



Mazviita Chirimuuta: Abstraction, Idealization, and AI's Platonic Problem

January 23, 2026

Machine Learning Street Talk
Podcast Transcript

Revision: 97d51a29

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

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

Contents

Overview	2
The Problem of Generalizing Neuroscience	4
Abstraction vs. Idealization: The "Kaleidoscope"	5
Platonism in AI: Discovering or Inventing Patterns?	6
When Simplification Fails: The Reflex Theory	8
Behaviorism and the "Black Box" Trap	9
Haptic Realism: Knowledge Through Interaction	9
Is Nature Protean? The Myth of Converging Truth	12
The Computational Theory of Mind: A Useful Fiction?	14
Biological Constraints: Why Brains Aren't Just Neural Nets	15
Agency, Distal Causes, and Dennett's Stances	16
Searle's Challenge: Causal Powers and Understanding	19
Heidegger's Warning & The Experiment on Children	21
References	27

Overview

Professor Mazviita Chirimuuta joins us for a fascinating deep dive into the philosophy of neuroscience and what it really means to understand the mind.

What can neuroscience actually tell us about how the mind works? In this thought-provoking conversation, we explore the hidden assumptions behind computational theories of the brain, the limits of scientific abstraction, and why the question of machine consciousness might be more complicated than AI researchers assume.

Mazviita, author of *The Brain Abstracted*, brings a unique perspective shaped by her background in both neuroscience research and philosophy. She challenges us to think critically about the metaphors we use to understand cognition — from the reflex theory of the late 19th century to today’s dominant view of the brain as a computer.

Key topics explored:

The problem of oversimplification — Why scientific models necessarily leave things out, and how this can sometimes lead entire fields astray. The cautionary tale of reflex theory shows how elegant explanations can blind us to biological complexity.

Is the brain really a computer? — Mazviita unpacks the philosophical assumptions behind computational neuroscience and asks: if we can model anything computationally, what makes brains special? The answer might challenge everything you thought you knew about AI.

Haptic realism — A fresh way of thinking about scientific knowledge that emphasizes interaction over passive observation. Knowledge isn’t about reading the “source code of the universe” — it’s something we actively construct through engagement with the world.

Why embodiment matters for understanding — Can a disembodied language model truly understand? Mazviita makes a compelling case that human cognition is deeply entangled with our sensory-motor engagement and biological existence in ways that can’t simply be abstracted away.

Technology and human finitude — Drawing on Heidegger, we discuss how the dream of transcending our physical limitations through technology might reflect a fundamental misunderstanding of what it means to be a knower.

This conversation is essential viewing for anyone interested in AI, consciousness, philosophy of mind, or the future of cognitive science. Whether you’re skeptical of strong AI claims or a true believer in machine consciousness, Mazviita’s careful philosophical analysis will give you new tools for thinking through

these profound questions.

The Problem of Generalizing Neuroscience

Prof. Mazviita Chirimuuta

00:00:00

What should we say as philosophers about the relationship between neuroscience and philosophy of mind? So how much of our ideas about how the mind works can we read off from the results that neuroscience is telling us. The results you get in the lab can be well established and fine. There's nothing wrong with those data, but there's more of a problem of generalizing from what you learn in the lab to outside of the lab with neuroscience. For cognition in the real world, it's precisely all of that complexity and all of that interactivity that is really important to how, for example, animals are able to negotiate their environment. It's not an argument that AI is impossible so much as why does it seem so possible, so inevitable to people. If you look at the history of the development of the life sciences of psychology, there are certain shifts towards a much more mechanistic understanding of both what life is and what the mind is, which are very congenial to thinking that whatever is going on in animals like us in terms of the processes which lead to cognition. They're just mechanisms anyway. So why couldn't you put them into an actual machine and have that actual machine do what we do?

Dr. Tim Scarfe

00:01:21

Yes. But anyway, Mazviita, welcome to MLST It's amazing to have you here.

Prof. Mazviita Chirimuuta

00:01:26

Thanks so much for having me along.¹

Dr. Tim Scarfe

00:01:28

So you wrote this book, *The Brain Abstracted*. It's an amazing book. Folks at home should definitely buy this book. It's really, really good. Tell me about this book.

Prof. Mazviita Chirimuuta

00:01:36

It was quite a few years in the making. I think officially I started writing it maybe 2018 and it came out in 2024. But it was really based on ideas that I'd been working on. Maybe since 2014, I started publishing some philosophy of science papers about computational explanation in neuroscience. And then going back beyond that, some of my own experiences when I was doing training in neuroscience on the visual system and I was using computational models of the era before there

¹The Brain Abstracted — Book Prof. Chirimuuta's book, the central topic of the interview.

was deep learning or anything that fancy, thinking about really what does understanding the brain through this lens of computation by saying that we have models which not only simulate the brain as biological simulation using computers and all kinds of things or weather simulation such and so forth, but actually kind of alleged to duplicate the function of cells in the brain, which is this kind of additional claim which is made of about computational modeling when it's applied to the brain as this unique unique structure which is not only a biological organ, but also a kind of computer itself.

Abstraction vs. Idealization: The "Kaleidoscope"

Dr. Tim Scarfe

00:02:51

The arc of your book is we have this problem with simplification. Yeah. Because as scientists, we want to build legible theories about how the world works.

