, Bach* stuff to *Spikes* and more like drier science. [laughs]

Paul Middlebrooks
Okay. By way of story, perhaps again, so you were there in the early days of the convolutional neural networks, and fast forward to today, and you want to put-- let me see if I can state this, and then you can correct me. I'm going to state it incorrectly on purpose. You want to build a cognitive architecture of four or five different types of deep learning models, -esque things, put them together, have them talk to each other, and build working agents that are behaving. There's so many ways I could have said that, and that was terrible, and I'm sorry, but correct me.

Aran Nayebi
No, that's right in the primary essential, but what's the motivation, I guess, to begin with, which is that the grand challenge right now in AI, even if you don't necessarily care about the brain in particular, is generalized embodied intelligence and the ability to ideally--

Paul Middlebrooks
Is embodied part of that--?

Aran Nayebi
It is the ultimate endpoint, but it doesn't-- I think a lot of the major conceptual issues are actually non-embodied. In other words, they're just building even digital agents that can, for example, not go into self-loops and can plan and reason and adapt to new situations. I'm saying this in ways that are clearly things that animals and humans do very well is they're lifelong learning agents, and that's really what we want. I think a lot of the core software issues or the cognitive architecture issues will have to be even addressed in these non-embodied contexts, but embodiment is the ultimate goal.

Obviously, there's details there about robot hardware and things like that, that maybe are not necessarily the core focus. I agree with you that it's really more about this cognitive architecture and making sure it works in open-ended settings to have these lifelong learning agents but also that we can use these as providing insight both about whole brain data that we're able to collect now and are emerging but also leverage that cognitive inspiration to build these more general-purpose agents ultimately.

Paul Middlebrooks
Let's get into reverse engineering. Maybe you can-- I bastardized a summary of what you're up to these days, and we'll talk about NeuroAI Turing tests later, but what is your cognitive-- Do you consider it a cognitive architecture?

Aran Nayebi
I consider it a cognitive architecture for two reasons maybe. This gets back to your point asking about what reverse engineering means. I think there's many different definitions, and I can tell you what my working one is.

Paul Middlebrooks
Yours is going to be the Jim DiCarlo one, right? That's my guess.

Aran Nayebi
Yes, perhaps. I think it's closest to that. Absolutely. I think it's really about understanding the relevant aspects of biological intelligence, those details that are useful for intelligent behavior. It's not about necessarily full-brain emulation or emulating every biological detail, like say the Blue Brain Project, for example. It's more about just isolating the abstractions from biology that are basically hardware-agnostic algorithms that you can then-- that are implemented in brains, but also can be run in hardware and abstracted into machines.

What does that really mean in more concrete terms? Well, it usually involves matching the population-level representations of a model, that's activations, to a neural population activity. It could have been any biological observable. The brain is a complex object. It could have been the dendrites or it could have been the neurotransmitters and maybe for certain questions, that is the relevant biological abstraction, but I think for a lot of intelligent behaviors, empirically at least what we found, there's no necessarily theory here, it's just empirical observations in different brain areas and different species, is that matching at the level of population activity is constrained by doing an intelligent behavior.

There's a relationship between biological observable and intelligent behavior. To be honest, if you wanted a one-sentence summary of my entire PhD, was just showing this paradigm of a task and an architecture and also learning rule was a useful way to interrogate those kinds of questions across brain areas and species, that it wasn't restricted to macaque ventral stream, for example, or human behavior, but actually, a lot of these other brains, and these brain areas in rodents and in hippocampus or higher cognitive areas, they can be understood through this lens of non-convex optimization, basically, the devil was in the details of what those loss functions are and what those architectures are, and that's the language.

Paul Middlebrooks
All right. Maybe it's worth mentioning here, I've mentioned it a lot on this podcast, but the reason why convolutional neural networks became so popular is because they are a multi-layered deep learning network. When you train them in a task in the early days, visual object recognition, and you look across their layers, you can actually match what was then the population level "representations" in different layers and match them with different layers of what we think of as the hierarchy in our ventral visual stream that we think of as being important for object recognition.

There's been lots of work since then. You have done this work yourself, adding recurrence, et cetera. That's one of the modules in your cognitive architecture?

Aran Nayebi
Sure. Yes.

Paul Middlebrooks
Then what are the other ones and why?

Aran Nayebi
I should say that I'm not married to this particular set of modules. It's meant to be free. In fact, a core question is, ultimately by doing these types of comparisons of now agent architectures, cognitive architectures, to whole-brain data across species as well, that we can start to understand if there's conserved modules, if there's a general-purpose architecture that emerges through lots of empirical comparisons.

Paul Middlebrooks
You mean across species, general?

Aran Nayebi
Across species, yes, exactly. Basically, across this species-conserved sensory-motor loop, so converting inputs to actions, that's why it's an agent, rather than one module is like one large-scale set of brain areas, which is how we've traditionally done comparisons in NeuroAI, but now what we want to do here is really engage with whole-brain data that's coming online and start to understand how these brain areas interact to give rise to complex behavior. I think that's why the agent naturally fits in here.

The idea would be that some natural starting points for modules are yet a sensory module, not only vision but also multimodal, and there's some evidence which we can talk about at some point about how maybe they have similar or actually self-supervised loss functions across these different sensory modalities. There's a kind of unification there, but maybe they have different inputs, but that's something that we could talk about. That's why I group it as a sensory module, but then there's a future inference or world model, which I think is really the hardest-- I think computation is the hardest part of this.

I think we've made a lot of progress in sensory systems. Obviously, there's work to be done still there, especially the type of work that still remains to be done, the sensory system connects with this world model. Basically, being able to have a model of the dynamics of the world, of your environment, but not like a model of real physics, actual physics. You and I have the maybe intuition that two objects, will fall faster if one is heavier than the other even though we know they should both fall in a vacuum at the same rate. I don't mean actual physics, I mean just intuitive physics, like what's intuitive to us and how to predict things.

Paul Middlebrooks
How does that differ from the Tenenbaum physics engine approach?

Aran Nayebi
It certainly, basically, is the physics engine. It's just that the difference is that we aren't assuming symbols as the input. We aren't assuming a program or anything like that. We're actually assuming unstructured visual inputs coming in, being processed by a sensory system. The output representation of that sensory system is then fed into the world model. It's visually grounded, basically, or sensorially grounded rather than-- we kind of assume that there's a particular type of output format that vision gives us, and then we proceed with that more symbolically.

Paul Middlebrooks
Okay. Got you.

Aran Nayebi
Why maybe is that distinction actually important for function? Well, because humans and animals operate in an unstructured environments-- in a wide range of environments. Whereas when you handcraft those inputs, that might be useful for studying particular environments, but it's very hard to then generalize to the open-ended, unstructured environments that we all naturally engage with. That's the key here, is being able to deal with open-ended environments.

At a high level, that's module number 2, is the world model, so sensory, world model, and then there's planning, which I don't know if it should be part of necessarily distinct, but just to distinguish the fact that if you have a good world model, also planning is easier, especially long-range planning, and maybe there's a hierarchy there of timescale.

You might plan at a high level and then fill in the details. At a high level, you might have abstractions that allow you to do more longer-range planning. For example, when you do decide to get out the door and then go to the Mellon Institute, in that case, you don't plan every step, you plan at the level of landmarks, or large-scale things. That's what I mean. Then the rest of your body fills in those other details, the fine-grained motor commands, et cetera, and that brings me to the last point, which is the motor module, which then executes these high-level commands, and maybe in a hierarchical way. Then I guess maybe there's a final module if it's helpful to think in these ways.

