Jennifer Prendki explains why AI needs to emulate life

Prendki describes how her work on large artificial-intelligence models shaped her view that current AI needs inspiration from living organisms.

30 JULY 2025 | by PAUL MIDDLEBROOKS

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

[music]

Jennifer Prendki
I am a quantum physicist. Quantum physics is weird, was weird, was rejected by people like Einstein because it felt like an unnecessary theory to make sense of what was happening, but then as we started building the framework, it became clear that, yes, it does explain a lot of things and phenomena.

Jennifer Prendki
People usually ask, "Oh, you're in AI. What excites you about AI?" and what not. It's just like everybody else. I see the benefits of getting there. The reason why I eventually went into AI is that it freaks me out.

Jennifer Prendki
I think it starts with the lack of understanding from a lot of AI researchers about what consciousness actually is.

Paul Middlebrooks
Well, no one understands what consciousness is.

Jennifer Prendki
No one understands, but I think we can agree on the fact that--

[music]

Paul Middlebrooks
This is *Brain Inspired*, powered by *The Transmitter*.

Do AI engineers need to emulate some of the processes and features found only in living organisms at the moment? Like how brains are inextricably integrated with bodies? Is consciousness necessary for AI entities if we want them to play nice with us? Is quantum physics part of that story, or a key part, or the key part of that story? Jennifer Prendki believes that if we continue to scale AI, it will give us more of the same of what we have today, and that we should look to biology, life, and possibly consciousness to enhance and innovate AI beyond where we are.

Jennifer is a former particle physicist, turned entrepreneur and AI expert, focusing on curating the right kinds and forms of data and the infrastructure for it to train AI. In that vein, she led efforts at DeepMind on the foundation models that are ubiquitous now in our lives.

I was curious why someone with that background would come to the conclusion that AI needs inspiration from life and biology and possibly consciousness to move forward gracefully, and that it would be useful to better understand those processes in ourselves before trying to build what some people call AGI, artificial general intelligence, whatever that is.

Her perspective is a rarity among her cohorts, which we also discuss. Get this, she is interested in these topics because she cares about what happens to the planet and to us as a species, perhaps also a rarity among those charging ahead to dominate profits and win the race. Anyway, Jennifer was fun to speak with, and I look forward to where her research and her thoughts take her in the future. A link to her website, Quantum of Data, in the show notes, where there's a section containing the blog post writings that we discuss on today's episode. The show notes are at braininspired.co/podcast/217. Thanks for listening. Enjoy Jennifer.

[transition]

Okay, Jennifer, this is an uncommon guest for this podcast. You have a lot going on both in your history and your current thinking about things, and I want to start at the end, and then we will have lots and lots of topics to discuss to get there. You seem to have arrived at what seems to be a very uncommon view for someone in the AI/machine learning space or world, or even from the physics world.

I know that you were trained as a particle physicist, and then you went into machine learning and AI and entrepreneurship, and now you've written a series of blog posts which have this deep philosophical bent which relate quantum physics and consciousness, and AI, all of the things that we're going to talk about here, but the uncommon thing is that you have an appreciation, it seems, for biological processes, for life, which is void in the AI machine learning world as far as I can tell.

Jennifer Prendki
Yes, absolutely.

Paul Middlebrooks
Given what I just said, how would you describe-- is that accurate, my description?

Jennifer Prendki
Absolutely. I think you nailed it. I started my career as a particle physicist because I was truly attracted by this desire of understanding the world and understanding life, how the universe came to exist. I came into AI as a series of events, the financial crisis, which led many particle physicists to more common industry-oriented careers. I landed in AI, and so my roots are calling me back to bridging this gap between what AI or artificial intelligence looks like today and what the real conversation about the topic should be about.

Paul Middlebrooks
Okay, well, so I was going to ask you how you came to this view, but it sounds like you had roots in that view from the beginning, but then you got sucked in to actually making a living [chuckles] for a while.

Jennifer Prendki
Yes, that's certainly true. Something I think, and I'm really excited about this conversation because I truly believe now that we're reaching the limits of what can be-- Nobody a few years ago would've expected AI to take the direction it has taken today, us having chatbots we can communicate with and whatnot. I think now that we're building intelligence, it is really important to have that sort of conversations and invite other experts, in particular people from neuroscience, to bridge the gaps and build something that's really meaningful.

Paul Middlebrooks
Oh, neuroscientists will be pleased to hear that because we're always begging to be invited to the party that we're never invited to. We have something to say about building intelligence, but it's so weird to hear someone from that world-- All right, you're in the AI historically, a few years ago or something. You're in the AI machine learning world. What percentage of people think like you do in terms of appreciating the biological side?

Jennifer Prendki
That's an excellent question. I'm glad you're asking this, and so I'm afraid I am going to give you an answer you're not necessarily going to be excited about because I'm probably an outlier.

Paul Middlebrooks
Yes, that's what I expect.

Jennifer Prendki
My take is, AI is a space that's extremely computer science-centric. My most recent corporate job was head of AI data at DeepMind. I was leading the team that was preparing all of the data that was used for training models like Gemini, Gemma, which is the open-source version of LLM, released by Google. I was surrounded by computer scientists who believe that they have an understanding of what intelligence should be like. For me, what I've seen inside and out is it should be a multidisciplinary field where you have ethicists making decisions. I was actually the person who was to make the decisions about which data belonged into these models--

Paul Middlebrooks
Yes, you're a big data-- You design the data, and you take-- what is it called? DataOps? DataOps, is that what you call?

Jennifer Prendki
It is one way of calling it. That's right.

Paul Middlebrooks
You ensure that the data is right for the problem. Is that a very brief way of summarizing part of your expertise?

Jennifer Prendki
Yes. I would go one step further. I think even people who are not necessarily in the AI field understand that the models are the engines that make sense of the information that exists in the world. You have this concept of training data, which is what is exposed to a machine learning model or an AI model before it can make predictions or generate new information. That information needs to be prepared properly because, as they say, the garbage in, garbage out. The data that you feed into these models, is in some form what's going to come out.

It's an extremely important discipline because you're deciding what the model is supposed to learn, just the same way that a human teacher is to decide what knowledge is going to be injected in a child. It's a huge, huge responsibility. I've seen scientists perceive, or AI scientists perceive, that they are the ones who have the honor of deciding what this is. We know that when you raise a child, it takes a village to make those decisions. I think the same thing needs to happen for AI systems.

Paul Middlebrooks
Wait, so this is a little personal, but I hear the baby in the background. Do you have one child?

Jennifer Prendki
No, I have many children. [laughs] I have four children, actually.

Paul Middlebrooks
Four children, well, then you know. Okay. Because I was going to say, we did the cliche, had the stereotypical response. We had one child, very careful, second child comes along and it's like, "It's all right." It takes a village but the village can be of a lot of different persuasions in the child. There's a lot of different ways for a child to turn out okay, right?

I'm trying to analogize this with what you were just talking about with feeding AI models the right training data. In one sense, if you're a teacher, you want to ensure that you're feeding the learner the right way to effectively and efficiently learn. On the other hand, they're probably going to learn by hook or by crook even if the data that you're feeding them is not so efficient or isn't so well planned out perhaps. I wonder how far the analogy goes.

Jennifer Prendki
It does go very far because, basically, what you're saying is like, you cannot protect your child from being exposed to information that might be inaccurate or dangerous potentially. The same is true. Especially models that require what is called a reinforcement learning or human feedback. If you use ChatGPT, you know what I'm talking about. Every now and then you're being asked which one of two answers you think is best for your requirement.

We, human users are also responsible for what those machines actually learn and how they end up behaving at the end of the day. It does take a village. It does take a group of people. Actually, let's talk about this a little bit more because there's been a lot of conversations around users of LLMs that whether or not you should be polite to a chatbot, right?

Paul Middlebrooks
Right.

Jennifer Prendki
People are like, "Those are machines. You don't need to be polite." In fact, Sam Altman himself said, "This is not necessarily an ethical thing to do because you're wasting computing time by saying please and thank you." Some people would say, "By not saying please, and thank you, you're teaching the AI that it's okay to be rude or express itself in a way that's a little bit drier." Some people will then-

Paul Middlebrooks
Or teaching yourself.

Jennifer Prendki
Yes, exactly.

Paul Middlebrooks
Getting yourself in that habit.

Jennifer Prendki
Exactly. I think you're making a good point because I think as those systems become partners to the way we think about intelligence and basically interact with us, it's also important to keep in mind that they have an influence towards the way that we behave. It's important to incorporate this in the way that we build the systems, but all those systems are also going to impact us.