Prof. Mazviita Chirimuuta

00:03:00

A lot of philosophy of science in recent years has picked up this topic of abstraction and idealization. So abstraction is sort of quite a general word which can just mean sort of ignoring details which are there in concrete real life situations. So it would be, you know, familiar to you from doing sort of Newtonian problems in physics where your teacher tells you, well, there's always friction in real life, but we'll pretend that the friction isn't there. So you're leaving out a detail which is known to be there in the concrete concrete system. Idealization means, sort of attributing properties to the system that you're modeling in science which are known to be false. So for example in genetics modeling, the assumptions made of infinite populations, these kinds of idealizations often make the calculations more tractable but of course there's no such thing as an infinite population in real life. In some way, an abstraction is also always a false representation, always an idealization. So sometimes the difference between the 2 can be subtle. How I put this in the book is that an idealization kind of points us to the thought that when we have a scientific representation, we're kind of presenting something which is kind of cleaner and better than the thing in real life. When we talk about some someone being idealistic, it's like they have a view of how things should be and unfortunately reality does not live up to that. So idealization in science is often to do with sort of representing things mathematically in a way which is kind of cleaner and neater than could be possible in real life.

Dr. Tim Scarfe

00:04:44

And, on abstraction, you said in your book that there's the lofty philosophical version of abstraction which is, you know, upstairs in the heavens of Plato, Yeah. Think you Or even Galileo. There's this idea that these natural forms exist which are disconnected entirely from the sort of the spatial or the temporal realms. And then there's the more deflationary view of abstraction which is simply that we just ignore details. Now, I'm speaking with my good friend François Chollet, again tomorrow, he's releasing the new version of the ARC Challenge. And think he does have this, and many AI researchers do, they have this platonistic idea. He calls it the kaleidoscope effect, which is that the universe basically is written in code and what we see is like a kaleidoscope when all of the rules of the universe just get composed together in different ways and all we need to do as AI researchers is kind of decompose back into the rules. What could possibly go wrong?²

Platonism in AI: Discovering or Inventing Patterns?

Prof. Mazviita Chirimuuta

00:05:39

So I watched some of the videos with François. I found it really fascinating precisely this kaleidoscope hypothesis because seeing that as a philosopher I thought that's Plato because François precisely says we have the world of appearance. It's complicated. It looks intractable. It's messy. But underlying that, real reality is neat, mathematical decomposable. This is precisely this sort of contrast between the world of forms and the world of being, sort of eternal stable truth, and the world of becoming, appearance, messy, flowing, complicated reality. And so it goes back thousands of years in philosophy. It's really interesting that this is an assumption not only that AI researchers, make often, but it runs through science as a kind of justification for the pursuit of mathematical representations. Even when they sort of depart from known facts about the concrete physical systems in reality, the idea that the mathematical representation is getting you more to the truth, the underlying truth of how things are as opposed to what I call the diff like the down to earth view of what abstraction is and mathematical representation is, that it's something that we do because of our complicit cognitive limitations. So instead of thinking that the abstraction gets you like the higher level of reality, just saying that we do abstraction because we're finite knowers. There's limits to how much complexity any individual person or group of people can actually encompass

²Francois Chollet — Person Creator of Keras and the ARC Challenge, mentioned regarding the "kaleidoscope" view.

in their modeling strategies or representations. And actually it's only by pretending things that are more simple than they actually are that we get some traction. So that's like the down to earth mundane explanation of why abstraction is so much used in science.

Dr. Tim Scarfe

00:07:35

Yeah. It's so pervasive in the deep learning world. I mean, I also interviewed the folks who pioneered this geometric deep learning blueprint.

Prof. Mazviita Chirimuuta

00:07:43

Right.

Dr. Tim Scarfe

00:07:43

And that's the same idea, basically, that, you know, the world is described with geometry. And all we need to do is imbue these geometrical inductive priors into deep learning models and then they can, you know, essentially by reducing the degrees of freedom to ones which are reliant of how the universe works, we get where we want to go. I think³

Prof. Mazviita Chirimuuta

00:08:03

the notion of patterns and real patterns, to invoke Dennett's term there, is a helpful 1. So 1 thing that you could say is going on here is that, yes, there's lots of complexity there in the natural world, it's apparent in the data, but like if you just denoise the data a bit, underlying there there's a real pattern and we should we don't have to be like Platonist and weird about it, but there's just regularity that is sometimes masked by noise. That doesn't seem like too metaphysically problematic, but 1 of the questions that I sort of posed to that as a challenge to that, you know, very moderate view, and I and I say this frequently in the book, is when you're saying that some of the apparent dysregularity in the data is irrelevant, that's your decision as a scientist. It's not relevant to you at the moment, but that it could be relevant to someone else. It could be really important to how that system works in the natural world for reasons that you're not aware of. So when we sort of classify the signal versus noise in our datasets, we shouldn't ignore the fact that those are decisions that we're bringing to bear on our investigation. We shouldn't assume that we're just reading off the signal, the real pattern that is there in reality, and that there aren't very many other significant real patterns there. And to the extent that we're probably also kind of creating pattern through

³Real Patterns (Dennett) — Paper Daniel Dennett's 1991 paper on ontology and pattern reality, referenced to discuss signal vs noise in data.

the through the very denoising process that we bring about.

When Simplification Fails: The Reflex Theory

Dr. Tim Scarfe

00:09:42

Interesting. Physicists aren't under any illusion, so they know that Newton is an idealization. And just to contrast, you cited reflex theory. Of course Pavlov and the dogs, folks at home will know about that. And Newton is still around, we still use that, but we don't use reflex theory anymore.