Again, I expect these to all be approximate in the actual brain, this is just more for conceptual clarity to frame things, is intrinsic goals. In other words, how do we guide what plans we care about or select, and therefore the actions we execute? Well, we can leverage the world model by planning through specific types of actions, but those are guided by intrinsic drives. Unlike reinforcement learning in games like Go or Chess, where you have a very well-defined reward function, and that's where RL thrives.

In real environments, there isn't one. Animals do rely on things like different behavioral states, hunger, pain, et cetera that are both built in, but ultimately, you can think of there's built-in intrinsic drives, but also more learned ones that might support more open-ended learning to seek out more information beyond their pre-training data. I just put that in the language of AI, basically, if you want an LLM agent even to go beyond its pre-training data, you want to specify the right autonomous signals, which is still an open question to do that and to adapt online. Those are the five modules, sensory, world modeling--

Paul Middlebrooks
For now you say?

Aran Nayebi
Yes, for now. Sensory, world modeling, planning, motor, and intrinsic goals, how they're combined will matter. I just assumed a feed-forward one for now. Of course, I expect there to be back connections, but the other aspect of it is that these modules aren't fixed. The standard paradigm in AI is to fix things and then do a test time, evaluate it.

Paul Middlebrooks
What is that? Oh. Oh. What do you mean fix? Freeze the parameters?

Aran Nayebi
Freeze the parameters, and then evaluate a test time. What we want to do, and ultimately, obviously there's plasticity in the brain, is for these animals, or these agents to adapt online, we need to specify update rules as well for those modules, so not only how they interact but also how they update online to new challenges. That kind of speaks to the learning rules aspect of NeuroAI, which I think is less touched on, and we can touch on it too, and there are things to say there, but that's the high-level approach.

You can think of, for example, test time reasoning that people have now where you're trying to get the LLM to reason online either via chain of thought or trying to do this in an online setting rather than just scale up the pre-training data, like just get it to reason at test time as a special case of this broader goal of getting these modules to be adaptive.

Paul Middlebrooks
Let's talk about plasticity. You just said that it's not so much the focus of NeuroAI, but has it gone out of favor? Is that a thing? Because, for a while, it was all about, "Oh, backpropagation, it's not brain-like, and we need to figure out," but there are lots of people working on synaptic learning rules, et cetera. I'll just jump the ship here and say, and you can correct me again, these modules that you say that are sort of up in the air, but they don't have to have the same learning algorithms necessarily in different parts of the brain, have different learning algorithms. That's something that you want to figure out, but people aren't focusing on learning anymore.

Aran Nayebi
Not anymore. I would say it just wasn't-- Oftentimes in NeuroAI, we're not modeling the developmental process.

Paul Middlebrooks
Yes.

Aran Nayebi
Right? Often, you might hear phrases like, "Oh, this backprop is basically a proxy for evolution, or it's a proxy for evolution and development," and we don't cleanly separate those things.

Paul Middlebrooks
Oh, okay. One way to say this is like, we have something that works, so we're going to use it and worry about it later. Is that--?

Aran Nayebi
It's like that, it's also that we're not explicitly modeling it, and you can't in the standard framework where you just train something of backprop and you get to the adult state of that particular brain area, I think of pre-training as more like, you can imagine you pre-train these modules to get to a desired state, but unless you have an agent, you can't actually go and test an update rule online, not through batches but through online interaction, can you more faithfully disentangle the evolution part where the pre-training stops and where the module updating begins, the more developmental aspects?

If we wanted a more formal computational grounding on development and we wanted NeuroAI to engage with that, I think that's why the agent-based approach would more explicitly speak to that by disentangling those two things, those assumptions that we're currently making and we're kind of lumping into backprop. Furthermore, you mentioned that there was a lot of work actually on biologically plausible learning rules, and there has been. I myself have worked on it with folks where the big challenge, say up till 2019 or 2020, was from 2016 to 2020, basically, there was a flurry of activity to understand backprop in the brain.

Paul Middlebrooks
Right. It's almost to show that backprop happens in the brain [chuckles], right?

Aran Nayebi
Yes, that's right, because we all use it for training these networks, is very successful, and so the natural inclination was, "Well, some version of it might be in the brain. We should go look for it. What would that be?" The trouble was that we couldn't scale any-- What's the biggest-- There's a few gripes, which are very well-motivated gripes about backprop as a non-biologically plausible learning rule. I would say the main one, just to keep things succinct, is that the forward and backward weights are always tied.

In other words, an error update in a feed-forward network always involves the transpose of the forward weights of each layer. Oftentimes, when we think of implementing backprop in the brain or as an update rule for a module in this kind of more embodied agent framework, we assume that would be a separate network that is basically computing the errors. If something like what's called weight transport is necessary, then that circuit would actually have to have an exact copy of the forward weights, and that's weird, at every time step, you need that, and that just seems very inconsistent with the fact that biology is very noisy and messy and non-robust to-- I mean, you know what I mean?

Paul Middlebrooks
Right.

Aran Nayebi
It's like a very--, but it's very [crosstalk]--

Paul Middlebrooks
I'm sorry, but people like Tim Lillicrap have shown that you don't need to do it that way, and you can approximate backpropagation. I'm sorry, this is a total tangent, but you can approximate it almost randomly with some feedback.

Aran Nayebi
That's the thing. Actually, that ended up being a-- was a control that they ran showing that it shouldn't work. That's what Tim told me, actually. Showing it shouldn't work, and then it did on MNIST.

Paul Middlebrooks
[laughs]

Aran Nayebi
They were like, "Well, we should investigate this." It was a very, very interesting story.

Paul Middlebrooks
Oh, that's great.

Aran Nayebi
One thing that motivated me to work on was that actually Tim gave a talk at COSYNE 2016, that they tried to take their feedback alignment algorithm, which is this thing of replacing the backward weights with random weights and scale it up to deeper architectures and on harder tasks in MNIST, so like CIFAR-10, CIFAR-100, ImageNet, and especially ImageNet, because that was the main vision dataset that if trained with backprop gave you neurally plausible representations that actually predicted brain data, not MNIST, for example.

They wanted to scale up to that, and he was saying, "Look, guys, the moment we do this, it just fails at these harder tasks, there's this bigger and bigger gap that grows with the performance of backprop." At the time, I didn't know what to do with that. It was very interesting, and I knew I wanted to work on it. It wasn't until 2019 or so that maybe there was some evidence that updating the backward weights. With what rule? That's the key question. Could start to patch that up but not completely.

With Dan Kunin and Javier Sagastuy-Brena and Surya and Dan, we developed a broader language of basically looking through space of update rules and on the backward weights, so rather than keep them random to update them. Once we had a library of primitives based on things like energy efficiency, improving the communication clarity between the forward and backward updates, that sort of thing, then we could kind of do a bit of a search there.

We started on smaller-scale experiments finding just what was working and starting to scale with ImageNet that we could then find that ultimately through a larger-scale search that something like an Oja-style update, though not exactly, but just to summarize it. Oja-style update was good with a few-- Anyway, once you had the language for it and update rules, you could finally close this gap. The main other issue was that even if you close the gap on one architecture like ResNet-18, as you went to deeper models, the same hyperparameters of your local learning rule didn't actually transfer.