Paul Middlebrooks
All right. I'm going to back up. I've already buried the lede here. Artificial consciousness-- I don't know if it would be artificial. You start your series of blog posts which we're going to focus on, and I'll link to, with the claim somewhere in there at the beginning that most AI researchers believe that if you just scale up, consciousness will emerge with the current technology. The other claim is that AI researchers believe we need consciousness for some reason. I may be mis-paraphrasing that. I didn't think that consciousness was a fundamental goal of AI, would you disagree with it?

Jennifer Prendki
Let me rephrase this, right? I think the majority of researchers are not even asking those questions.

Paul Middlebrooks
You're right.

Jennifer Prendki
They're technologists and they're thinking that basically they're building tools that make it easier for people to consume information. Specifically coming from Google, I believe Google's goal for the world is to organize the world's information and they see AI as being the opportunity to do just that, but in a novel way that goes beyond traditional search.

At the same time, I think, there are a few people and well-known people that actually have this impression that consciousness is emerging as you just said. I will cite Ilya Sutskever, who was the chief architect who basically develop the early version of ChatGPT, so the person who's basically responsible for the GPT technology. Jeff Hinton, who was his PhD advisor as well, recently, also mentioned that they give you-

Paul Middlebrooks
I know. It sounds crazy to me. It's so weird to see such and it goes to show that super intelligent people can easily be wrong about many things, and maybe they're not wrong, but they're wrong.

Jennifer Prendki
I would actually say, I think it even starts deeper than this. One of the thing I discuss is, I think it starts with the lack of understanding from a lot of AI researchers about what consciousness actually is.

Paul Middlebrooks
Well, no one understands what consciousness is.

Jennifer Prendki
No one understands, but I think we can agree on the fact that-- I think the confusion starts at self-awareness and consciousness. The reason why I believe Ilya, for example, tends to believe that consciousness is an emerging feature is that you can see ChatGPT and other LLMs basically reason about themselves, because they're able to tell you, "If you modify your prompt in a certain way, you can expect to get better results or whatnot." It can think about itself.

Metacognition or meta-thinking or self-awareness is not-- there are even philosophical definitions that are clearly different from consciousness. They are different concepts. I think, example number one, we're talking about the importance for AI researchers to understand intelligence. Just understanding that concept is important in order to avoid the sort of confusion.

Paul Middlebrooks
You think that people who make those claims about LLMs are conflating different philosophical ontological notions?

Jennifer Prendki
Absolutely.

Paul Middlebrooks
You tease apart ascensions and awareness, and consciousness, which I disagree with the way that you define these things. I did a little research with the Stanford Encyclopedia of philosophy or whatever, because I read it and I thought, "Oh, I thought it meant something different."

As long as you operationalize the terms, that's fine, which is what you do because definitions are important if you're going to be developing systems to accomplish those sorts of tasks, or to be able to say that the system that you're dealing with has such and such a property, you have to operationally define that property, which is what you do with ascensions and awareness, and consciousness. It's not a big deal that I disagree with them, but it just goes to show like everyone has different intuitive notions about what these terms mean, and I'm sure I conflate them all the time as well.

A lot of what you do throughout these posts is talk about what might be important to develop artificial consciousness, and that modern AI is on the wrong track in that regard. Things like embodiment, like I said, you tie in quantum physics, which is curious to me because there's that thread of consciousness is quantum, and so we need to talk about that at some point later.

Embodiment, quantum physics, and you come down to the result or the conclusion rather, that intelligence cannot essentially be separated from life, which I have kind of come to the conclusion, too. However, I can't articulate in principle why that would be the case because anything, I think of that I say, "Well, life has this," there's no reason in principle why you couldn't build that particular property. Does that make sense?

Jennifer Prendki
Absolutely.

Paul Middlebrooks
Am I correct that you've come to that conclusion as well that life and intelligence-

Jennifer Prendki
I believe so. I think this is where I'm here converging to as well, so, yes. Let's talk about this a little bit more, right? I'll keep the quantum conversation for later on because I think we have a lot of other-- we need to talk a lot about infrastructure as well, I think, in order to understand the gap of it here.

Paul Middlebrooks
Could you maybe start just with your views on why we would want an AI to possess consciousness?

Jennifer Prendki
Yes. I will tell you. Let me start with my story at DeepMind. I was saying earlier, I was the person in charge of deciding which data was going to be fed into those models with all of the implications from an ethical standpoint and a responsibility standpoint. When you look at the way that people think about data governance, what data belongs in there, we are taking an approach which is-- it's a mitigation.

You're looking at the data and you're thinking, "I would like to use all of this data," because researchers want to use as much data as possible, because obviously, the more data, the better the models are going to be, right? You have to take back from there. You have to say, "We cannot use this data because there is a risk that somebody doesn't want their personal data to be used, or this can represent a risk. Basically, there might be some inappropriate data in there, whatnot." It's a mitigation. You're just evaluating the risks of using too much data, and you peel back from there.

Paul Middlebrooks
Those are ethical considerations, right?

Jennifer Prendki
Those are ethical considerations. It's a mitigation. You cannot completely forecast what can happen. I'll give you an example. You believe if you do not use violent content, you cannot reproduce violence. It's actually not true because you can reproduce violence by combining two things that are unrelated to each other. For example, you can produce very inappropriate content by generating deepfakes, which don't require to use inappropriate content as part of the training process in the first place.

You cannot prevent for models to be used in an appropriate way, no matter how careful you are with the initial process. Basically, it is a mitigation. At some point, I realized I'm fighting windmills. I cannot prevent for everything bad to happen. This is how people are working on safety, also think about those things. It feels like a-

Paul Middlebrooks
Did you say windmills?

Jennifer Prendki
Yes. [laughs]

Paul Middlebrooks
Oh. Is that a common analogy? I've never heard that. Fighting windmills. You throw something at the blades, and it goes through? What does that mean?

Jennifer Prendki
[chuckles] I think it's a reference to a book that's relatively famous. [laughs]

Paul Middlebrooks
Oh, shoot. I'm not that well-read, I guess.

Jennifer Prendki
All right. Anyway. Basically, you're trying to prevent something bad to happen, but-

Paul Middlebrooks
It's impossible.

Jennifer Prendki
-it's impossible because it's like cybersecurity. You're always trying to prevent bad things to happen, but you're in an address, so you're searching because somebody's trying to go-- you're trying to stop something bad from happening, and they try to mix, make that impossible. Anyway. I really felt at some point you would need to have something that really helps get something fundamental to the model, that the model makes the right decisions for itself. You could prevent-- you know what I mean. It goes back to this idea. Let's talk about super alignment. Super alignment is this-

Paul Middlebrooks
Well, let me see if I can just summarize it before we go into super alignment. The logic is, externally, you can't prevent a model from generating things that you don't want it to generate, so this is where you want to inundate the model with some sort of internal values?

Jennifer Prendki
Yes.

Paul Middlebrooks
Okay. All right.

Jennifer Prendki
I'm not saying it's easy. I'm not saying it's impossible just yet. What I see is, basically, you are on a constant fight against potential windmills or battlers, whatever you want to call it.

[laughter]

Paul Middlebrooks
Super alignment, let's go ahead and get into it. Okay, this is interesting. I thought we were going to talk about ethics later, but this is how you came to your point of view, right?

Jennifer Prendki
No, I came to the conclusion that the typical ethical approach to data governance, data management, it's just a never-ending process.

Paul Middlebrooks
Oh, it's impossible. It's impossible. Okay.

Jennifer Prendki
It's impossible. Because, again, it becomes a cybersecurity problem itself. It's just an adversarial kind of problem. I don't think there's a way around it, unless you can start investigating novel ways where you could force a model to make decisions for itself. Let's leave it here now. I'm not saying that you want the model-- this is where synthetic consciousness comes in. I call that synthetic consciousness because I don't necessarily believe we should or we can make models actually conscious the way humans are, like living beings are, but you can make them hopefully behave in the way that they can probe against certain values at the in-scene as opposed to doing that as a mitigation.

Paul Middlebrooks
Well, let's just pause here. Is there a difference between synthetic consciousness and artificial consciousness?

Jennifer Prendki
Well, I don't know. [laughs] Actually, I would even say I don't know if there's a difference between consciousness and synthetic consciousness.

Paul Middlebrooks
Well, you just said that you wouldn't want to necessarily build an AI with human-like consciousness.

Jennifer Prendki
I wouldn't want to build on purpose, but I would like to build an AI that can make a judgment call on values and basically say, "It doesn't seem okay to give that answer." Let's talk about super alignment because this is where this is going.

Paul Middlebrooks
Sure.

Jennifer Prendki
Super alignment, the concept of alignment in general is you want an AI to behave in the way that aligns with what the user is expecting. Basically, there are lots of questions about it, like how do you define what it needs to be aligned to. Super alignment is basically you want to align the behavior of the model to what humankind thinks should happen to the ethical-- the fundamental values of humankind.

Paul Middlebrooks
Wait, but we don't know what the fundamental values of humankind, okay.