Prof. Mazviita Chirimuuta

00:10:03

Yeah. Yeah. So this is a chapter that I present in the book as a case study of how oversimplification can get scientists on the wrong track. So the history of science is, like, 2020. We're looking at a theory about how the brain worked, was really dominant for a few decades at the end of the nineteenth century, beginning of the twentieth century. It's, yeah, familiar to us in the with Pavlov, with this idea that we can explain behavior in terms of reflexes which get conditioned and there's obviously learning involved with that. The most ambitious version of the theory said that all of the functions in the brain are basically versions of, reflex arcs, so sensory motor loops. So, very prestigious and sort of well regarded physiologist like Charles Sherrington was heavily invested in the reflex theory. But he admitted in his his book, *The Integrated Action of the Nervous System*, that this notion of a simple reflex is an idealization. It probably doesn't exist in real life, and yet this is the key that's going to kind of unlock, neurophysiology. It's gonna help us decompose and make sense of all of these different interactions that could be observed experimentally. So what seems to be going on there is that scientists were sort of taking that age old method, which is that it's a good heuristic to seek parsimonious explanations, to use Occam's razor, and the obvious thing to do was like, let's assume there's this thing that's there's a simple reflex, and then running with it way too far, actually never being able to explain the amount of data that they had initially thought that they would with it. And it's not clear how long the reflex theory could have gone on for if it hadn't been for the computational theory, sort of coming in during the second world war era, and basically providing an alternative explanatory framework which was also quite neat, and I would say provides its own kind of idealization toolbox.⁴

⁴The Integrated Action of the Nervous System — Book Charles Sherrington's classic physiology text, cited as an example of idealization.

Behaviorism and the "Black Box" Trap

Dr. Tim Scarfe

00:12:23

So a very popular thing in cognitive science is to say, well, if something behaves the same way as a cognizing human for example, we can maybe we might draw inferences that it has consciousness and has many other cognitive faculties. But there's always this kind of almost ignorance of the actual mechanism of the object of study.

Prof. Mazviita Chirimuuta

00:12:45

I think behaviourism is it has a bad name, but it's not that discontinuous with a lot of thinking which is sort of normal and still acceptable insights, which is to treat things as black boxes. This is precisely what the behaviorist said. It's like, the mind is opaque, it's hidden within the walls of someone else's individual subjectivity. As scientists, all we know are the inputs and the outputs, and we'll just track those. And that's like a version of what you just said if, well, if the inputs and the outputs, the behavior of this system are looking like what we know to be a conscious system elsewhere, well, let's just treat them as all of the same class of objects given that the only available information is the inputs and outputs. I think that kind of reasoning can be fine in certain contexts, but it's a philosophical leap to say that access that we have to our own thoughts and the presence or absence of subjectivity that we're aware of with other people is irrelevant to making these decisions or judgments about what other kinds of systems can have consciousness. So I think it's much too quick to just go behaviorist and say, well, there's no relevant difference between x and y, even if 1 is a person, 1 is a machine, just because we can say that there's some similarities in inputs and outputs.

Haptic Realism: Knowledge Through Interaction

Dr. Tim Scarfe

00:14:20

I think if I remember correctly, at 1 point you drew an imaginary kind of graph where you said on 1 axis we have science realism, which is where our scientific theories actually represent things in the world, then we have empiricism, is the idea that facts we receive tell us something about the world. And then there's this more interesting access, which I think you're very inspired by, which is this kind of constructivist idea. Right, yeah. Can you explain that?

Prof. Mazviita Chirimuuta

00:14:45

Yeah, so the constructivist path, sort of, which is different from the scientific realist and empiricist 1, sort of really runs with the idea that we are active makers of knowledge. It shouldn't be confused with the kind of constructivism that we have in some kind of, like, more extreme branches of sociology of knowledge which say that all scientific theories are social constructs and not constrained by phenomena that have been observed in nature. So it's So I'm not saying that scientific theories are merely constructed in the way that like poems could be a work of imagination and so forth. But the idea is that there's this interactivity between humans, groups of scientists, their plans as epistemic agents going out into the world with an agenda to find stuff out about certain phenomena in order to achieve certain goals, often technological, applied science goals, and there's some pushback from the things in nature themselves that they're investigating. But the idea that knowledge is always the product of this interactivity, so we cannot discount that there is a human framing side to this. We can't go along with this idea that a scientific theory is just like reading off the source code of the universe as if the human way of conceptualizing those phenomena had no bearing on the theory as it ultimately turns out. But we also can't discount that the theory that arises is constrained by how things happen to be that is worked out through that process of experimental interaction.

Dr. Tim Scarfe

00:16:31

You said, I think you were inspired by Immanuel Kant, so he

Prof. Mazviita Chirimuuta

00:16:35

had

Dr. Tim Scarfe

00:16:35

this transcendental idealism and please bring that in, but that somewhat informed your own view, is this haptic realism. Yeah. Can you introduce that?

Prof. Mazviita Chirimuuta

00:16:45

Yeah. So that's saying that knowledge comes about through this process of interaction. So this notion of haptic realism is emphasizing that it's through engagement. So haptics being like the sense of touch. The contrast here with us is with an ideal of knowledge which is based on this idea that we can know things in a disengaged way. If you think of, vision as the archetype for of knowledge, what happens when we look around at our surroundings, and use sight as a source of knowledge, we can get into this mindset where it seems like we do not have to interact with things in order to know them. We can just kind of absorb information passively, and then because we're not bringing about our representations in a kind of active way, it would seem to us, and I'm not saying this is how vision works, but it's a kind of conceit that often comes about if you use this very visual model for knowing. John Dewey called it the spectator theory of knowledge, so this is a clear predecessor from what I'm saying here, is that like we just look around, we absorb how things are, our knowledge is sort of entirely objective. It's almost like a god's eye view on reality. But if you think that scientific knowledge in particular is more kind of touch like, you can't ignore the fact that we sort of run into things. We have to pick things up, engage with them, ultimately change them in order for us to acquire knowledge of them. So you cannot discount the fact that we're kind of meddling with things in the process of, bringing about our knowledge. And another sort of dimension of this haptic metaphor is that our hands are not only a sensory organ, but they're also the means by which we manipulate things. So manipulation means precisely working with the hands. And so I think that really captures, if you like, the double face of scientific models. They're both of, means of acquisition of knowledge in the way that hands are also sensory organs. We have to find things out about the world through the sense of touch. But they're also means for changing things, for doing things.⁵⁶

Dr. Tim Scarfe

00:19:02

We speak about this in evolution. What would happen if you could just rerun evolution? Yeah. You know, what would happen if we could just have a parallel universe and the entire enterprise of science just ran again? And what you're alluding to is that it wouldn't be completely different, maybe there are some guardrails, but it is actually quite divergent.