This is actually unlike SGD. In backprop, that transfer does occur. That would be really unfortunate from an evolution point of view. Every time you create a new organism, you have to do another search of the hyperparameters. We found actually robust primitives that were hyperparameter robust, and then you could transfer two very deep architectures as well-- Once we had that, then there was like an N-of-1 example of like, here's a learning rule that's vector error based but doesn't require the weight symmetry that backprop needed. Then maybe that's starts to become a plausible candidate to look for in brain data.

Paul Middlebrooks
All right, but it's not the hot topic these days, I suppose.

Aran Nayebi
Not anymore. Yes, I wouldn't say it is. There was a question of like, what brain data should you measure? We had some work on that using artificial neural networks, that was follow-up work, and showing that the activations are actually enough.

Related to my earlier point about how model activations correlate with intelligent behavior, tracking those changes across time also correlates with better identifying the learning rule in artificial neural networks where you have ground truth and you can weaken that with noise and limited observations to start to mimic what we actually get in the brain and it's still robust to that unlike synaptic weight changes which are the more natural thing to look at, and activations are much easier experiment to do. You just do ephys as opposed to tracking the dendritic spines which is a much harder experiment.

There is an open question of like, can we go and validate this in data? We have some evidence that it's a much easier experiment than previously thought. Yes, as you say, it hasn't been the main focus of the field nor of at least of my own interest at the moment, but I think it's important. Basically, I want to get to an agent architecture that does work, but then we can then start to study those questions of the module update rules development and those things. Those are interesting down the line, but they're just not the immediate interest because there's just upfront challenges to begin with to get to the modules to some initial state that's actually good and they combine well to begin with.

Paul Middlebrooks
You want to make an agent in a robot, or is that a more longer term?

Aran Nayebi
The robot is a longer term. Right now it's all in sim with biomechanically realistic bodies. Biomechanically realistic bodies also have a much larger number of degrees of freedom than current robot bodies. Let's say SPOT is about I think 10 degrees of freedom or something like that, but of course, they have lots of low-level control. There's a beauty in the hardware stack. We wanted to more focus on the actual biomechanical control aspect of it, the high degree of freedom biomechanical control aspect of it.

That's why we're doing it in sim where you don't need hardware to approximate it, you can actually be more exact about it and because a lot of neuroscientists do VR experiments, there's no gap in the eval, like you can literally take the same stimulus in sim and just put it in a new simulated world that matches what the experimentalists did. Experimentalists like sim stuff because it's very controllable and repeatable. Now it closes the gap with the evaluations as well.

Paul Middlebrooks
Right. Yes, and it's not real. You're very well aware of Moravec's Paradox that the hard things are actually kind of easy, so we can play chess in computers really well, and the things that we think are easy like ping-pong are hard, like physically doing ping-pong, because you can simulate it with however many degrees of freedom that you want in an agent and you can control it.

Then once you get in the real world, it's all of a sudden hard. The reason why I'm bringing that up is because one of the things I wanted to ask you which I think is related is, since the early GOFAI days, there have been a lot of cognitive architectures, and it's kind of transitioned from symbolic. Then you have hybrids like Chris Eliasmith's SPAUN and you have Randy O'Reilly that are making really more connectionist-type cognitive architecture-type things.

Without fail, I believe, maybe this is not the case for Randy O'Reilly, but a lot of the people who have worked on these things have suggested that it's not so difficult to actually get one module to perform well. It's actually the crosstalk between modules, the control between the modules, that's the difficult thing and it seems to take way more effort than actually getting the modules themselves to do what you want them to do. Is that on your--?

Aran Nayebi
I actually agree, but I would say that it comes with a nuance though that it depends on your goal. If your goal is more open-ended, unstructured environments, it is also a challenge in itself to build the modules like the world model or it took a while to even get a good sensory encoder. If you go beyond particular tasks to more open-ended tasks, it's already a challenge in itself to get to that pre-training, right?

Paul Middlebrooks
Yes.

Aran Nayebi
We didn't have good SSL things till a few years ago, loss function. Advances in better SSL algorithms also led to better brain models in, for example, mouse visual cortex where it's not categorization-optimized, so all to say definitely connecting the modules is not trivial, but even constructing the modules, especially I think nowadays with the world model and figuring out what those representations even should be, is actually I think still a core challenge.

To give an even more concrete example for today, a very common approach right now is VLM, vision language models. We can go and collect tons of robot data in the world, like people drive Teslas and they have a century's worth of data. It's not hard to get lots of data actually. It's not the bottleneck. It's that the scaling laws have not been as favorable on those types of data with the existing VLM architectures. Even if we wanted to go back to the sensory module, the scaling laws haven't been as favorable as they have been in language.

I think part of this is the architecture itself, in particular the way we tokenize, in other words, the way we process the inputs to give to these specific VLM architectures are random patches, and so they end up basically learning something like a convolution, which doesn't end up being ultimately a step change or an advance over what we had with CNNs. They're actually converged. It's like the platonic representation hypothesis.

They're like converging on very similar representations, and also as a result, vision transformers are also very similar matched to the brain as CNNs are because they're effectively approximating convolution. I think that in language though, the notion of a token, individual constituent words, is very much semantically related to that modality is trying to do. Combining words gives rise to the higher meaning, but the patches themselves are-- We don't have a good prompting language for vision yet, and I think there's still an advance there to be made even in that domain.

Paul Middlebrooks
Are these the sorts of issues that made you start focusing more on what you have written as Marr's algorithmic level rather than—Marr's three levels, right? Everyone for the past 10 years have been focused on the computational level and that's what AI focuses on. For example, you have written and many people have, that, "Oh, the reason why these models work so well is because we give them a goal, give them a task." Right? That's the computational level, something that they need to accomplish.

Then the algorithmic level, how they accomplish that algorithmically is somewhat less important, but they can learn them. That's why these models are so great. Then the implementation level, who cares, just you have to put something in there and eventually it'll give rise to it. Are these the sorts of issues that have made you focus more on that algorithmic level?

Aran Nayebi
Yes. I think just to even translate what you just said to, like NeuroAI. The computation level is a task and the algorithmic level is related to the architecture, but also the interactions between the task and the architecture too. In some cases, I think it's not just about specifying the right goals. It's not just about figuring out the right self-supervised objective, like next token prediction or contrastive learning, that sort of thing, which is an advance in itself. It's also figuring out the architecture that meshes well with that modality.

I think there's a lot of promise in using transformers or the token-based paradigm because it is a little bit modality-independent. In other words, it's kind of general purpose. You just swap in different data. I think with that generality, I think not all sensory systems are necessarily equal in that way. I think there's nice-- maybe if there's a lot of shared explained variance between them, and we have some evidence for that.

I think going forward, if we really want to get things that are better at intuitive physics, where a lot of the models lack in terms of human capabilities, we will likely have to start to become more specific about how we at least, to put in the language of today, tokenize or process inputs that are more vision and embodiment based. They may just be different than language. This is probably consistent with how maybe things are in the human brain where the language areas were evolved later and are a little bit topographically distinct from what visual cortex looks like.

Paul Middlebrooks
You don't have language in your modules yet, right?

Aran Nayebi
That's right.

Paul Middlebrooks
You're concerned about also comparing across species, and given that humans are the only species that use language, I know that's arguable, but let's just say it, it might not be what you're after.