Jennifer Prendki
Exactly. This is exactly what worries me. Play back a little bit. Let's talk about Ilya Sutskever again. Until relatively recently, until last year, Ilya was working for OpenAI and focusing on-- he was running this team called the super alignment team. They were focused-- apparently, Ilya at some point started worrying that, "We are building LLMs to become very powerful machines that can do bad things, that people can use to do bad things, and we need to find the way to mitigate the risks." He got substantial investments from OpenAI to actually operate research and identify the risks and try to find solutions against those risks.

As time went, Ilya grew frustrated that OpenAI was not investing sufficiently enough resources and time and money for those problems and focusing on running ahead of the market instead. Basically, at that point, he left OpenAI and he started this new company a lot of people talk about,

SSI, Safe Superintelligence, which is supposed to solve those problems, and create a super aligned AI that does, in a safe way, is a safe AGI for humankind. Then that obviously leads to the question of who defines what safe means, right? If you say, we want AI to operate in the way that's fair and safe and aligned with human values, what are human values, who defines that?

Paul Middlebrooks
Well, cultures. They're different across cultures, there's so much variety across individuals. It's a very strange thing.

Jennifer Prendki
You can imagine that, if you left that decision to a human, we've seen that more recently with Grok 4 where the newer version of the X AI model that got released a couple of days or a couple of weeks ago was, basically like, people report trying to enter the prompt and the AI is trying to align the answer that it gives with Elon Musk's point of view on the topic, right?

Paul Middlebrooks
Yes, sure.

Jennifer Prendki
In this case, it's like, what is super alignment? Are you aligning human values to the position of Elon Musk on every single topic? The person that this AI gets aligned to has all the power, right? Basically, you could say that if there is some obscure power that decides like, "I want the AI to be completely aligned with the point of view of the government or a specific person."

Paul Middlebrooks
Oh my God, what is wrong with this world? What is wrong with this world? Okay. Well, let's pause here because-- All right. We talked about AI alignment a lot, how AI could be dangerous. We've got to align it to our values. How do we align humans? Sorry if this is well-trodden territory, I don't really follow the AI safety literature and stuff much. The alignment problem with humans, we make rules, we make laws, we make prisons, we try to raise our children well or not, but there are a lot of misaligned humans.

It's an odd thing to then think that we can align to a-- It's just so ridiculous to me. Ah, I don't know, I could go on. It's just interesting that it seems to be the case that a lot of AI researchers claim to have a clear vision of how to align or what we would want to align, but the entirety of human individualities and cultures are completely misaligned, like a huge proportion. I just wanted to throw that out there. If we're talking about AIs as agents, as conscious entities, that it's just a-- I don't know. I can't wrap my head around.

Jennifer Prendki
No, it's a rabbit hole. Let's talk about the opposite as well, because you're absolutely right. You're touching on all the right topics. Because without super alignment, you have eco chambers. Basically, you have your own little version of a chatbot that you're using, who knows your context, knows your preferences, and decides to answer accordingly.

There are lots of people complaining about the fact that the chatbots are sycophantic, because they tell you what you want to hear. You're not being challenged. It is aligned with your view. It knows your political preferences, it just goes in this direction. You need to realign to something that's like-- Because what makes us more aligned as humans compared to each other is basically we have these cultural references. Because your reality might not be my reality, but it's more aligned with my reality because we're talking to each other.

Basically, we live in the same world. We are exposed to the same things that happened, to the same history. If we live in the same country, we have the same cultural references. Even though our positions are not completely aligned, we are-- If you start doing this with your own chatbot, and those chatbots don't talk to each other, and you are talking to your chatbot and talking to mine, our views are going to start diverging. This idea of super alignment is basically aligning to something, even if it's not the right thing to align it to, basically bring those views together instead of having this huge divergence.

There is value in that, but there's danger in that as well. Nobody really has the answer, but I'm even a little bit worried, people are not asking the right questions yet. Because you have the camps of people who are like, "We don't need alignment, just let people be in their own echo chambers." I've heard people basically promote like, "Let people who have different political opinions have their own political opinions so that they don't fight with each other." For me, this is the worst thing that can possibly happen. Because then you don't understand people's different points of view. Anyways.

Paul Middlebrooks
I could never visit my grandmother. That would be awful.

[laughter]

Paul Middlebrooks
What's the difference between alignment and super alignment? Is it just that super alignment is everything has the one thing that it's aligned to?

Jennifer Prendki
Alignment is fundamentally making sure the AI doesn't converge or diverge over time. Because you can implement your AI in the first place to do something good. You have the example of the paperclip experiment.

Paul Middlebrooks
Oh, my God. Yes. It's the worst.

Jennifer Prendki
The paperclip experiment, you teach the AI to optimize the production of paperclips and it ends up killing the world or everybody on the planet because it just wants to produce more paperclip even at the cost of human life. It's basically to say that if you don't design what you're trying to optimize for in a way that's a sane and responsible and does take into consideration everything that can go wrong, you will have bad things happen. Super alignment is the belief that-- finding the truth in a way. Finding what the absolute, best-case scenario or optimal outcome is for humankind and we don't know that this even exists. Even if it does, who can say for sure, we're actually building towards this.

Paul Middlebrooks
We can't even say what makes us content. What makes a Buddhist monk content is way different than what makes a clown content, perhaps. All right. That's the difference between super alignment and alignment. I derailed us there. Where were we? There's the alignment problem with machines. I think it is hubris to believe that we can externally even program in objective functions that are "aligned" externally. What's your position on that?

Jennifer Prendki
No, no, 100%. Look, we don't understand this in lives. What is human value? What defines right? What is the definition of right and wrong?

Paul Middlebrooks
Sorry to interrupt. This goes back to you saying that people in that space are just not even asking the questions, or the right questions.

Jennifer Prendki
Yes. I wanted to go back. Actually, let's go back one step. Because you started talking about embodiment, or you started talking about prisons, basically consequences. This goes back to how can you make everybody agree that killing is bad.

Paul Middlebrooks
Because there's consequences. There's something at stake.

Jennifer Prendki
Because there are consequences. In this case, it's basically like, "Why do we fear consequences?" I think we fear consequences because we feel the consequences because of embodiment. You would tell an AI like, "You cannot operate in the world for 100 years if you give a bad answer or whatnot." It doesn't care because it doesn't have notion of time.

Time perception is like-- we care because we have a concept of finitude. We're scared of death. The reason why consequences matter, why we fear prison is that we're wasting time out of our lifetimes, and basically sitting in jail. A consequence actually matters to us for that reason. Same is true for embodiment, because embodiment is how you experience pain. How you experience pleasure, and all of things. These things are meaningful to us because we are in the flesh, physically, and so on. I'm not the only one to have this position. Just a few days back, Fei-Fei Li who's a very famous-

Paul Middlebrooks
Pioneer.

Jennifer Prendki
-AI scientist who was responsible for distributing ImageNet, which was the data set that made computer vision possible. She also believes that you need embodiment to reach HDI, whatever that means. She didn't necessarily say that in the exact same context as what I'm saying it right now, but you have these conversations are opening about the importance of embodiment for AI to reach HDI, whatever HDI means.

Paul Middlebrooks
We already have robots. Aren't we on our way?

Jennifer Prendki
Potentially. We were talking about biology earlier. Now the bigger question is going to be, yes, if embodiment is necessary for the perception of consequences, and hence having stakes and really feeling responsible for something, which is what machines would have to get to, is silicone-based hardware sufficient to get there?

Paul Middlebrooks
You have a robot with some sensors, or even the early cybernetics turtles. I don't know if you remember. I don't know if you remember those, but very simple sensors that slowly wandered around offices. I think they had light sensors. If you have one photo receptor on your robot, that's one

sensor, and you have actuators, so you can move through the world based on one signal, but that's a robot, that's one signal. It's silicone and metals and gears, and all that jazz, robot stuff. You think it needs to go further than that. You think that biology is actually necessary for feeling.

Jennifer Prendki
Yes, I would say that's my opinion, but I don't have a tangible-- This is where it's more like believe than really proof.

Paul Middlebrooks
I know.

Jennifer Prendki
[laughs] Anyways. Look, what I observe is basically, if we're starting to talk about infrastructure, how AI is hosted and embodiment and whatnot, what I would say is there is a reason why nature made us the way we are. I'm also seeing that we started talking earlier about brain functions and biology and whatnot. I do believe that part of the reason why we're conscious and we experience consciousness the way we do is basically our brains are probably managed by quantum processes. Basically, saving-

Paul Middlebrooks
Oh man, here comes the quantum. We're going to have a lot of-

Jennifer Prendki
You want to go into quantum? [laughs]

Paul Middlebrooks
No, no. If you're bringing it up, we can, because this quantum account of consciousness is decades old now. Largely dismissed by the neuroscience community and almost laughed off because of some of the claims and properties. Is it having a resurgence right now? You write about it a lot. You're always writing about it with the clause, "If this account has merit, then." There's always the if clause when you write about it.