⁵Transcendental Idealism (Kant) — Philosophy Immanuel Kant's philosophical framework, relevant to the discussion of how we perceive reality vs things-in-themselves.

⁶The Quest for Certainty (Dewey) — Book John Dewey's 1929 Gifford Lectures, first major work critiquing the spectator theory of knowledge that Haptic Realism challenges.

Prof. Mazviita Chirimuuta

00:19:19

Yeah, there's certainly like contingency in the history of science, you know, where people start out, cultural factors which prompt them to people to ask certain kinds of questions and not others. So a view that quite similar to what I say about haptic realism in the book is by Hasok Chang, who's a professor of philosophy of science at Cambridge. And he has a view which he calls realism for realistic people, that's the title of his new book. And he is an out and out pluralist about science. So he says that because there is contingency in the history of science, it means there are paths not taken, but we could maximize the acquisition of knowledge if we just like explored as many of those different paths as possible, which isn't something that I say in the book myself because I think there are also reasons why it makes sense to narrow views and parts of inquiry, and also we don't have like unlimited resources. But yeah, sure there are opportunity costs that come along with like taking a certain path and there are others not pursued.⁷

Is Nature Protean? The Myth of Converging Truth

Dr. Tim Scarfe

00:20:23

In the enterprise of science, there might be a trope or an idealization that we're getting closer to the truth. Yeah. Do you think that's the case? Do you think as the enterprise of science just progresses that we're getting closer to the truth? Or could we be in cul de sacs basins of attraction and so on?

Prof. Mazviita Chirimuuta

00:20:40

That's very much associated with scientific realism. So there's this view that there is 1 way nature is and science succeeds in so far as scientific representations conform to this 1 way that nature nature is. From my from my view, sort of takes very seriously the idea that nature could just be sort of inexhaustibly complex. So if you ever try to sort If you ever pin it down in 1 representation, there are ways also that it could be represented inexhaustibly many different varieties of ways that you can investigate it, and then also ways that any 1 representation is lacking. So there's a kind of inherent sort of lack of convergence that that picture brings about. 1 of the ways of expressing this is to say that nature is protean. There's this mythological character, called Proteus, who was a shape shifter. This sort of this, yeah, mythological being that lived in the sea and it would he would keep changing his shape. You couldn't if you could pin

⁷Realism for Realistic People (Chang) — Book Hasok Chang's pragmatist philosophy of science, cited as a compatible view to Haptic Realism.

him down, he would answer you a question and tell you the truth. But the thing was you had to pin him down. And I think this is a really nice illustration of like what's going on with our interactions with nature as scientists. Nature is sort of inexhaustibly complex. There's all kinds of patterns and things going on there. It can be pinned down and we can get true answers, but it but when we sort of release our grip, it will carry on shape shifting, and there's lots of other ways that it could be. So, yeah, 1 final theory. I'm not so convinced by that. This is very much a view that I think makes sense if your basis for your theory as a philosopher of science is really the biological sciences, which is where I'm coming from. If you're a physicist, it seems much more natural to think that there is 1 fundamental set of laws of the universe which is gonna be nailed down once and for all and could explain everything. Biology just throws up lots and lots of examples, particularities. It tends to be less considered sort of less intellectually satisfying in comparison with physics. Oh, you can just spend all your time in biology doing stamp collecting because there's this thing and there's this thing and there's this thing. How do you tie it all together theoretically? But on the other hand I think that if you take that particularity and that shifting quality to biological phenomena, then actually it just forces you to think about knowledge differently.

Dr. Tim Scarfe

00:23:20

In in your book you you spoke about a trajectory I suppose of possibly failures of simplification. We just spoke about reflex theory. 1 of the big things is this metaphor of cognition or the brain perhaps as being a kind of computer. Yeah. And you spoke about the early roots of this from reflex theory to cybernetics and computationalism. Can you sketch that out?

The Computational Theory of Mind: A Useful Fiction?

Prof. Mazviita Chirimuuta

00:23:44

So the connecting thread is really this idea that what cognition is, is something that is machine like. That what I know going back to the seventeenth century, this is a view associated with the philosopher and physicist and physiologist Rene Descartes who said, we need to give up, we need to go along with this idea that everything that happens in the body is explicable in terms of quite simple mechanistic forces. And this, you know, idea that biological systems are machine like has obviously been hugely influential in the different branches of science. The reflex theory was 1 instance of that. People often said machine like reflexes and making comparisons with sort of Newtonian decomposition. With the computational framework, have an actual machine, a digital or analog computer, which could be compared with brain processes. I mean, cybernetics is an interesting stage along the way because they were buildings of little devices that had some degree of autonomy made up of and and supposed to be emulating versions of negative and positive feedback and as hypothesized to occur in the body. But, yeah, I would say at the core of this research idea is that if what's going on in the body is ultimately a mechanistic process, then by redoing engineering with this non living system which is capturing some of the core operating principles that we find in biology, then we can use that device as a map, as a as a resource to then reinterpret what's going on in the biological system. You saw that with, for example, McCulloch and Pitts in their 1943 sort of landmark paper of interpreting neuronal cells as logic gates and then saying, yeah, you could build a computer out of neural nets. This is the origin of neural nets as we know them today. This is the birth of the idea. But then using that notion that neurons are logic gates to then interpret what's going on in physiology. So what I describe in, and it's in chapter 4 of the book, is this sort of back and forth thing of of making devices which are somewhat inspired by biology, and then using those then as the lens through which to review biology again. And I say that the advantage and the appeal of this process is that it allows you to or gives you kind of license to ignore so many things that are happening in the brain and nervous system which are just not shared with non living machines. Like all of the biochemistry, all of the ways that neural tissue is shaped by vasculature and interacts with the immune system, and all of that sort of background stuff that if you're a theoretical computational neuroscience, you can say, I'm only interested in the computational properties of the brain. I don't need to care about all of that messy biological detail. So it gives you a kind of tunnel vision which a scientist can be fine to have tunnel vision. You can't take in everything at once all of the time. But what I take issue with is the kind of ontologization of that saying that because compu-

tational neuroscience is this successful field of inquiry, we know now that the brain is a computer. I think that is not an inference we should make.⁸