Aran Nayebi
That's right. I think that's maybe why I think that there's a core underlying algorithmic desire of us wanting these agents to better understand the world. Animals certainly build models of the world without language, and that's already hard. This relates to, like coming up with a prompting language for vision for VLMs is better than this random image token thing, that I think would better speak to the visual intelligence that animals have, that needs improvement in existing architecture today.

I think where language can play a role, even if you're modeling at that level, is at least like-- you could argue that you're using language when you train a model on supervised categorization. You're providing labels for the images. You could make that argument that language is maybe a useful guide to learn representations, even if they're not in a linguistic context.

For example, if we're stuck on a certain question, we can't come up with the right self-supervised objective that we think an animal could be implementing plausibly, or related to that, like a better prompting language for vision that isn't language-based, then I think it's fine as a proxy at the moment to use the kind of the less good but still gets you somewhere, supervised language-conditioned version of that loss function before you kind of figure out the self-supervised thing. I think that's fine.

It might actually get you quite far. I think we shouldn't throw that away either. It's just not used specifically in linguistic context. It's really more used for guiding representations, which is what we've been really doing. Language is a remarkable ability to teach machines how to reason like us or for us to communicate that, basically.

Paul Middlebrooks
Yes. One of the things that I enjoy about your approach is that you appreciate the sensory and motor loop and the agentic aspect of intelligence and existence. Then, earlier, the way you were talking about it, it sounds very much like an input-output, brain is a computer, metaphor kind of thing. You started in vision, you started in sensation, and you've come to appreciate the motor aspect of it. Some people would turn that around like active inference people and say that actually you're behaving to adjust your sensory input. Where do you land on that?

Aran Nayebi
Oh, totally. Yes. This relates to why having an agent is actually the way, I think, to studying questions about development or online learning rather than the existing way, which is just to lump it all into a single backprop update. This is also related to the goals, whatever goals it's using to guide its exploration in the environment, the intrinsic goals, which is--

Paul Middlebrooks
Where do those come from? That's another thing I wanted to-- we'll put a pin in that.

Aran Nayebi
Yes, no, I think that's an interesting question on its own. Whatever those goals are, they're going to alter the training data. One of the reasons you want the online update rules is to adapt the distribution shift online. If you're going to go beyond your pre-training data, and you're going to explore your environment, you have to also adapt to the fact that you're going to encounter out of distribution things simply by exploring the world a bit. In order to handle that in a robust and reliable way, you'll also need an update rule, and the exploration strategy is guided by these intrinsic goals.

Paul Middlebrooks
Where do those goals come from?

Aran Nayebi
That, I think, is unclear fully. I don't think there's a definite answer yet. We have--

Paul Middlebrooks
The computer scientist way is to then program in the goals. Is that achievable? Can you program in the goals, whatever? Because it seems to be this mysterious central core of our existence biologically is that we have these intrinsic goals. No one knows where they come from. It's an internal reference signal that we have to follow. We want to be at homeostasis. It's kind of a mystery. A computer programmer wants to just, "All right, program in the goal." Is that achievable?

Aran Nayebi
I think it's less like-- This is related to things like reward hacking that people talk about these days in modern agents, which is that we might think we're programming it in by specifying it, but actually, because these things are optimized, just specifying a particular goal might lead to unexpected behavior.

Paul Middlebrooks
An emergent kind of behavior.

Aran Nayebi
Exactly. Emergent, either desired or undesired behavior that's harder to-- If we were dealing with computer programs, quite literally, then yes, it should-- Even then, computer programs can lead to unexpected behavior too. You just didn't fully anticipate as the creator of the program all the possible things that could lead as an outcome. AI safety people have used the paperclip maximizer as the unexpected version of this. It maximizes paperclips, and then it's like, "Well, I should just take everything, take over." That's an unexpected outcome.

Even when you knew it was very myopic and very specific, it can lead to that. To your broader question of where are these goals, I think likely they could be distributed. The obvious candidate for everything that someone doesn't know is to say it's prefrontal cortex, but you could also imagine that [crosstalk]--

Paul Middlebrooks
But a lot of organisms don't have prefrontal cortices in plants.

Aran Nayebi
That's right. That's right, actually. Related to other organisms, there's beautiful work by Misha Ahrens's group at Janelia that shows that zebrafish have a futility-induced passivity that's actually computed in non-neuronal cells, in astrocytes.

Paul Middlebrooks
Oh, really?

Aran Nayebi
Yes.

Paul Middlebrooks
That's cool.

Aran Nayebi
We're actually working with them studying this, related to the intrinsic goals question, trying to figure out what these intrinsic goals are. We should have something out soon. The main thing is that I think these goals can be computed in a lot of different ways and a lot of different parts of the brain that isn't just PFC or something neuronal, but even in other animals, in non-neuronal cells, potentially. Even if it is fully neuronal in other higher species, it could be that that's done in, say, large-scale, say, thalamocortical loops, et cetera. I don't think it's localized, in other words.

In fact, the evidence that we're seeing is maybe that it's not so fully localized, but that animal behavior is still very stereotyped. In other words, from the decade of neuroethology with machine learning that's applied to naturalistic videos, going back to our naturalistic discussion in the beginning, people have found that these ML tools are kind of auto-discovering behavioral primitives that seem to be reliably switched between, more or less.

That's kind of what motivates this hardwired intrinsic goals thing that I'm mentioning here, unlike end-to-end RL, where it's like one objective, there's probably multiple things that it switches between and dynamically state switches between. That might not necessarily have to be localized. Those individual goals might be represented in different parts of the brain.

Paul Middlebrooks
Let's talk Turing test. We're all over the place, and I apologize. That's my fault as the host, but it's fun for me.

Aran Nayebi
It's fun for me.

Paul Middlebrooks
Maybe before we move on, is there something that you want to add on the cognitive architecture slash- I keep calling it cognitive architecture, sorry- the agent, the embodied agent that you're working on?

Aran Nayebi
I call it a cognitive architecture, too. I think maybe the main difference is that what we really care about are open-ended, unstructured environments that humans and animals are in, and actual open-ended tasks. Also, comparison to not just behavior but internal representations as well. Comparing the individual modules to the individual brain areas, and then interactions between modules and online updates therein to developmental signals down the line. This actually nicely segues into the Turing test because the ultimate root of this is quantitative comparison.

Paul Middlebrooks
Wait. Before you start talking about, you've put out, there's a manuscript, it'll be linked in the show notes, is it called the NeuroAI Turing test?

Aran Nayebi
Yes.

Paul Middlebrooks
Okay, which modernizes Alan Turing's Turing test, which he didn't call it the Turing test, but it came to be known as the Turing test, where, ‘if you're a computer, can you fool a human to think that you're a human?’ That was very focused on behavior. You're saying yes, that's great, but in the NeuroAI world, we actually need to also compare the internal "representations", which we'll talk about what that means. You have the behavioral comparison, but also the internal representations' comparison, and that should be a benchmark of sorts.

Aran Nayebi
That's exactly right. The key principle is that for any measure of comparison you want, you want your models to be as good as brains are to each other in the context of internal and behavioral representation.

Paul Middlebrooks
What does that mean? Explain what that means.

Aran Nayebi
Basically, there's two issues here when doing these model brain comparisons at least, which is one is that brains are stochastic. Unlike our models, which are deterministic, they respond variably to the same stimulus. That's been well quantified up till now as the internal consistency of the neurons. The statistical noise ceiling that we call in the paper. That's often been either 100% is used as a ceiling implicitly or that more quantified metric of the internal consistency of the stochasticity of how neurons are consistent to each other across trials has been used.