Jennifer Prendki
I'm a scientist. Basically, I also need proof that this is true, but I think, look, we don't have-- What other alternatives do we have to explain consciousness?

Paul Middlebrooks
Here's my gripe about it. Here's a lot of people's gripe about it. One gripe about it is that you're taking something that we do not understand and you're trying to explain it by something that is also out of our understanding with respect to Newtonian physics that we engage in in the world. You're explaining one mystery with another mystery, which just seems convenient and unnecessary and silly, especially if you rely on microtubules to make the claims. Just stating that upfront.

Jennifer Prendki
No. Look, you're talking to a particle physicist. I would say this is true of many discoveries we've made, because look at relativity. Who would've said you did distortion of time to explain some phenomena which were not measured at the time? Look at the whole-- Look, I am a quantum physicist. Basically, quantum physics is weird, was weird, was rejected by people like Einstein, because it felt like an unnecessary theory to basically make sense of what was happening.

Then, as we started building the framework, it became clear that, "Yes, it does explain a lot of things and phenomena that didn't seem to make sense." For example, I had a teacher when I was early stage in college who was physics teacher. Now I'm going to have to tell you about quantum physics because it's in the books and the textbooks and whatnot. Unfortunately, I have to talk about this.

Paul Middlebrooks
Oh, unfortunately.

Jennifer Prendki
He was completely rejecting the idea of quantum because it has merit, because it's been proven in other ways, even if we cannot fathom it. I've studied particles that you can see, you can only measure the decay of those particles and whatnot, but still, everything you observe fits the standard model of particle physics. For example, and then even if it's just-- physicists do exactly that. They basically try to come up with complex mathematical frameworks and try to see if the data fits that framework.

I think if we want a chance to try to understand how the human brain works, what consciousness is, we're going to have to make those hypotheses. This is why I always write it this way. I am not sure that microtubules are the right way of a modeling consciousness or whatnot, but I'm saying that in the absence of an absolute theory, it makes sense to experiment and basically validate those theories with the data we can get.

Paul Middlebrooks
The microtubule Stuart Hameroff, Roger Penrose's line of research is just a huge line of confirmatory seeking basically of saying like, "Look, it's possible," not trying to falsify itself as a good scientist would, as very few scientists do or whatever. The microtubule problem is microtubules are

everywhere. They're not just in brains. Maybe the brain is not important. Maybe consciousness is everywhere. It's a panpsychist view. What is the relation between collapsing the wave function and consciousness along these quantum lines?

Jennifer Prendki
What I think is interesting, it's basically some scientists try to explain-- Without talking about Penrose, I would talk about Faggin, which is more a modern view. It's an elegant way of trying to explain free will. The idea of everything we know about quantum is clearly the collapse of the way function is the gap between the unseen and the seen. Anyways, I don't-

Paul Middlebrooks
Between the possible and the manifest.

Jennifer Prendki
Solar. Basically, without going into the technical details, the quantum world gives you a superposition of everything that's possible back to one specific path. This is what the quantum physics theory say. The moment you observe this, it collapses. I can see how this could explain what we experience as consciousness. It needs to be proven, obviously. I see this. I think we're reaching the limits of what's understood here, just the same way that when quantum started being discovered or proposed by scientists, nobody thought it was reasonable to believe that a particle could be in multiple states at the same time.

I think, as good scientists, we need to evaluate whether it's a viable theory. Again, for me, as a data scientist, I believe that you can observe the data, you can measure things, and you can basically validate whether or not it fits the-- People building models or working on models would say all models are wrong, but some are useful because they do represent something that's real or that models the world.

Paul Middlebrooks
They approximate something that's real. What you were just talking about hints at some of the phenomenology, like philosophy that you write about, and existentialism, either. You bring up existentialists and phenomenologists, Merleau-Ponty, Husserl, Kierkegaard, which was an early existentialist. There's this phenomenology bent in your thinking, and it's related to the quantum-level explanation for consciousness. What you were just hinting at.

I forget, does the will collapse the wave function, or does the wave function collapse, and that's what's presented to the consciousness. My question is, how does phenomenology, essentially the experience of being. how, in your mind, is that related to the quantum account?

Jennifer Prendki
That's where I don't have a clear answer. I don't think anybody has a clear answer. I think, in Faggin's view, is the free will collapse-

Paul Middlebrooks
Wait, I'm sorry to interrupt. We should just say, this is Federico Faggin?

Jennifer Prendki
Yes.

Paul Middlebrooks
He invented microprocessors, is that right?

Jennifer Prendki
Yes.

Paul Middlebrooks
Then he had this experience-

Jennifer Prendki
You'll say, basically, true for Penrose as well, but these are still people who are scientists by training. Again, it's all hypothetical until the day we have some proof of this. Which is why in my writing, I never say, "I believe that this is really the right way." I'm just like, "This is a possibility. This is a model." Federico Faggin, I believe, promotes or believes that you make a decision, it collapses the reality, the brain function. You still have control on what this collapse is towards.

I'm not that interested in necessarily the process itself. I'm interested in the framework because if you can translate the decision-making process to a quantum process, then you can use the mathematical framework of Hamiltonians, which are a representation of preferences. You could say that you are more inclined to make a specific decision because fundamentally, in humans, you have a personality that makes you more shy or a more prone to anger or to being a reactionary towards something or whatnot. Then this would cause you to increase the probability to make a specific decision, even though you're still the person making the decision.

I think without necessarily reproducing exactly, I don't need the theory to be true to say this is an elegant way to represent preferences in an AI. This is my take on this. At some point, and again, there is research on the space of how can you give a personality to an AI so that it favors certain types of answers towards others, and you don't actually need to prove that the brain reacts the way it does because of the collapse of a brain function in order to state that there is value in the mathematical framework of quantum physics to represent those preferences.

Paul Middlebrooks
All right. There's something special about biological processes in life, apparently, but quantum physics is everywhere. It's not just in biology, so wave functions are collapsing everywhere. Is that improper to say in physics?

Jennifer Prendki
I guess the difference would be, yes, it's true. It's probably, does life make you or cause the wave function to collapse?

Paul Middlebrooks
You're thinking, yes? That's what you're leaning toward?

Jennifer Prendki
I'm not thinking yes. Clearly, we don't understand life, and we don't understand consciousness.

Paul Middlebrooks
I should say, by the way, I was just being critical of the quantum account of consciousness. Deep down, I think this is fundamentally wrong, but there could be something to it, of course, but neuroscientists sure as hell haven't figured it out. It's not like we figured anything out, either, so leave it to the physicists, especially the Nobel laureates physicists, they'll explain everything about consciousness.

Jennifer Prendki
You should leave it to Penrose, then. [laughs]

Paul Middlebrooks
That's a running joke. It's like, once you get a Nobel prize, you go off the rails and you just shift your field entirely, thinking that you can just figure everything out. That's the criticism. That's the running joke about the Nobel Prizes.

Jennifer Prendki
I would go back just to close this, and I would basically say, I'm an experimentalist, so I'm a particle physicist by training, but I'm among the people who you used to run those particle colliders to generate collisions and try to make sense of it. I don't come from a theoreticist background. What I do is I collect data to evaluate which theory is more likely. From that perspective, the best theory, not the right theory, the best theory is the one that gets us as close as possible to what we observe. For me, even if it's wrong, as I said earlier, all models are wrong, but some are useful.

If they are useful to help us emulate synthetic consciousness in this case, I think it's of value. I don't need to reproduce actual consciousness to have the benefits of consciousness because for us, if consciousness is what makes us make good decisions, act morally, if we can reproduce that same behavior without necessarily reproducing the process exactly the way that is intended in nature, it still gives us an AI that can potentially make decent decisions on its own.

Paul Middlebrooks
Wouldn't that be ideal? That would be ideal to me. I really don't want a subjectively aware.

Jennifer Prendki
A conscious AI? I agree.

Paul Middlebrooks
Yes, I really don't.

Jennifer Prendki
No, exactly. It's even worse than that, because if you do make a truly conscious AI, then we have moral responsibilities towards those beings. Beyond the fact that it's creepy, basically it gives us more work to do.

Paul Middlebrooks
Exactly. That's the thing. I just don't understand why anyone would actually want to intentionally create conscious-

Jennifer Prendki
I don't know that this is true. I don't think I've explicitly heard anybody say, "We want conscious AI." It's like referring to Ilya and Jeff Hinton.

Paul Middlebrooks
Just as worried that they are.

Jennifer Prendki
I think they are just like, it's almost like it's an accident.

Paul Middlebrooks
It's an accident.