Biological Constraints: Why Brains Aren't Just Neural Nets

Dr. Tim Scarfe

00:27:25

Yeah. I mean I don't think connectionists typically argue that I mean they would say it's a different mechanism. Yeah. But they think that there's some kind of functional equivalence. And that's the thing, because so many folks in AI at the moment, they are interested in biologically plausible architectures. So what if, like the cyberneticists did, what if we have more autonomy, diversity, agency and so on? And they fundamentally think that, guess they make the assumption that the world is a machine and if we replicate it with sufficient fidelity, then we can reproduce the behavior.

Prof. Mazviita Chirimuuta

00:28:02

To what extent are the mechanisms of the brain inherently bound up with the fact that the implementation here is in living tissue? So I think it's really this sort of tantalizing evidence about how the extent to which sort of brain processes and signaling between neurons, not just the electrical specialized signaling that neurons do, but biochemically, it's kind of outgrowths of signaling that's happening elsewhere in the body all of the time. So that there's nothing that we shouldn't think of neuronal cells as distinctively cognitive as opposed to the other cells in the body, but that they're extensions of the ways that, cells signal anyway. And if neuronal function is so much just a manifestation of what's happening with metabolizing cells anyway, that makes it more of a stretch to say that a machine that's not living could have the same functionality. Yeah. I mean, no one's trying to sort of build artificial neural networks with living cells.⁹

Dr. Tim Scarfe

00:29:17

No. No. But, I mean, there's an analogy in neural networks. There's this thing called the lottery ticket hypothesis, which spoke about pruning. Yeah. And what the researchers found is that you train this big dense neural network, and after it's trained, because you need the density for stochastic gradient descent for training tractability, after it's trained, you can prune away 90% of the connec-

⁸A Logical Calculus of Ideas... (1943) — Paper The seminal McCulloch & Pitts paper treating neurons as logic gates.

⁹The Lottery Ticket Hypothesis — Paper Frankle & Carbin's paper on pruning neural networks, used as a counter-analogy.

tions and it still works the same And, maybe evolution and our biological instantiation, maybe it's the same thing. We've been through this billion plus year training process and all of these things we think are important like the instantiation, the auto poiesis, the agency and so on, maybe those are vestigial and we can now just kind of snip, snip, snip and we can just kind of create this abstract version. It seems reasonable. What do you think?

Prof. Mazviita Chirimuuta

00:30:13

It's just vestigial. I mean, I think we really need to take seriously the economy that is there and of biological information processing. Do a lot more with a very limited energy budget running our brains every day than is like artificial neural networks are really really expensive to run. It doesn't strike me that biological cognition could get away with being that wasteful. That surely to keep things sort of blowing up in terms of like energy being consumed for information processing biologically, that there must have been a fair amount of pruning on the way.

Agency, Distal Causes, and Dennett's Stances

Dr. Tim Scarfe

00:31:01

I kind of think of agency as being a bit of a spectrum from, you you can think of it in a deflationary sense as being this autonomous thing that's the cause of its own actions. Right. And then the deep philosophical sense is that there's this intentionality and you can control the future and whatnot. And it rather speaks to like the physicist's view is that, you know, all of these, you have this light cone, you have all these micro interactions. Yeah. And, of course, that's beyond our cognitive horizon, so we develop ideas of representations where we can have these distal relationships between things that are in our mind and things that are far away in time and space. And I suppose you think of that as being another form of idealization. But the fascinating thing though, when I think about agency I think about it in terms of like apparent causal disconnectedness. We agents because you have consistent beliefs and ideas and you're not just an impulse response machine that's your being actions aren't determined entirely by this situation, you're a person. Yeah. And I kind of perceive that as a kind of causal disconnectedness.

Prof. Mazviita Chirimuuta

00:32:08

Yeah. Yeah, I agree. I mean, the what I say in the book in that chapter, I set out and I say this to be sort of very metaphysically neutral about what representation is,

what intentionality is, But at the same time, not what I directly wrote in the book. I think I agree with you that there is something very important about connecting the notion of agency and intelligence with this thing of like being responsive to what is actually very distal. It could be distal in time and space, it could be like distal because it happened a long time ago, but this is what biological memory is. Things that happen like to you when you were a baby affect how you are now. Physical systems, like non living physical systems, they're much more constrained in their in their actions, and I don't mean that in like action, but just what they do, what happens to them, by what's proximal to them. There's like the distal is always screened off by the proximal, if that makes sense. Like the Whereas for you, all of these things that happened in the past could be as relevant as anything that happens in the room right now, or your ideas about the future would be relevant to what you're saying right now. So yeah, this this notion of being sensitive to what's not immediately driving you in your surroundings, I think that's a really important, like, thing to latch onto in, like, delineating at least the class of systems that we wanna call cognitive to ones that we would say are sort of merely physical, not intelligent in any important sense of the word.