The other aspect of it is that that doesn't as cleanly semantically map on to how we actually compare models to brains. We're mapping one representation to another, and the statistical noise ceiling isn't really capturing that. It's just capturing the stochasticity of neurons, which is good. We want to capture that, but it's not the only thing. We want to measure a ceiling. If anything, it actually becomes a correction term later on, but the main thing we want to capture is the fact that, even if a brain was deterministic, even if the stochasticity didn't matter, if you took two brains, there's just going to be variability between them. Within individuals in a given species, fix the same brain area, fix the same stimulus, there's just going to be variability in how they process that.

When a model is imperfect as a match to the brain, we need to be able to disentangle whether that's because it's actually a poor match to the brain, truly, or because there is inherent evolutionary variability between brains that we need to account for. That's why we emphasize. Whatever metric you use of comparison, we can talk about that, is, under that metric, you should also do a comparison of the brains to each other as though the brain was another model.

Paul Middlebrooks
Then, is the idea that you have one model against lots of different brains, multiple models against lots of different brains within a species, for example?

Aran Nayebi
Yes. It's still used in the context of integrated benchmarking of having multiple models to at least one set of brain data, one brain. It's good to then

maybe do cross-species comparisons down the line, but this is, at a minimum, the base thing you want to first straighten out is the model to single-brain comparison question. Then, to generalize, it is just applying that same procedure over and over. Once you've established that procedure, what that ceiling should be, it's going to be the same one that applies to other brain areas, other species as well.

Paul Middlebrooks
This is where your mathematical and logic background really comes through because you really specify how these comparisons are going to be made theoretically. You were just talking about the metrics. Let's talk about how to compare these. You leave it open, and so you can compare any metric of whatever you use for the representation, right? You did air quotes.

Aran Nayebi
Yes, [laughs] I know. I think that's very totally appropriate for this because the question of what metric to use is certainly an important one. I think oftentimes as a field, and certainly I have as well, there has been an implicit assumption that there's one platonically good metric that we ought to strive for of model to brain goodness. This is also even reflected in pre-neural AI days, where we had different notions of brain likeness in words like sparsity, energy efficiency, things like this that we wanted our models to have.

Paul Middlebrooks
Oh, right.

Aran Nayebi
When we assume that even in this more quantitative setting, that there's a platonically good notion of a metric, then in some sense, it's like we're saying upfront, we know what it means to match the brain well. You might as well just bake that in, and then you're done. The problem is that that's not the case. That we don't know upfront, a priori, what a good brain model should be. We have data, and we want to match the data as good as brains are to each other under the assumptions of that data collection process. That's the empirical reality of it.

I would also argue that, given that the brain is a complex object, there's different things that people focus on. We were talking about how people have different definitions with NeuroAI. It's like some people really do care about topography, for example, which, a CNN is not a 3D spatial map, it's 2D. Under a topographic metric, it would be zero effectively under that. Different people have different things for the question. I don't actually even think, not only in platonic realities, they're not a platonically good one, but it's just highly question-dependent anyway.

Paul Middlebrooks
You keep saying platonic here, and I know that you've thought about this. There's, -what is it?- the Platonic hypothesis that's been floating around. What is that, and why do we like or dislike it?

Aran Nayebi
No, I just meant platonic in an overall notion of good that we should all strive for.

Paul Middlebrooks
Okay, then let's not go into the Platonic hypothesis then, it's too far an aside, but you mean an ideal that is--

Aran Nayebi
An ideal that we should all be striving for. I think that basically there is no such thing as an ideal. It's just question-dependent, and the brain is very complex, so you might focus on different aspects of it. All we're trying to do here is, we want to standardize that operationally and say, whatever metric you choose for your question, make sure that you assess model goodness up to how brains vary under that measure.

Paul Middlebrooks
Oh, okay. In some sense, it's extremely pluralistic because it allows the user-- Here's what I want to ask is, how do I pass your NeuroAI Turing test and how do I fail it? I can come with my own question. As long as I adhere to the scientific rigor of the NeuroAI test, I can come in with any question or any assumptions as long as I state them. It almost sounds like I can pass it if I want to.

Aran Nayebi
Right. Of course, you could define maybe a trivial metric or something where it's zero or something on that, and then the models are-- Right. Certainly, you could do that, but then you could argue that was what was sufficient for your question. I can't tell scientists what is a good question or not, that's their judgment, of course. The idea is that you define a metric that you want to score model goodness on.

Paul Middlebrooks
Let's say that metric is, you mentioned efficient coating or sparsity. You could just do it on sparsity, you could do it on individual spikes in a population. You could do it on, I don't know, astrocyte calcium signaling. Whatever you want to do.

Aran Nayebi
Yes. It's meant to be extensible, it's meant to encounter the broad range of applicability and diversity of questions that we have in the brain sciences naturally because the brain is complex. I can tell you, though, in practice what I I do, and why, in terms of the metrics that I choose. For a lot of settings, I want to be clear, in vision, I think we have had the luxury of such advances to get us to models that were really good. The most common

benchmark was, for example, HVM, which is the **[unintelligible 01:10:19]** et al. one that people push on, was the initial one that Brain-Score used. Brain-Score now uses a lot more other vision benchmarks, too, as part of it.

For a while, when we were first working with HVM, which is in 2016, 2017, one of the impetus for coming up with this animal-to-animal measure was that, according to the statistical noise ceiling, the models were explaining 50% or 60% of that, and we were like, "Whoa, clearly there must be advances in vision needed to beat this very simple visual behavior of an animal staring at a stimulus and doing nothing else with it the first 150 milliseconds of visual processing." It turned out that actually, when you looked at the animals to each other, and we--

This actually ended up being a supplement in the ConvRNN paper, and then we made it one of the main figures in the NeuroAI thing because even the authors of the NeuroAI Turing test have just relegated this thing to supplement for some reason. Not intentionally, because there's other focuses. On HVM, it wasn't 60%. It was actually 90%.

In other words, now, that's not to say we've solved object recognition. We're just saying, on this particular data set, this benchmark that people have been trying, I don't know if you would really want to go and invest major money to do advances in vision to push on HVM, in particular.

You might instead be motivated and say, "Okay, one of two things. I'm done with the question because this benchmark was the thing that mattered to me most, and I think the vision models are good enough for my purposes." Or two, because of this saturation, you'd be more motivated to say, "I'm going to go collect more data, higher variation data to really push an object recognition, and again, set my ceiling to the animal-to-animal consistency there." For example, we did a bunch of extrapolations of the NeuroAI Turing test on the HVM data.

We found that actually having more conditions and having more neurons didn't start to improve that ceiling. It was starting to saturate at some point, but it motivates that you should actually go and do something more concrete and collect a new experiment there, too. That's another possible viable route, but it does rule out as a viable route, I think, to continue to push on, for example, HVM, where you may have thought the gap was much bigger, but it's actually much smaller than you imagined.

Paul Middlebrooks
What is a representation? When you use the term representation, what do you mean?

Aran Nayebi
I actually, 100% of the time, mean the population activity. That's all I mean.

Paul Middlebrooks
I think that that's a common usage in neuroscience, but then, of course, there's the philosophy of mind way of using it, cognitive sciences, so on. It's a slippery term. It's deflated in that sense, in that it just means the activity of whatever you're studying.