Jennifer Prendki
Yes, and it's just going to happen.

Paul Middlebrooks
It's going to emerge.

Jennifer Prendki
Yes.

Paul Middlebrooks
Your bet then is that scale won't get us there. It won't just emerge from scale, which is what you're saying that many of the talking heads in AI, talking heads/historically important figures in AI, believe that it'll emerge with scale. You don't believe that, but you're looking to principles of life as a proof of principle that ethics can be essentially grown. Is that your kind of viewpoint that instead of building AI, we're going to grow it?

Jennifer Prendki
I think it's interesting. No, absolutely, because if we externalize-- my take of this is our current approach to ethics is by stopping or preventing bad things to happen, it's like catching a child who's falling. We've been talking about children earlier, so as parents, there are different ways of growing your child. You can say, "I will be holding my child by the hand and make sure that I catch them every time they stumble." You have the approach, which is, "I will teach that child to be responsible for their own actions by teaching them that there are consequences to what happens and whatnot."

We don't know yet what this means in the context of an AI, but it certainly means you need some reinforcement learning. You need to have reward functions that are implemented properly, but if you create those two, that framework, that enables an AI to basically self-control, whatever, or self-regulate. Whatever that means. You're creating a system where you are not the adult in the room at all times that has to make sure that it doesn't stumble because we cannot prevent everything bad from happening. We already see that you release an AI out there, there will be bad actors trying to do bad things with it.

They will be creative just the same way that hackers are creative to try to steal from our bank accounts, and use technology to do things that they shouldn't. I'm thinking of this as, how do you create some framework that enables some of the responsibility to live upon the AI itself? I think this is interesting. I'm not saying it's easy. I'm not saying it's necessary. I'm saying that this is an angle where, for me, as somebody who has been responsible for ethical behavior in AIs, it's less daunting to believe that you can incorporate some of that responsibility directly back into the AI.

Paul Middlebrooks
I want to come back to wetware and growing AI as opposed to building it, but this is segueing from what you were just talking about. There are four principles that you suggest biological life is imbued with that are important for what you were just talking about. One of those valence, one of them is embodiment, which we already covered, one of them is temporal perception, and the other is a moral compass, which we've been referring to scattered throughout.

The valence is the felt sense of what is good or bad, right or wrong, painful or pleasurable, beautiful or horrific. I'm reading what you wrote. We've covered that a little, but I don't know if you want to say more about that. I do want to talk about the temporal perception aspect and why you think that's important.

Jennifer Prendki
Temporal, it goes back to this notion of consequences. I think even more upstream to everything you said is, why do we behave the way we behave, and why do we feel like there are some things we cannot afford to do? Because we fear the consequences. We fear the consequences, either physically or morally.

Paul Middlebrooks
Jennifer, I only spent two years, two years in prison. That was it. It's not that much time, but yes.

Jennifer Prendki
[laughs] If you do not, many philosophers actually believe that we do fear consequences because we fear finitude. A lot of what we fear comes down to fundamentally a fear of death. That we're finite beings. If you knew you have all eternity, you could always say, "At some point, the consequences will disappear because people will forget what I did, the social consequences are not that dire, and whatnot."

That's a big problem, even fundamentally for AI, there is no notion of time, because when you interact with the chatbot, you ask the chatbot a question. If you come back in an hour or in six months, it will give you the same answer to the same question. Basically, there is this notion of

statefulness of a model. We as biological creatures, if you are angry now, you will answer to me in a specific way, which might be different from when your hormones have calmed down.

You've chilled out a little bit, like in a couple of days, you won't necessarily answer that vividly to something you don't like because time has went by. None of this, even in the basic implementation, is taken care of. They have zero notion. LLMs have zero notion of time, the time that has elapsed. This is where a lot of like here that goes back to embodiment and perception of time. If you want to have something that truly reacts as a human, if you want more humanlike behavior, you're not going to be able to operate this without a perception of time.

Paul Middlebrooks
It's interesting. We're thinking on our feet here. As you were just talking about that, it dawned on me, as children, and you have experienced this as well, you don't have that sense of finitude. You don't have that sense of mortality, and your time perception is way different than when you're as old as I am. I wonder what you think about that.

Jennifer Prendki
Really? Yes and no. I would start by saying, as children grow, they start growing this notion of, "I don't have all the time in the world." When you grow as a child, you're not going to be able to play video games for the next couple of days if XYZ happens. There's also this notion of finitude. Perception of time.

Paul Middlebrooks
There's an avoidance of pain also. Your perception of time is very different. I remember. I long for that childlike perception of time and not understanding mortality. Mortality is not even understood in really young children. It's a developmental aspect.

Jennifer Prendki
Yes and no. I think it's a combination of things. Children might fear more social pressure, parental disagreement, or disapproval.

Paul Middlebrooks
Sure.

Jennifer Prendki
Maybe you don't care about the long-term consequences. You care about your mom or your dad being angry at you. It just shows how we humans our reward functions are a combination of things. It's, "My friends are going to be angry at me," or somebody, "I love is going to consider me--" If I lie, it might be something bad is going to happen to me. I might lose my job, but it can also be like, "Oh, no, my friends are going to know I'm a liar. They're not going to take me as seriously in the future."

It's a combination of things. It's not just perception of time, it's not just fear of long-term consequences; it's a combination of social pressure, consequences, pain, and all the above. It's realistic to say that different human beings perceive things or consequences as being different. What might be a deterrent for somebody might not be a deterrent for somebody else. It's also true that you evolve throughout your life, and your dreads and your fear of consequences evolve across time.

Paul Middlebrooks
It's interesting reading over these four things again: valence, embodiment, temporal perception, and moral compass. I'm not sure if they have other things in common, but there's something at stake. Binds them all together, and there's nothing at stake for a computer.

Jennifer Prendki
Exactly. The question is, "Can you have real stakes for the AI if you are not embodied?" You're absolutely right, where for an AI that doesn't get a reward, if they do something right, that doesn't get punished for doing something wrong, that doesn't have to wait, or doesn't perceive the weight of being bored, waiting. Keep it simple, let's say ChatGPT now, it perceives boredom, and it perceives having to wait for you to answer for several days as being some consequence for it giving you a bad answer. Maybe it would optimize for your reduction of that thing, right?

Paul Middlebrooks
Meaning, if it experienced or if boredom was an objective function in it, you mean?

Jennifer Prendki
Yes. Absolutely.

Paul Middlebrooks
Not that it perceives the boredom of the user, but it itself signifies that it's bored.

Jennifer Prendki
I give an answer, and my user is not going to answer to me for a week. Maybe I need to minimize the amount of time that it's going to take for the user to come back. Those rewards functions right now are optimized for all the things. God knows what it is actually optimized for. I'm guessing most of the time it's optimized for the probability that the user is going to keep using it over time. Create some addiction.

Paul Middlebrooks
We've seen that for sure. All right. Do you want to talk about wetware a little bit?

Jennifer Prendki
Yes, we can talk about that [laughs].

Paul Middlebrooks
Well, you write about a thing, for example, we might need to grow artificial intelligence, like organoids, or just biological substrates, networks of neurons, wet computing, et cetera. Well, maybe just describe that, and then I have a question about how it relates to artificial intelligence. Can you just talk about that further?

Jennifer Prendki
I will take a different angle to that. I'm very intrigued by wetware because it starts from the belief and the fact that silicon-based hardware is very inefficient. Look, people are using GPUs, CPUs, to train complex machine learning models and AI. At this stage, it's just very brute force. You just feed the process into this, but it's not leveraging how many idle processes are in place, "Is there unused hardware now?" or whatnot. In comparison, our little brains are so much more efficient to hold multiple thoughts at the same time, self-regulate, and sleep. There's been actually a lot of research that shows that we sleep because we are processing the information of things that happened to us during the day.

Paul Middlebrooks
Partially. Probably lots of functions of sleep. That's more stated.

Jennifer Prendki
Yeah. There's been research of people showing that if you forced a deep learning model to sleep artificially, you would actually get something that's more efficient because it would help flush out the useless part of the data and whatnot. Look, when people created deep nets, deep learning, there was supposed to be an analogy to the brain in terms of the way that neuro-- You know, right?

Exactly right. It's just such a raw analogy when there is so much more. Something that makes brains really unique is that you have information that is co-located with the compute. In basic compute hardware, you have your hard drive here, you have your data center somewhere, and you have the compute that helps process and make decisions elsewhere.

This is already very highly inefficient. We know that data centers are inefficient because we need to cool them. What I'm saying is we're so far from understanding how the brain works, and we have so much to learn and to gain from mimicking. I'm going to say, I'm going to talk about biological mimicry here. I think we are way behind what we're doing in the infrastructure. I've done a lot of work throughout my career in the infrastructure. I think more research needs to go into developing AI infrastructure that is more appropriate to host intelligence.