Dr. Tim Scarfe

00:33:53

Very cool. So we're trying to I suppose partition the world into logical units that we can understand and agents are a great version of that. And Daniel Dennett of course, he had the 3 stances, he had like physical stance, the design stance and the intentional stance as a way of kind of like building useful explanations of how things behave introduce those. But you said that you didn't quite agree with that because to Dennett, it's a hierarchy which means the intentional stance perhaps has precedence over Oh, the other

Prof. Mazviita Chirimuuta

00:34:28

so actually it's kind of the reverse. It's it's as if the physical stance has an ontological priority, like that's what's really there, but it's useful to use the design and the intentional stance.

Dr. Tim Scarfe

00:34:42

Very interesting. But but you said for for you, you don't really have that prioritization, you're open minded.

Prof. Mazviita Chirimuuta

00:34:47

Yeah. So that's part of the sort of metaphysical neutrality that I set out with the chapter is to say, okay, let's not go in with the assumption that low level physical causes are the, like, the primary causes of everything. Yeah. It's a it's a way of of, if you like, taking intentional phenomena at face value, intentional in the sense of like bearing representations, and I think 1 of my criticisms in that chapter is this agenda which is our philosophy of mind to say that, okay, if representation is real, we need to be able to tell a physical story about how it comes about. And I'm just saying, why not why go along with that project if there's no and this is actually going a Dennettian view is is, you might see it as that If talking about representations and intentionality is useful within the sciences, why not just take that in face at face value and not say that that needs to be established by making it coherent with some causal story about what's going on in terms of non intentional physical interactions. So that was the position there.

Dr. Tim Scarfe

00:36:07

With Putnam's Rock, right, you know, he said that you can take any open physical system and you can configure it in such a way to have the same types of information processing and then why wouldn't that have all of the, you know, all of the cognitive properties that

Prof. Mazviita Chirimuuta

00:36:23

When you're making a claim that, you know, the brain is a computer and that that explains cognition, what grounds have you got for saying that any arbitrary physical system is actually implementing a computation just from looking at its physical dynamics? If it's purely a question of mapping the physical dynamics to a computational formalism, then any physical system can afford a mapping of that sort, whether it's a rock, whether it's the sofa, whether it's my stomach as opposed to my brain. And so I So that's a challenge to the computational theory of mind, that it's assuming that brains implement computations just because we can sort of model them computationally, but we can model all kinds of things computationally, what makes brains special?

Searle's Challenge: Causal Powers and Understanding

Dr. Tim Scarfe

00:37:13

So what about this idea of whether computation itself has causal powers?

Prof. Mazviita Chirimuuta

00:37:18

I don't think it does. Computation itself is mathematical formalism. It's like exists it's it's a it is a mathematical structure. Things that have causal powers are concrete physical systems. So I just think they're different kinds of things.

Dr. Tim Scarfe

00:37:36

So Searle famously argued that the reason why we can't build strong AI is that computation doesn't have causal powers, it's implemented in So what does have causal powers are the machines actually implement the computation. But couldn't you sort of say, well, there is still a causal graph. Perhaps you would argue that computation isn't a node in that causal graph, it's just some kind of an aspect of it.

Prof. Mazviita Chirimuuta

00:38:02

Yeah, I mean, I think it just goes back to this issue that computation in and of itself is not the kind of thing that could have causal powers. I think Searle's point, and this is in the Rediscovery of Mind on this, was an interesting 1. It was kind of maybe kind of subtle and it kind of gets lost in the wash of like AI back and forth and Searle bashing, which happens a lot But it was about the kinds of ways that we form explanation in the sciences. And his What his point was that cognition, if it's anything, is something as part of the physical realm, the realm of causation. The assumption of the computational theory of mind, and he argues that this is very dominant within cognitive science, is that you can explain this phenomenon, is a phenomenon of the concrete physical world, through this non causal thing, which is computation, and suddenly there's no gap that needs to be closed. And I and I think that's that's a fair point, is that there's something inherently, that needs, like, further justification of why, of all of the things that happen in the concrete physical world that demand explanation, why we reach outside of the concrete realm of physical causation into computation in order to explain this thing, cognition.¹⁰

¹⁰The Rediscovery of the Mind (Searle) — Book John Searle's 1992 book on consciousness and the mind, cited regarding the critique of computational explanation.

Dr. Tim Scarfe

00:39:35

Another argument Searle was making was how machines couldn't understand. Yeah. And, of course, he was talking about things like semantics. Yeah. Do feel that they could understand?

Prof. Mazviita Chirimuuta

00:39:46

So what do I think? I think certainly there's more to human understanding than that. I think that a thing about human cognition and animal cognition in general is that my view is that it's not a set of discrete modules that work separately from 1 another. I think language is bound up with sensory motor engagement, and how we, likewise, how we perceive the world is shaped by linguistic concept formation and everything like that. So the idea that you could just sort of detach off a language faculty, have it replicated in an LLM that doesn't have the other bits of our cognition and it doesn't have embodiment, doesn't have the capacity to engage with the world, and that it could have understanding in the same way that we do. I find that implausible.

Dr. Tim Scarfe

00:40:44

Interesting. But again, if do the galaxy brain thing and we say we can embed robots in the physical world, give them sensory motor affordances all the rest of it, you know, are many replies to the Chinese room argument about this, like the robot replying type. Yeah, yeah. You know, would they have a little bit more understanding?

Prof. Mazviita Chirimuuta

00:41:02

Yeah, I mean, that's getting more along the lines of the kind of thing that could have understanding. I think there's also more relevant stuff in the background of what it is to be a biological thing, in that some things are inherently meaningful to you and relevant to you because they're connected with the demands that the challenges of your environment place to you. Like I was saying before, you know, life is a Being alive is a way of being always in a situation that is problematic to you. So saliency and meaning, I think, are connected to that. It's not to say that you couldn't have robots with more and more precarious lifelike situations and maybe, I don't know, it's not beyond the realms of possibility that then you start seeing understanding in ways that are more like that as well.