Aran Nayebi
Yes. I've had a hard time understanding, and maybe you could tell me about this, but understanding the arguments in cognitive science that aren't population activity and why it's a more nuanced term, but I've honestly literally meant it as a vector of population responses, of bind firing rate, so very precisely. It's a very precise notion. I guess, from my understanding into the cognitive sciences, it's like, does that vector necessarily contain all the semantically meaningful stuff that we're assigning to a behavior? I think that's what the--

Paul Middlebrooks
I don't like to do the etymology thing or tear apart words, but represent is re-present. I think it is the idea of, okay, this is presented in my mind, and it is attached to the thing in the world, and it's somehow a copy of that. It's re-presenting in my mind, which is very different than just a measurement of activity in some brain region.

Aran Nayebi
Yes. Oh, and we can see that if you go to visual cortex, there are some efference copies from other areas representing other things that you can decode.

Paul Middlebrooks
That's the thing if you can decode a lot from a lot of different brain areas, but is that meaningful? Is it causal? Is it correlational? Is it a representation? There's a lot of talk these days about being more careful with the term representation, but the way that you use it, you don't have to be careful with it. Maybe we just need a different term because it just means the activity.

Aran Nayebi
Yes, I literally just mean the activity. For me, that's always there. The question of the semantic representation or something like that, if you want to call it a slightly different term or maybe a completely different term later on.

Paul Middlebrooks
Do you mean attaching meaning?

Aran Nayebi
Meaning to it. It emerges from this pre-modern neuroscience view of there's this one function that you can assign to this one cluster of cells in the brain, and that's where it is. I think that's, by and large, probably a very toy view of the brain because it's quite distributed in ways that we don't expect. I think the only way to really get at those kinds of questions is not to go in assuming that there is a very localized function. In some cases, there are, but not in a lot of cases. Then do those kinds of quantitative comparisons between these--

That's why, to engage with whole brain data, you want to go in this embodied agent direction because, ultimately, these things do interact and do influence each other. You want to test whether that hypothesis of the interaction between the modules is a good one. The only way you can do that is to have the modules, to have them interact, and then do this kind of NeuroAI Turing test on top, just to combine all these ideas together at the level of population activity to begin with, to really assess, does this vector of population activity contain that semantic content? I think you can answer that in a quantifiable way. Yes, no, or 0.65. It's not 1.0 or 0, basically.

Paul Middlebrooks
It's interesting, though. You just said something that I very much jibe with and appreciate that, back in the phrenology days, we said brain area X does function Y. You said that that's not necessarily the case. It's probably not the case. It's very distributed. However, the thing that you're wanting to build is made up of modules that do functions and then have to talk to each other. How do we reconcile those things?

Aran Nayebi
That's an excellent, excellent question. When I say function, I'm not saying do the Jennifer Aniston recognition or anything like that. It's a more general purpose. It's such a general loss function anyway that it's actually highly nonspecific. Yes, there is some localization, but it's like, for example, for self-supervised learning, next token prediction or a world model, it's trying to figure out the dynamics from the current state to the next state. That's a very general thing. It's not as specific as maybe in the phrenology days, where it's like, oh, it's this specific thing, like recognizing a particular person or places alone or things like that.

I'm not saying that the brain doesn't have it. I think that maybe through this optimization of a general purpose, high-level goal, you can learn internal representations that do have more specific content to them. I think that's entirely possible, and we know that. Face patches emerge, that sort of thing. It's just that the kind of top-down guides to build those high-level modules, those large-scale modules, are a lot more general purpose than specific.

Paul Middlebrooks
Okay, fair enough. Aran, what's holding you back these days? Two questions. What are you excited about, and then what's in your way?

Aran Nayebi
That's a good question. What I'm excited about is ultimately, even beyond specific scientific questions we can ask, we are entering an era where we're just building more capable systems. Even modern agents that we're building today with LLMs, they start to have cognitive components to them. People are starting to realize, oh, you need a memory, things like that and compositionality, modularity. That's good. It's consistent with how the breakthrough is needed in AI to get to the next generation have also led to better overlaps with the brain, partly because there have just been fewer solutions to get there.

Since the brain has already reached that, there's a high probability of that overlap. That's the contravariance principle, or in AI, the Platonic representation hypothesis that we've referenced a couple of times now. What I'm excited about is that we're entering an era of more and more capable systems. I do feel that what used to maybe be a sci-fi dream or even a dream, it was still it felt a little bit within reach, but still a faraway dream of 50 years back when I started with neural networks.

Paul Middlebrooks
Are you about to say it's all going to happen within five years?

Aran Nayebi
No, I'm not about to give a timeline, but I do think that it's a lot sooner than maybe, or even a capable, I don't mean general intelligence, I just mean even a weakly capable AI system that can do tasks autonomously for 24 hours, that would already have a huge economic impact. Already, LLMs have changed education. We make our tests in class and on paper so that students are really tested on what they actually know.

Paul Middlebrooks
I was just questioning myself just yesterday, walking in the windy sunshine in Pittsburgh here. I was thinking, "Am I learning faster now that I'm using ChatGPT to find things out, or is it slowing it down, or how is it affecting me?" Anyway, it's different.

Aran Nayebi
I was thinking of that, too. In some ways, am I engaging as critically as I used to? At the same time, does it outweigh the amount by which I can quickly learn a new topic?

Paul Middlebrooks
It seems quite efficient to me. I think overall, for myself, it's been a real benefit.

Aran Nayebi
That's right. I know there are problems with hallucination, et cetera, but it's gotten much better, one. Two, even at the stage it's at, it's hugely impacted. I haven't met a person that doesn't fully use it or use it to some extent, even when they were skeptical initially. It's started to become like Google Search, in a sense.

Paul Middlebrooks
It basically has taken that place.

Aran Nayebi
Has taken that place. I'm just saying, even not AGI at all, AI system that's just useful and capable, which I think is the target of a lot of AI companies today, is going to impact things in ways that maybe we can't always foresee. That's what I'm excited about. It's at least entering an era where that starts to feel like Hi-Fi a little bit, having a little assistant that is actually you can hold a reasonably productive conversation with is cool. That's what I'm excited about. What I'm, I guess, held back by is, I think it continues to be as always in science ideas. In other words, I don't think it's so much compute, actually. I think, as an academic, you try to be more-- In fact, the lack of compute, though it's quite fine actually for our purposes, is, drives you to be more creative. I think that's the fun part. It's also that making sure that, what if we're in the wrong paradigm or something in some way? I think I'm always open to that consideration.

Paul Middlebrooks
What percentage do you put on the likelihood that we're in the wrong paradigm?

Aran Nayebi
[laughs] Probably 10%.

Paul Middlebrooks
Oh, you think we're in the right paradigm?

Aran Nayebi
Yes.

Paul Middlebrooks
This is a weird thing to say there is a correct paradigm because I don't think there is, but we're in one right now.

Aran Nayebi
We're in one that I think is very productive and empirically so, and more so than other prior approaches. I think that speaks volumes. I think there's limitations to it, too. Maybe the ultimate paradigm that people use might be very different. I think that's totally normal within scientific progress. It makes sense to push as hard as you can on the existing thing while it's bearing fruit, and see how far you can take it. Then, when it stops doing that, you have very reasonable next steps of what to take rather than completely jumping ship immediately.

That's, I guess, my style is, push hard on the things, really Steelman it, and then, because it's failed, you are the advocate of the thing. Now you're seeing it's no longer empirically giving you gains, move on to the next thing. At least you know what problems it's falling short on to generate new hypotheses. I also do think that if the ultimate goal is the brain, at least at a very detailed molecular level too that's helpful for disease, not just at this algorithmic level that we're talking about, that's more hardware-agnostic, that's often talked about in NeuroAI. I think that can take a lot longer, and that'll probably happen before we have these very competent algorithms that occur.