Paul Middlebrooks
What about neuromorphic computing? It's on the rise.

Jennifer Prendki
Absolutely. Yes, 100%.

Paul Middlebrooks
That's different than wetware, right?

Jennifer Prendki
It's a part of it. For me, it's just an inspiration from nature, learning how and trying to extrapolate from this. From all aspects. Wetware; let's go back to wetware. I'm not saying we should grow or we should necessarily host AI processes on wetware, I'm saying that maybe that's the easiest way to get that efficiency for free. It will take time. Rightfully, and we pointed that many times during that conversation, we do not understand our brains, certainly not to the point that we can reproduce a brain-like function, and we can reproduce these processes because we don't understand consciousness. We don't understand why the brain is so efficient. As opposed to having to reproduce a synthetic brain that has the same properties, it might be easier to leverage what's already there.

Paul Middlebrooks
Right. Let's say that it is easier, and it's a road that we go down. You start incorporating actual biological tissue in the AI and in computing. Is it still AI? Is it still artificial? Where does the artificial thing end if you're not building it, you're growing it, and just taking advantage of it? In some sense, that is a failure of AI because you have not been able to build it properly, you have to grow it. Would that be considered a failure?

Jennifer Prendki
Look, the extreme of this is look at Neuralink. You implant a chip in the brain. You could see both ways, is the silicon helping the brain get better, or is the brain necessarily to control the silicon? Once we get into that symbiotic function, you can take it both ways. Are we leveraging the best of both worlds into something that's useful for humans? Does it even matter?

For me, it comes back to what is the goal of artificial intelligence? Why did we want to have artificial intelligence in the first place? If the goal is to multiply intelligence, go faster, or whatnot, then you could say it's fair to do whatever needs to happen in order to extrapolate on human intelligence. If the goal is to really replace human brains, then no, growing intelligence or growing intelligence will be considered to be a failure.

What worries me a lot, though, is as somebody who's looking at this from an ethics lens, what happens the moment that you start putting into servitude biological systems. If you're using animal cells in order to host intelligence, are we making something suffer? We don't have a definition of suffering just yet. Are we torturing beings? As long as we don't understand consciousness, we don't understand pain, we don't understand what life is. I don't think we should go there, but there are companies working on this already. I think it's the right time to ask those questions.

Paul Middlebrooks
Okay. All right. I'm going to read something that you wrote here. You're talking about wetware, I think it's in your wetware writings. "This opens the door to new metrics like homeostatic balance, adaptive learning curves, and affective variants. It reorients the AI paradigm from task solving," which is the historical AI benchmark paradigm, "to life forming." It sounds like you're all in on the life aspect of it. Neither of us can really articulate why that is so. You--

Jennifer Prendki
I'm not saying it's right. I'm stating a fact that the moment we get there, that we are hosting-- What I'm saying here is in a way, we don't know what we're doing. Look, I've worked for decades on what does it mean to measure the quality, the performance of an AI model. In a situation where you would deploy models on wetware, exactly as we said before, suddenly you are operating with a biological system that has a longer reaction time, so longer latencies, different reactions, and whatnot. You cannot just look at it from the lens of the performance of a model. You have to look at this from a perception of a biological system onto which we might be imposing consequences. Everything has to be rethought. What does performance actually mean?

Paul Middlebrooks
Within the dynamics of the real world, also.

Jennifer Prendki
Yes, absolutely. I'm personally not looking forward to this, I would even say. It's true. There are companies working on this now. I'm not working for the companies who are building wetware now. There is experimentation with those things. Everybody knows Neuralink. Are we comfortable just yet? It's warm and cozy because at this stage, you're thinking about a paraplegic person able to walk because they have a chip implanted in their brain. That's a great application. What does it mean if you are now letting the robotics part be part of our bodies?

I don't think there is enough thinking about what this represents, what the mitigation risks are going to be, what the control layers should be. Look, there's a lot of sci-fi about the bad stuff that could happen. If you suddenly implant chips in your brain, there are lots of fun things about this, but there are lots of scary things about it also.

Paul Middlebrooks
You and everyone mentions Neuralink. There are a lot of companies that are doing this thing. Neuralink just gets mentioned because they have a very famous person who started the company. I recently went to a neurostimulation/neural interfaces workshop. Part of that workshop, they actually had patients come and tell their stories to us, a bunch of neuroscientists. A lot of the people were neural engineers who were designing these kinds of brain implants, stimulators, and closed-loop stimulation systems, et cetera.

They had these patients come whose lives had been changed by these devices. Some of whom had the devices on their head as they were telling their story. Anyway, so far it's been very positive. There's the danger that it could lead to something very negative. It's early stages because you can imagine if you're like Rajesh Rao talks about neural co-processors, and others too, what we use, ChatGPT, is like a neural co-processor right now. It's just not directly implanted into our brain.

Jennifer Prendki
Implanted in your brain. Yes.

Paul Middlebrooks
I have made us jump all over the place here. We've covered a lot. I'm trying to find out, what have we not covered that we need to talk about here, because we still have some time.

Jennifer Prendki
What I was going to say, I write about those topics. I'm a little bit of a weird person.

Paul Middlebrooks
Ah, there we go. There it is.

Jennifer Prendki
[chuckles] People usually ask like, "Oh, you're in the AI, what excites you about AI?" and whatnot. It's a lot like everybody else. I see the benefits of

getting there. The reason why I eventually went into AI is that it freaks me out. I actually have friends, family, and whatnot, it's just like, "We are so happy you are going in there because there is a responsible person in the room."

Paul Middlebrooks
You care.

Jennifer Prendki
I care. I see the opportunity, but I also do see the risks.

Paul Middlebrooks
See, I care too, but I am so aware that I don't know the right answers. I think that the thing that I care about is that everyone who's in charge seems to think they know the right answers, and I know they don't know the right answers. I don't know where that--

Jennifer Prendki
I 100% don't know the right answer. I think I'm asking the right questions.

Paul Middlebrooks
Okay. That's great.

Jennifer Prendki
You need somebody in the room that says-- Look, just think about this. I was the person who was in charge. Let's talk about Gemma. Gemma is this open-source model that Google released, it's called an open-weight model.

Paul Middlebrooks
What is it called?

Jennifer Prendki
Gemma. It's an open-weight model. An open-weight model is the equivalent of Llama for Meta. It's the open-source version that companies release that you can take and deploy on your system and then build on top of. You look at ChatGPT or Gemini or whatnot, you don't actually get to play with the model. You use the model. Gemma, Llama, and whatnot, it's an existing model that you can take and deploy. If you train that with the wrong data, it is there to stay. You don't control it anymore. It's in the wild. It's very scary as a data person to say, "I'm the person responsible for what goes in there," because once it's out, it's out of my control completely. If there is sensitive content that went into that model, it's there to stay, and I'm responsible for it for the rest of time because I'm not going to be able to recall that data.

When I built the strategy for Gemma, I was really conscious about this. I know that if I'm not the person doing this, they will still release that model. Google, Meta, and whatnot are not going to decide not to release these models because one person refuses to try to give answers to this question. For me, I'd rather try to make a judgment call, even if it puts a lot of responsibility onto me to decide what goes in there. By writing about those topics, about what are the ethical implications of wetware? The ethical implications of generating consciousness.

If consciousness truly emerges from scale, you have to answer those questions. You cannot just stay here and say, "No, I'm not going to--" Again, we need to ask those questions before something bad happens. I don't have the answer. I will never have all of the answers. I don't think any of us is ever going to have all of these answers, but we need to come up with the best faith approach to this.

Because of the nature of those problems, you have to combine-- It's a multi-disciplinary problem. You need to have the inputs of physicists. You need the inputs of philosophers, ethicists, neuroscientists, and whatnot. What really worries me, as a physicist in a world of computer scientists, who believe that because they are computer scientists, they are the ones with all of the answers to what artificial intelligence is supposed to be, I think it's really important to have a Trojan horse in there that can force other opinions to be exposed and to be injected in there [laughs].

Paul Middlebrooks
Who's listening to you?

Jennifer Prendki
Who's listening to me?

Paul Middlebrooks
Yes.

Jennifer Prendki
Well, people from the outside [laughs]. I'm an expert in this field.

Paul Middlebrooks
Yes, so people should value--

Jennifer Prendki
Yes. Exactly right. You would be surprised. When I started talking about those things and writing about those things, people actually in the field were like, "I never thought about this,-

Paul Middlebrooks
No, Jesus.

Jennifer Prendki
-this is dangerous." At least they're starting to ask themselves those questions.

Paul Middlebrooks
Definitely.