Heidegger's Warning & The Experiment on Children

Dr. Tim Scarfe

00:41:58

Let's talk a little bit about Heidegger. So he spoke a lot about our relationship with technology, and I know this is something you've been thinking about. Can you tell us about that?

Prof. Mazviita Chirimuuta

00:42:06

So 1 of the things that Heidegger said, and in many ways he was a grandiose and unpleasant person, but his grandiosity kind of manifested in how he could have read the importance of philosophy to things about the modern world today. So 1 of the things he said was that technology and, cybernetics for him, we could say AI for us today, was the culmination of a metaphysical tradition. Right? So it's because the history of philosophy set out on the certain path that it did that we are here today now with these technologies which seem to be having such a transformative role in our lives, and to the point that we feel like not in control of these technologies. So I'm not a kind of philosophical determinist like that. I don't think, you know, just because some philosophers and ancient Greeks said certain things that this is why we have AI today. But I think there are some features of his The contrast he draws with the philosophical tradition in his own account of what it is to be a human person that give us, like, an interesting perspective on, like, what is being assumed in that path towards AI. So 1 of the things he really insists on is human finitude. Like, we are inherently finite bounded knowers as communities, and we're communities of knowers. But that the philosophical tradition has encouraged a kind of leap that encouraged the idea that something about us as knowers kind of crosses beyond the boundaries of finitude into like a universal, boundless realm of knowledge. And I I do think that the very idea that a non situated, sort of non embodied absorber of facts like an LLM that kind of just sits there and sucks in all the information in the world, that that could somehow be a counterpart to how we know things as human beings. I think that's an instantiation of this lack of acknowledgement of human finitude because it's human finitude coming from a culture which is expert at some things but not other things, and whose knowledge is grounded in a discrete set of sensory experiences that are not accessible to other people, that boundedness of knowledge. To say that that is not inherent what it is to be a knower, that a purely disembodied, absorber of facts is what we are, I think that sort of revealing that lack of acknowledgement of finitude which is there in that tradition that he criticized.

Dr. Tim Scarfe

00:44:54

We're kind of moving into this technologically embedded world, and it's changing our nature in some way.

Prof. Mazviita Chirimuuta

00:45:01

This tradition that Heidegger criticizes from that path from philosophy to technology, of this aspiration to sort of transcend embodiment, transcend materiality, to create for ourselves this leap into a almost spiritual world of pure information. And it's interesting how, like, infrastructure is presented to consumers, people like me, as like behind the scenes immaterial, not really connected with real world constraints. Like, it's the cloud, it floats above us. There's no there's no almost cost to it. It's weightless. That's not how technology infrastructure works, but it seems like that's what we'd like it to be. We'd like this idea that all of this information age that we have us around today is not connected with real world stuff like actually building computers and shortages of resources and consumption of energy and all of those things. We like to think of it as disconnected from that in ways that it's obviously not.

Dr. Tim Scarfe

00:46:13

I agree that it's not because clearly we live in the physical world but there's an apparent disconnection. There was the digital divide of course in the 1980s and it was very difficult even to work and get a job if you didn't know about computers. Now, even to get a driving license, it's all done online. The legal landscape is not physical anymore, it's And even like the European GDPR regulations, as Floridi says, you're a digital entity and it's diffused and so on so, you know, Facebook and whatnot. It certainly feels like we live our world, we live our lives in the information world. Yeah. And it's becoming more and more kind of confusing in that sense.

Prof. Mazviita Chirimuuta

00:46:58

So living in the information world, do you mean by that kind of the information that we get through devices, through kind of this diffuse spread of technologically connected things is more salient to us than our experience with like our concrete here and now situation.¹¹

¹¹So You've Been Publicly Shamed (Ronson) — Book Jon Ronson's book on online shaming, mentioned regarding social pressure in the digital world.

Dr. Tim Scarfe

00:47:18

You know, there was that Jon Ronson book about you've been shamed or something like that you know, basically saying that there's almost more controlling pressures in the social media world than there is in our physical world. Sure.

Prof. Mazviita Chirimuuta

00:47:30

Yeah. I mean, think all of that is possible because human beings are imaginative creatures. Like, we live in imaginative worlds, and it's not so much the social media is always in our face, well, maybe it is, people looking at the screens, but our way of thinking through our own life history and how we project and everything else, there's a whole world that we're also constructing around that. So I think we're co creating this digital world, and it it doesn't work unless we're imaginatively and emotionally invested in this. So my view goes back to this Kantian idea, which is also goes back to this idea of human finitude, that there's something wrong with thinking that what you are as a knower is the kind of being that can float free of your environment and just sort of regard it from above, and like take in all the information that's there as it is by itself without your impact on the world. And my point is that we're not those kinds of beings. We only acquire knowledge through this arduous process of interaction, which means that we cannot claim to have that god's eye neutral view on things, but that it would be It's a mistake about knowledge and about ourselves as knowers to think that that is an aspiration that makes sense for us.

Dr. Tim Scarfe

00:48:56

Is it possible for both things to be true at the same time? I'm not making any like weird claims that we're not actually really physically situated. But it's almost like we're increasingly disconnected in many aspects of our mental life.

Prof. Mazviita Chirimuuta

00:49:15

From

Dr. Tim Scarfe

00:49:16

From our direct physical experiences.

Prof. Mazviita Chirimuuta

00:49:18

Yeah. So I think that's true in terms of like what we pay attention to. I mean people look at their phones instead of looking out the window when they're on a train, not looking at the people around them. So there's this where our attention goes can be wherever the Internet wants to place us. But the phone is still like a concrete physical device that with little light emitting diodes that, you know, produce like photons in our retinas. Right? So I think it's about again, I go back to saying it's about the imaginary around that around that, that in our minds, this isn't just a device that just is showing images right now, but it's this portal to this other place where we can be disconnected or dissociated from where we are right now. But I think that need to live in an imaginary, to be disconnected, is probably always there and manifests in different ways in human culture, through fantasy, mythology, all kinds of things.