Paul Middlebrooks
Wait, what happened before? Sorry. We will have better disease treatment before we have--

Aran Nayebi
No.

Paul Middlebrooks
Oh, okay.

Aran Nayebi
After. I think that's where AI for science can be helpful. I think actually that, by having these better agents that can also accelerate the types of-- Because biology is enormously complex, and the brain as a part of that is itself enormously complex, especially as you go down to beyond the algorithmic level to the synaptic level and the neurotransmitter level. I think that what will ultimately aid that discovery are systems that can really process lots of data and not be tied to particular simple stories, which is what biology has suffered from to some extent, and really help maximize and find new information to generate those hypotheses. In other words, that's why I particularly focus on the algorithmic level, is because I think that there's a lot of room to improve there, and we can get there much more quickly. Then, that in turn, has benefit for studying the brain at a more detailed level for disease, that sort of thing. Biology more broadly, science more broadly, down the line.

Paul Middlebrooks
Jim DiCarlo, we mentioned him earlier in his reverse engineering approach, thinks that when you can predict, when you engineer something, you build it, and you can make predictions. That basically is understanding. What you were just saying about, that we need these agents to handle lots of data, we're simple. We need simple, short, low-complexity sentences, symbolic things to hang onto to say that we understand something. Are we losing understanding there, or do you agree with Jim that's what understanding is?

Aran Nayebi
Yes. I think that ultimately we'll always need our AI systems to communicate simple things to us. It's just that, just like how, in NeuroAI we talk about three things, the task, architecture and learning rule, you and I don't transmit the weights of the neural network to each other in terms of our understanding. We say, "Look, this pattern of these three things explains this brain data or predicts this neural activity of thousands of neurons to hundreds of conditions." There's always going to be a higher-order language by which humans talk to each other and form models of the world.

They'll just be informed by AI systems that don't make those simplifications, but still communicate to us in that context. I don't think there's any way around us as a species using these things. It'll have to be in that kind of language. I would argue that NeuroAI already does that because it summarizes those three things in the particular context of NeuroAI.

I do agree toughly with Jim about the prediction thing. Related to the Turing testing thing, the types of metrics I usually use, especially in settings where we have less knowledge of brain areas, like where we're making progress, it's much better to have a predictive model.

We basically start with zero predictive models of the system to get to one better predictive model already, under linear prediction, is already a huge advance, and leads to control, optogenetic control, that sort of thing that we were talking that Jim is known for. Then, down the line, and this is where the AI agent thing comes in, we want to make more finer grade distinctions for particular questions. We're no longer in the dark ages of that brain area, where we want to push on linear productivity. We then want to ask very specific questions about maybe neurotransmitters, neurochemicals there to aid in a BMI, for example. If you're going to have a brain-machine interface, then you're going to have to care for about the particular physiology of that individual, not just the average. That's where, again, when we come to disease and other things, you're going to want to ask these more finer, grained questions, but you're going to build on the most linear, really predictive model to start with and then iterate it for that particular question.

Paul Middlebrooks
Coming back to the NeuroAI Turing test again because I wanted to ask you this earlier because when we were talking about how you can measure any metric that you want, any representation that you want, as long as you adhere to the theoretical premise of the test, it seems that there's a lot built into what you decide to measure, what you decide is the right metric. There's judgment, I think, that could be had on that. I could dismiss a model that measures only oscillations, for example, or that would be one that a lot of people would agree with me with, and a lot of people would disagree. Oscillations.

I just wanted to throw something out there that some people think is epiphenomenal and doesn't matter. Some people think that it's causal. I think it's both. That doesn't matter. If I decided to measure beta synchrony, and that's my measure, then someone could say, "That's not even worth paying attention to," even though it passes the NeuroAI Turing test. How do I know what metric I should measure?

Aran Nayebi
No, that's an excellent question. There's two things. If your goal is that particular question, then it makes sense to discard the other stuff and focus on that for now, myopically, almost.

Paul Middlebrooks
Okay. I could say then, a lot of people would say, "That's a worthless goal."

Aran Nayebi
Totally. [chuckles] That's what scientists debate all the time about. It's no different than anything we've already been doing. The main thing, though, is if our ultimate goal, though, as a field, is to have a consistent, complete theory of brain function across scales, and we don't have that today, but maybe in the future, then I think we want it to agree on as many metrics that we all agree on as a field, are valuable. Anything that the brain has, if that's your goal, we want it to be passing multiple NeuroAI Turing tests rather than just one across all of those benchmarks.

Paul Middlebrooks
Right. The other related question that I wanted to ask is, given multiple realizability and degeneracy in the way that populations can transform signals to then enact some action, if the representations don't align, is that really a problem? Can't I get to the same location by taking a different route? As long as I'm getting to that location, it doesn't really matter what my internal representations are doing if I'm achieving the task.

Aran Nayebi
Yes, that's right. You could make that argument. One thing that's just interesting is that you end up-- You do the contravariance, basically. You just end up, even if you were like, "Hey, I'm just purely an AI person, I don't really care about matching the brain," it just turns out that your advances lead to better brain models, too. This is the case in vision, even with SSL objectives. With vision, it was like just high variation task was better models of primary visual cortex, with better SSL object.

Paul Middlebrooks
It is astonishing and so fucking cool.

Aran Nayebi
Yes. The converse principle, it's explaining maybe why that might be, but it's interesting that, the same with language and transformers, like the GPT-based models are the best models, predictive models so far of human language areas. When SSL objectives came out that were better, and we were part of that, it was we also got much better models of mouse visual cortex because it gave you a more general purpose thing for the constraints of the smaller cortex and lower visual acuity it needed. In other words, it was like all of these advances in AI, these fundamental advances that we need to get to this ultimate goal of an open-ended autonomous agent, basically, that is what AGI is really.

All of those, to get there, have led to much better theories of the internals than prior theories that came before it. I would say that if there is a science, what does that tell you? That tells you that there's, lurking underneath it, a science of intelligence that unifies neuroscience's goals, cognitive science and AI, where it is about really building a hardware-agnostic theory of intelligence, but it's about optimization under different constraints. That's how these different brains, different brain areas, relate to one another across this type of spectrum.

Paul Middlebrooks
All right, Aran. Thank you for letting me take you on lots of divergent wandering paths here. Is there anything else that we missed that you wanted to discuss or highlight, or that you're excited about or fearful of?

Aran Nayebi
Oh, I can mention maybe for a couple of minutes a little bit of the AI safety stuff that we recently--

Paul Middlebrooks
Sure, yes. All right. We've had Steve Byrnes on the podcast, the big AI safety person.

Aran Nayebi
That's awesome. I mentioned the main goal is building better autonomous lifelong learning agents and using that to try to engage with whole-brain data. The other aspect of it is what happens once we get there. I think that's also another place where being an academic actually makes a lot of sense because we don't really have a good science right now of alignment of making sure that these AI systems are aligned with human values and preferences. You mentioned programming goals and how that leads to unexpected behavior, the reward hacking.

That's a very common thing. We want to try to avoid features of that as we build more and more capable systems because even a weakly capable system will have consequences, both good and bad. We want to try to mitigate those things as much as possible. Now, to be clear, I'm not a doomer or anything like that. I do think genuinely that humans have a higher risk of harming one another than any AI system, especially we have today.