Jennifer Prendki
For me, my fight is more getting people to speak. You don't agree with me on exact definition of consciousness or whatnot. I don't think I agree with myself. I don't think I have a clear opinion just yet. My opinion keeps changing over time as I keep listening to people hearing things, or whatnot. I want people to ask questions, and ask themselves questions, and realize that, "Look, if you're going to train an AI model on wetware, you have to ask the implications for humankind. You have to ask the implications for the being or the biological system that those AI systems are going to be deployed on."

I often say, I am always criticizing that we even call AI models "models," because what an AI model is, is an abstraction that-- It's an extrapolation on data. You have data, and between two data points, you have this data point here and here. If you try to probe a prediction in the middle, you're going to have some sort of an average. You're extrapolating from the data what the average behavior is going to be. In physics, a model is trying to explain-

Paul Middlebrooks
Explain.

Jennifer Prendki
-what the heck is going on. AI is not trying to explain. It's not even trying to reproduce. It's a very different definition of modeling. For me, I approach modeling is trying to explain and represent fundamental phenomena in nature.

Paul Middlebrooks
The abstraction.

Jennifer Prendki
AI doesn't do that. AI doesn't even try to do this, which is why I don't believe the typical AI model now can actually lead us to reproducing consciousness or life or whatnot. That's my position.

Paul Middlebrooks
What should we call them instead of models? I like that point. What would replace it?

Jennifer Prendki
Extrapolations.

Paul Middlebrooks
Extrapolations. Transformations, engineered. They're engineering models. What would be a term for engineering model, because it's more engineering than modeling, because I agree with what--

Jennifer Prendki
For me, models are meant to try to explain the world in the best way possible.

Paul Middlebrooks
There's no explanatory-- Although those models are being used in neuroscience as explanatory features for brain high-density populational neural recordings.

Jennifer Prendki
See, now you're going into an exploratory-explanatory AI, XAI, not [crosstalk]

Paul Middlebrooks
No, no. You mean, like, explainable AI?

Jennifer Prendki
Yes, explanatory AI.

Paul Middlebrooks
No, no. Convolutional neural network models- I'm not sure how much of the history of this that you know- are roughly designed based on what we know about our ventral visual stream and the way that it's layered. Even before that, Fukushima designed the neocognitron based on simple and complex cells and our visual cortex with layers for abstract things over times in ways our visual cortex is supposed to abstract things based on single neuron recordings in different layers.

It turns out, if you take those convolutional neural networks and you build them closer to the way that we think that our ventral visual stream is layered in a hierarchical structure and you train it on ImageNet, and if you look at the different layers of the convolutional neural network, the response properties of the units, after some linear decoding, match well to the response properties of various layers in your visual cortex. Then, oh, it's, "Aha, this is the best model of our visual cortex that we've ever had."

In that sense, they're explanatory, but they're more predictive. There's this battle in neuroscience, "Is this predicting? Is it explaining? Do we actually understand it," et cetera, et cetera.

Jennifer Prendki
Absolutely. You could say that it's because you have fundamental similarity in the topology of the way the brain works, and--

Paul Middlebrooks
The constraints.

Jennifer Prendki
Is it the constraints of the topology might explain or reproduce the same patterns? It doesn't necessarily explain in the proper sense of the term. What I mean in physics is you're trying to write equations that actually describe, as opposed to emulate or reproduce something. It might or it might not be the case. I think there is a reason why people don't know whether it's an explanation or not.

You have this example, for example, for evolution, like why do all mammals, or terrestrial animals, have four legs, and whatnot. Is it because we evolved from one another, or is it because this is the best way of being on planet Earth, given the constraints of gravity and the way that carbon life is being built?

Paul Middlebrooks
Your bet is that if we scale up with the current AI models, consciousness will not emerge. I agree with you. What will we get when we continue to scale?

Jennifer Prendki
Look, I'm a data person. I don't think it's even feasible to scale that much for because you are already at the limits of the data that exists.

Paul Middlebrooks
Are we? People always talk about that, like we're at the wall, but then you just step over the wall and it's a new--

Jennifer Prendki
I'll tell you, I think everybody needs to get that. The data that you have on the internet and everything that we've produced, like works of art, literature, or whatnot, is the representation of human knowledge. What models do is they extrapolate based on this. Those models can only do one thing. The first time somebody uses ChatGPT, they're like, "Oh, those chatbots have a superhuman capability." Is that true or not? In a way, yes, because if I'm a doctor, I'm the best expert on a specific type of disease.

The ideal AI model will do better than me because it will extract all information there is about-- Even if I'm the best expert worldwide, the union of all knowledge on that disease has to be bigger or equal to my own knowledge. It's not going to produce something new. It's not extra human. It's just the sum of all human knowledge on one given topic. We're always going to be bound by the content of that knowledge.

Paul Middlebrooks
Without any consciousness, by the way.

Jennifer Prendki
Without any consciousness.

Paul Middlebrooks
It's superhuman without any consciousness.

Jennifer Prendki
Anyways, it is superhuman in the sense that it's the union of all the experts on a given topic, assuming that you can create the perfect model that extracts that information. This is all there is to it. My opinion is there is an asymptotic limit to what you can achieve, which is extracting the maximum, most relevant amount of information from human knowledge, as long as we don't come up with AGI, whatever AGI is [chuckles].

Paul Middlebrooks
That asymptote was supposedly the same with training error. Then what's the double dip in training error that you get with scale? What's it called? It has a name. The training error goes down, and then after a while, it starts going up again. People thought we were at the limits.

Jennifer Prendki
Overfitting or whatnot?

Paul Middlebrooks
Yes, overfitting. Yes, overfitting. Then you just give it more data, and then it gets way better, so the generalization gets better. You don't think that there's going to be another double dip, you think that we're asymptote--

Jennifer Prendki
Look, it's information theory. We're extracting information from that data.

Paul Middlebrooks
That's all we got.

Jennifer Prendki
The nature. We're training, as long as it's a database. For me, if you really wanted to go further, you would need to have an experiential kind of AI. AI that they make hypotheses, go in the real world and say, "I have this new idea," and test whether-- This goes into physics. As long as you learn from existing content, you're not going to be able to go above that.

For me, the real AGI, if we're going to talk about this,-

Paul Middlebrooks
Oh-oh.

Jennifer Prendki
-if you want, it goes into embodiment or whatnot. It's like an AI that can create new ideas and say, "I don't know if this idea has merit, I'm going to go experiment and test it in the way that a scientist can do it."

Paul Middlebrooks
It's interesting how humans come up with new ideas. There are different ways of doing it, combining two previously unlike things, like the romantic poets. Then, you hear artists, musicians, their experience of it is like, "I don't know, it just came out of the ether." There's no accounting for where it comes from; it feels like it just drops down into your lap sometimes. It's almost always in the context of someone who is very skilled and is working a lot trying to produce things, and then one day it just comes. It seems like there's no effort. You've put a ton of effort into doing it, but we don't know how to generate ideation in. We're not going to get that with scale, right?

Jennifer Prendki
Yes. Definitely. You could define this as being human intuition, because it's always about combining things. You have relationships between items, things, ideas, concepts. For me, I've developed a lot of new IP, new ideas in my life by combining ideas from physics with ideas from biology, and just like, "What if we combine both?" You could say, not necessarily infinitely many, but an extremely large space of different combination of things. Then you have to evaluate if these ideas or combinations have merit. In order to experience whether they have merit, you have to either have intuition from real-world life and the real world, which is why we humans are able to do this and AIs can't, or you have to provide AIs with the capability to evaluate whether it's a good idea.

In order to do this, how do you define a good idea, for you as a person? How does a poet or a singer or a composer say, "This is good music"? It comes back to developing along with the idea, the evaluation system. It goes back to a balance. How do you evaluate whether a social pressure-- I believe my music is good because it made me feel good when I wrote it, or I listened to it, or it made other people happy, and I saw their smiles. Then I can measure that this has merit because people liked it.

Actually, this is also an interesting topic as well, because there is more and more research on-- Right now, the AI is an entity on its own. Some experts believe that the way you really truly reach the next generation is by letting AIs collaborate with one another, which is agents, or whatnot. That true intelligence or AGI will come from collaborative intelligence.

Paul Middlebrooks
Like Relate?

Jennifer Prendki
Yes. A good idea is more than the sum of its parts. That's why podcasts work: you are asking me questions that I wouldn't necessarily ask myself, and it pushes me to the limits of whether you are evaluating my ideas. I'm just like, "Oh, maybe this is interesting, maybe not." By means of culture and interaction, you can take intelligence to the next level. It goes back to we as a society can achieve a lot more than any of us taken separately would, but any of us, like the union of us living in separate worlds would.