Dr. Tim Scarfe

00:50:28

And we can debate whether it is imaginary or not with I interviewed Chalmers about his Reality Plus book and he was talking about, you know, whether virtual worlds are real and stuff like that. I mean, certainly with social media, it is still real. You know, like these people Instagram, they're having real experiences, and I'm vicariously experiencing through my phone. So the question is whether that's a sort of like a truncation of our mental life or whether it's expansive.¹²

Prof. Mazviita Chirimuuta

00:50:50

Yeah. I mean, this is something that I think becomes a question of ethics and phenomenology. This like the part of philosophy that really studies experience and what we draw from it in its own sake. So there's there's always an opportunity cost. People if you're looking at your frames absorbed into 1 thing, then you're not absorbed in other things. And ethically, is it right to do that? I think that's really important. We're running a big experiment on the next generation because kids, young young children nowadays spend a lot less time looking at people's faces than they used to. Will they be socialized in a way that will allow them to lead happy lives later on? We don't know.

Dr. Tim Scarfe

00:51:40

Is that something you worry about?

¹²Reality+ (Chalmers) — Book David Chalmers' book on virtual worlds and philosophy, mentioned in contrast to the discussion.

Prof. Mazviita Chirimuuta

00:51:41

I do, I do, yeah.

Dr. Tim Scarfe

00:51:44

Tell me more.

Prof. Mazviita Chirimuuta

00:51:46

Well, just because it's a massive experiment. I mean, it's obvious that from developmental psychology that young children are predisposed to be sort of paying attention to social interactions with the people around them, the gaze and faces and features and all of that. If they have less opportunity to make those connections, have those experiences at a very young age, it seems like that's bound to have an effect on how they relate to people later on. Don't see how it couldn't.

Dr. Tim Scarfe

00:52:20

I mean, there's a thought experiment that I had a debate with an AI doomer and because I'm a big externalist, I think we're embedded in these cognitive as you do. And I gave the example, imagine that a child was brought up in a hermetically sealed chamber and you gave them a computer and the internet and they could still learn to do lots of clever things and so on. But we're becoming a bit more like that. It's not quite a hermetically sealed chamber but we're mediating through the internet and so on. And that could go 1 of 2 ways, right? As you say, it could dramatically kind of truncate our mental life and cause lots of problems, or maybe it just won't be a big problem.

Prof. Mazviita Chirimuuta

00:53:02

I mean, there were already experiments done on monkeys in the 1950s of depriving them of maternal contacts, and it didn't work out well for those monkeys. Mean, should we be doing that experiment on children, sort of depriving them of what is instinctively the kind of social engagement that you know young humans seem to need to develop normally.

Dr. Tim Scarfe

00:53:26

Amazing. Mazviita, thank you so much for joining us today It's been a great conversation.

Prof. Mazviita Chirimuuta

00:53:30

Thank you very much, Tim. I've enjoyed it.

References

[1] **The Brain Abstracted**

<https://mitpress.mit.edu/9780262548045/the-brain-abstracted/>
Book Prof. Chirimuuta's book, the central topic of the interview.

[2] **Francois Chollet**

<https://arcprize.org/>
Person Creator of Keras and the ARC Challenge, mentioned regarding the "kaleidoscope" view.

[3] **Real Patterns (Dennett)**

https://ruccs.rutgers.edu/images/personal-zenon-pylyshyn/class-info/FP2012/FP2012_readings/Dennett_RealPatterns.pdf
Paper Daniel Dennett's 1991 paper on ontology and pattern reality, referenced to discuss signal vs noise in data.

[4] **The Integrated Action of the Nervous System**

<https://www.amazon.sg/integrative-action-nervous-system/dp/9354179029>
Book Charles Sherrington's classic physiology text, cited as an example of idealization.

[5] **Transcendental Idealism (Kant)**

<https://plato.stanford.edu/entries/kant-transcendental-idealism/>
Philosophy Immanuel Kant's philosophical framework, relevant to the discussion of how we perceive reality vs things-in-themselves.

[6] **The Quest for Certainty (Dewey)**

<https://www.amazon.com/Quest-Certainty-Relation-Knowledge-Lectures/dp/0399501916>
Book John Dewey's 1929 Gifford Lectures, first major work critiquing the spectator theory of knowledge that Haptic Realism challenges.

[7] **Realism for Realistic People (Chang)**

<https://www.cambridge.org/core/books/realism-for-realistic-people/ACC93A7F03B15AA>
Book Hasok Chang's pragmatist philosophy of science, cited as a compatible view to Haptic Realism.

[8] **A Logical Calculus of Ideas... (1943)**

<https://link.springer.com/article/10.1007/BF02478259>
Paper The seminal McCulloch & Pitts paper treating neurons as logic gates.

[9] **The Lottery Ticket Hypothesis**

<https://arxiv.org/abs/1803.03635>
Paper Frankle & Carbin's paper on pruning neural networks, used as a counter-analogy.

[10] The Rediscovery of the Mind (Searle)

<https://mitpress.mit.edu/9780262691543/the-rediscovery-of-the-mind/>

Book John Searle's 1992 book on consciousness and the mind, cited regarding the critique of computational explanation.

[11] So You've Been Publicly Shamed (Ronson)

<https://www.amazon.com/So-Youve-Been-Publicly-Shamed/dp/1594634017>

Book Jon Ronson's book on online shaming, mentioned regarding social pressure in the digital world.

[12] Reality+ (Chalmers)

<https://consc.net/reality/>

Book David Chalmers' book on virtual worlds and philosophy, mentioned in contrast to the discussion.