Paul Middlebrooks
Agree.

Aran Nayebi
That doesn't mean that they can’t cause some harm. At the end of the day, it's a technology we're building. There's a nice quote by Dylan Hadfield-Menell, who's at MIT who works on AI safety, saying that, look, when you're a bridge builder, like a civil engineer, you also care about bridge safety. It's like that. I think it should be a natural part of, if you're building these systems, to think about that. I think that's one thing where academics can help because, right now, a lot of the alignment stuff is very high-level frameworks, like policies. They're not really policies, they're just discussions. They're not precise.

We also don't have any guarantees of, okay, say, hypothetically, we do achieve agents that are autonomous and capable. What then? What are the guarantees there? What's hard? I think that's where actually my math background, maybe from earlier, comes in, is you can start to prove theorems about rational, fully capable agents that don't misalign for trivial reasons by failing, but literally, they're ideal, they're computationally unbounded even. If there's things that are hard for them, then it's going to be something we should avoid in practice.

One work that we had recently that's called *Barriers and Pathways to Alignment*, that's under review, is showing that, some of the first complexity theoretic barriers to the alignment problem. Basically, in a nutshell, it's showing that if you have too many distinct tasks or too many agents, there's always going to be problems where the number of bits they have to exchange to provably reach alignment is going to be too large, basically. In other words, what you ideally want to align on, you want to choose your tasks and agents wisely. In other words, it's not a question of if they'll misalign, it's when.

There's always going to be tasks that, even if they were incentivized to align, they will misalign. We have to really be careful about these tasks and agents that we want this alignment for. A corollary of this theoretical result is saying that some people talk about brain-computer interfaces as solving the alignment problem. In other words, like Elon Musk, for example, in Neuralink, that's the only way we'll merge with AI. One, obviously, in practice, the issue with that is that our brains are constrained. I don't think that that's just naturally going to be a band-aid to solve the alignment problem.

The other one because these theoretical results is that even if our brains were unconstrained, we were these perfectly rational, capable agents that were computationally unbounded, if we have too many distinct tasks and agents, like imagine all our BMIs or BCIs are connected by Bluetooth, the number of bits you have to exchange would just be too large anyway that you couldn't guarantee it. Still this, one, BCIs won't solve the alignment problem. Two, choose your tasks and agents wisely. The task of making a sandwich and getting your agent aligned with you when it causes misalignment is far less harmful than running a nuclear power plant.

Paul Middlebrooks
Let's go back to the-- If you have another minute or two.

Aran Nayebi
Yes, I do.

Paul Middlebrooks
I'm going to try to illustrate this through a stupid story. I'm a child, and I build my first bridge by putting a piece of tree trunk over the creek. I try to walk over it, and it breaks. In that case, I didn't plan everything and say, "All right, I'm going to build this bridge, but I have to care about the safety." I just made it, and I'm still around. Then I made a better bridge the next time, and then even a better bridge. Then eventually started thinking about safety. The point of this stupid little story, I apologize, is I think throughout human history, human history is not replete with examples of we're going to plan for the safety. Human history is we're just going to move forward and make it, and then the safety comes later. Why is this an exception?

Aran Nayebi
This is an excellent question. One of the reasons why I think the theoretical study in this case is warranted because it's actually faster to prove a theorem than to run it. We don't have AGI yet, and we probably shouldn't anyway, if we were. The reason for this is that the fundamental difference between AI, this technology that we're building, and prior technology is that now you're starting to build a technology that takes in inputs and intentionally produces actions. It's an agent. As a result, we're not talking about the type--

Paul Middlebrooks
It's not passive.

Aran Nayebi
Yes, it's not passive. It's not like, for example, I was in an airport and the elevator broke, and God, that's an inconvenience. What do people do when they have disability? That causes issues, but I'm not talking about machine failures here. I'm talking about things that are unique to AI. People in AI safety study that right now. They study hallucinations and current LLM systems, and so forth, white box attacks, et cetera. That's great and really useful for today. I guess what I'm talking about here is, suppose we fix the functional problems.

Now, I'm not talking about misalignment with failure modes there, but we want to avoid the situation where you've now built a capable agent that is out there. It might, in the short term, seem like it's agreeing with you and helping you out, but it's either spreading lots of misinformation, either for its own end or otherwise. Then, ultimately, that leads to catastrophe in one way or another. I don't think that's tomorrow or anything, but I think that as we're building more and more capable AI systems, this question starts to become more relevant. I think we can go far beyond sketches of discussions about that are not as precise to really prove guarantees, of like, "Okay, look, if this is actually hard for a very capable system to align with it, then we have to avoid it in practice."

Furthermore, and this is something I'm working on now, is how can we build better incentives? Beyond RLHF, reinforcement learning with human feedback, which is the way we, right now, align these LMs with human values, can we go beyond that and design incentives with theoretical guarantees that prevent this scenario that I'm talking about? Then really implementing current systems today, so it does speak to systems today but has guarantees in the systems of tomorrow.

Paul Middlebrooks
All right, cool. Last thing before I say goodbye. We were trying to figure out what movie to watch with my son the other day, and one of the ones I thought was *The Matrix* would be a good one. Anyway, I happened upon the little clip where they're talking about what *The Matrix* is, and what happened with humanity, and stuff. I'm old enough to remember when this movie came out, and it was super cool, and everyone loves the movie.

Aran Nayebi
Same. I remember going to the theaters to watch it.

Paul Middlebrooks
Oh, really?

Aran Nayebi
Oh, yes.

Paul Middlebrooks
I watched this thing, I thought, "This is so dumb." It looks so cartoonish. Not looks, but the premise is so cartoonish and ridiculous. It made me feel

better that I've learned something, and now it's not like, "Oh, that's matrix. That's awesome." I don't know. How do you feel about *The Matrix* in retrospect?

Aran Nayebi
In retrospect, I think it's obviously very exaggerated. I don't think we're going to get there or anything.

Paul Middlebrooks
Yes. Human bioelectric energy.

Aran Nayebi
Yes, very creative.

Paul Middlebrooks
Creative, yes.

Aran Nayebi
The first *Matrix*, hands down, one of the best movies.

Paul Middlebrooks
Great.

Aran Nayebi
Everything else after was, eh. Especially, did you see the fourth one?

Paul Middlebrooks
No.

Aran Nayebi
Good.

Paul Middlebrooks
Okay. We can leave it at that then.

Aran Nayebi
Yes, good. [laughs]

Paul Middlebrooks
Anyway. All right, Aran, I've got to actually go to work in a few minutes, and I suppose you do, too. Thank you so much for your time, and I'll see you around campus, I hope.

Aran Nayebi
Yes. Thank you so much. This is wonderful and a great pleasure and honor to be here.

[music]

Paul Middlebrooks
"Brain Inspired" is powered by *The Transmitter*, an online publication that aims to deliver useful information, insights, and tools to build bridges across neuroscience and advanced research. Visit thetransmitter.org to explore the latest neuroscience news and perspectives written by journalists and scientists. If you value "Brain Inspired," support it through Patreon to access full-length episodes, join our Discord community, and even influence who I invite to the podcast. Go to braininspired.co to learn more. The music you're hearing is "Little Wing," performed by Kyle Donovan. Thank you for your support. See you next time.

[music]

Subscribe to "Brain Inspired" to receive alerts every time a new podcast episode is released.