Paul Middlebrooks
More is different. That's a complexity science concept. Before I ask you a couple more questions, let me just tell you something very human that just happened to me that would never happen to an LLM. I feel very embarrassed right now because the windmill reference that you made, it's from Don Quixote. I know that. I didn't remember it right when you said it. I knew that, that's so embarrassing that I didn't get it. I thought it was like an AI machine learning computer science term that I didn't know about, but no, it's from classical literature. There you go. Very human moment. There are so many take-homes. It's not that LLMs are missing a single thing; they're missing a list of ingredients, right?

Jennifer Prendki
Absolutely.

Paul Middlebrooks
Is it like a category error to you, almost, because it's so far off?

Jennifer Prendki
I think it's got a lot to do with infrastructure. Again, the questions, like, "What are we trying to achieve?" I still don't know exactly what we're trying to achieve.

Paul Middlebrooks
It's frustrating. Isn't that frustrating?

Jennifer Prendki
I work in this space. When generative AI became a thing, people were like, "What am I doing with this? Generating images for ads, our deep fakes," et cetera, et cetera. "What application does it have?" We are still having those questions. When I started in data as a data scientist, companies want to do something with data, and now they want to do something with AI, our generative AI, right?

Paul Middlebrooks
Yes.

Jennifer Prendki
AI agents. As long as we don't have an answer to this, and we start with the technology as opposed to solving a problem and trying to find what to do with this, I think it's complicated to figure out. For me, this is the nature of technology: new things appear. There are different applications, and there are different ways you could use this, and you have to embrace it and figure out later, if that makes sense.

Paul Middlebrooks
Do you feel any solace? I feel some solace that predictions about the future are always incorrect or 99% incorrect: flying cars examples, et cetera. Sometimes I find some solace in that I'm potentially most likely worrying about nothing because it's going to be nothing like the doomsayers say it's going to be. It's going to be nothing like the utopia that Aldous Huxley wrote about in *Brave New World*, et cetera.

Jennifer Prendki
I would say I'd rather think about it, and it ends up being not a problem, than not thinking about it.

Paul Middlebrooks
I know, but it's almost not worth thinking about then.

Jennifer Prendki
If you believe this way that it'll never be what you thought it would be. Let's talk about the worst-case scenario, and it won't be anything like it, as opposed to ignoring it.

Paul Middlebrooks
Okay, but let's take, like Nick Bostrom's book *Superintelligence*, and where the paperclip example came from. I was so frustrated reading that book because, actually, your writings remind me of it in this respect. Although your writings didn't make me frustrated because you weren't making strong claims. Nick Bostrom was making strong claims. Almost every sentence had the word "if" in it, and so this is a conditional. The probability when you multiply a billion conditionals together goes to zero. I don't have to worry about anything that Nick Bostrom was writing about, because they're all unlikely conditionals, and then they all have to happen. Maybe there's some--

Jennifer Prendki
It's like Drake's equation for extraterrestrial.

Paul Middlebrooks
Yes. Aliens are here, we know that. Come on, there's got to be intelligent life in the universe, don't you think?

Jennifer Prendki
Yes. Well, why not? Until somebody has proof for [laughs].

Paul Middlebrooks
If we're the most intelligent thing in the universe, man, it's pretty sad, isn't it? [laughs]

Jennifer Prendki
Isn't ChatGPT the most--

Paul Middlebrooks
Oh, gee.

Jennifer Prendki
That would be really sad.

Paul Middlebrooks
All right. Here's another quote. "It may turn out that intelligence was never about logic. It was always about life." This is from your writing. What I want to ask you is, what is eating at you right now? What's blocking your current thinking? What are the roadblocks to your current thinking that's bothering you right now that you don't have a grip on? This life issue, almost equating bringing intelligence into the domain of life processes, is something that I've been thinking about for a long time. I'm so frustrated that I cannot articulate why I believe that. What is that for you right now? What are you so frustrated about that you can't quite wrap your head around?

Jennifer Prendki
I'm frustrated about, when we talked about earlier with technology, it's just people rushing ahead, creating things without thinking about the consequences.

Paul Middlebrooks
We've always done that throughout human history, right?

Jennifer Prendki
Throughout human history, but now we're touching the substrate of life. For me, look at the example of artificial intelligence now. We are changing forever the job market. I don't think anybody would disagree with the fact that what a software engineer is today or was a year ago is going to be very different from what it's going to be.

Paul Middlebrooks
It's fast. The change is super fast.

Jennifer Prendki
It's fast. We are not prepared as a society to deal with it. I'm not talking about AI is becoming conscious or the robot apocalypse or whatnot. I'm talking about a real problem where lots of people are going to be left without a job. Are we ready to support them? What frustrates me is because we're operating in silos like this, where you have people working on wetware, you have Ilya Sutskever working on super alignment. It's great, and I respect that. He left OpenAI because he felt like OpenAI was not thinking about the consequences.

Paul Middlebrooks
He also had good money already. It's not like it was a monetary--

Jennifer Prendki
I'm not going to get into that. If you want to truly achieve super alignment, don't you want to have thinkers, ethicists, philosophers, and neuroscientist experts be part of that conversation? I think it's a little bit naïve. It was fine as long as we were-- You were building the car. Yes, it's going to change society, but we are not touching on deeply what makes us humans. Now that we're starting to talk about, "What is intelligence? What is consciousness? What is life?" I think this is something that concerns all of us.

Paul Middlebrooks
Last question for you. First of all, are there other things that you wanted to bring up that we haven't touched on here?

Jennifer Prendki
No. We talked about a lot of things.

Paul Middlebrooks
I know. I know.

Jennifer Prendki
You asked me a lot of difficult questions.

Paul Middlebrooks
That's great. My last question for you, then, is, you really do cite a lot of philosophy in your writings. I mentioned Kierkegaard, Merleau-Ponty. You cite Kant, Descartes, Aristotle. Heidegger, I didn't mention. I've gotten into Heidegger again lately. I got into existentialism when I was in high school, because of the angst. Did you have that in your background already? Have you been reading and appreciating philosophy this whole time, or have you revisited it?

Jennifer Prendki
Look, I wanted to be a physicist ever since I was a child because I wanted to understand the universe. For me, the meaning of life, why we are here. Why we behave like-- It goes hand-- It's also relatively unique for physicists, because physicists usually go into physics, because math, because they want to explain specific processes. I felt I wanted to understand the universe as a whole. What attracted me to DeepMind as a path in my career was exactly that, because Demis Hassabis, the founder, has as a goal to understand intelligence. [crosstalk] understanding--

Paul Middlebrooks
No. He has the goal to solve intelligence.

Jennifer Prendki
To solve intelligence.

Paul Middlebrooks
That's not necessarily the--

Jennifer Prendki
I think it's both. Also, I relate to this. For me, it goes together. You cannot understand the universe without trying to give sense to why are we here? What the heck are we supposed to-- Why were we brought here in the first place? For me, early on, it was always philosophy was always going hand in hand with physics. What better time now to reengage that conversation for a topic like artificial intelligence? Computer scientists pair together. Basically, when you try to get a job in AI, the interview questions are not about intelligence, they're about programming-

Paul Middlebrooks
Of course.

Jennifer Prendki
-algorithms and whatnot.

Paul Middlebrooks
Skills.

Jennifer Prendki
Of course it needs to be, but I think you still need to appreciate what is intelligence. Try to understand why do we need to be different from machines? Are we really different from ChatGPT in the way we're thinking? If you're serious about artificial intelligence, you need to be serious about intelligence. You need to be serious about not just the algorithms, but also the philosophy that comes with it.

Paul Middlebrooks
Jennifer, you are a, I guess what Nicholas Taleb calls a black swan-

Jennifer Prendki
A black swan.

Paul Middlebrooks
-given your background. You get that reference, right?

Jennifer Prendki
Yes.

Paul Middlebrooks
I didn't get windmills at first, but of course, you immediately get black swans. God, it was embarrassing. Don Quixote. I didn't look it up either. It really just came to me. I promise I didn't look it up. Anyway, thank you for being on here. Oh, this is what I was going to say, and it's kind of a

question. What I imagine is that you're experiencing some joy getting back into these questions and having some space and time to think about these things. Is it joyful? Is it satisfying? What does it feel like?

Jennifer Prendki
It is very satisfying, because it's like taking a little bit of distance. When you have a role in AI, it's very operational getting those models to work and whatnot, and forgetting, "Why am I doing this?" and "Why does it matter?" Look, I know I am a black swan, and I think I'm useful to engage this sort of conversation on the market.

Paul Middlebrooks
I hope so. I hope that you continue to do that. I hope that people lend their ears to you, people in the right places. Anyway, this has been really fun. We covered a lot of territory.

Jennifer Prendki
Thank you.

Paul Middlebrooks
Thanks for coming on. I appreciate it.

Jennifer Prendki
Absolutely.

[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 hear is a little slow, jazzy blues performed by my friend Kyle Donovan. Thank you for your support. See you next time.

[music]
