/mlst

Why does the Chinese Room still haunt AI?

March 24, 2026



Machine Learning Street Talk

Podcast Transcript

Revision: 887de1d5

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

Doom Debates Recap and LLM Turing Completeness 3

Potentially Infinite vs Actually Infinite Memory 5

Algorithms, FSA Controllers, and the Training Problem 7

Can We Train on Turing Machine Programs Directly? 9

Hybrid Systems, Scaffolding, and Novel Programming 12

The Chinese Room Argument 15

Searle on Physical Realization and Causal Structure 19

Emergence, Wolfram, and Computational Subjectivity 22

Friston, Self-Awareness, and the Path to Consciousness 26

Causal Structures, Concept Role Semantics, and Understanding in AI 29

Consciousness, Computation, and Philosophical Puzzles 32

The Nobel Prize for Deep Learning 34

AI Harm, Ethics, and Public Policy 37

References 41

Overview

Keith Duggar and Tim Scarfe return for the second edition of their hosts-only philosophical steakhouse, picking up right where the last episode left off – with Keith fresh from his appearance on Liron Shapira's Doom Debates show. The conversation starts technical and stays there for a good while, working through a precise argument about why standard autoregressive LLMs are not Turing complete in the technical sense.

Keith lays out the distinction between potentially infinite and actually infinite memory, explains why the algorithms learned through gradient descent on fixed-context-window architectures come from the finite state automata class rather than the Turing machine class, and confronts the practical objection head-on: even if this is technically true, does it matter? Tim pushes back with RAG systems that expand effective memory, and the two work through whether closing the read-write loop could unlock a richer algorithm space.

The middle hour is a sustained engagement with John Searle's Chinese Room argument. They dig into the gap between syntax and semantics, why gears churning out responses lack something even if they pass behavioral tests, and what Searle means by requiring the right causal structure for understanding. Bishop's Dancing with Pixies reductio, Wolfram's computational boundedness, and Friston's temporal-counterfactual depth model of self-awareness all get pulled into the conversation. Keith makes one of the more interesting concessions: he cannot rule out that a sufficiently complex program with the right causal structure could cross the threshold into genuine understanding.

The final act covers the Nobel Prize controversy around deep learning (Keith declines to judge but notes academia's political problems), Chomsky's critique that neural networks are not scientific theories, and a framework Keith developed on Doom Debates: focusing on AI harm rather than AI doom as a way to build broader policy coalitions. Tim distinguishes AI ethics from AI safety and both agree the speculative existential arguments, while philosophically legitimate, should not drive policy at the same level as the concrete social harms already visible.

Recorded Friday 11th October 2024.

Doom Debates Recap and LLM Turing Completeness

Keith Duggar 00:00:00

You know, a mathematical model of a black hole is not a black hole. A weather simulation is not weather. Okay? A computational process simulating the human mind is not the human mind. There's this gap. There's this gap between our abstractions, between our computations, between our math and physical realization, and it's fascinating. Welcome back to Tim and Keith's philosophical steakhouse. Exactly. Our our our fireside chats. I think it's been, almost a month since we did the 1 that or maybe 5 weeks even that that that did the 1 that led to the, you know, this this, like, doom debate analysis, which led to me going on doom debates and whatever. Right? So that was pretty funny.[1][2]

Tim Scarfe 00:00:49

Yeah. So it was your fault that I didn't get to bed until about 03:00 in the morning the other day because I wanted to watch the doom debate. And I thought you did a very good job. I mean, you know, obviously, the comments were really positive. I think everyone went from hating Keith Duggard to thinking that Keith was intellectually honest and not such a bad guy after all. So, no, I thought it was really good. And props to you, Liron. I, I'm very impressed with how you handled yourself. The whole situation was great. Unfortunately, I did get the impression that both of you left feeling that you'd won the debate. What what do you think? Well,

Keith Duggar 00:01:23

so I wasn't I mean, I actually wasn't in it to to win that that debate. It's more like I'm always and nobody ever believes me, but but my goal is always the pursuit of, you know, knowledge. Right? And and or sharing and and whatnot. So I think I think we think we made a lot of progress in the sense that he agrees and and hopefully most people agree now. You know, there's there's a technical sense. Like, there's a very technical sense in which LLMs and again, like, you gotta be careful about how you define that, you know, and whatever. Not LLM systems, not not o 1, not but I mean, just an LLM, you know, sort of a auto re autoregressive fixed context window type thing. I think there's a sense in which people understand that's not Turing complete, like, in this in this technical sense. Right? There's still these open questions or folks aren't that aren't as convinced as I am, let's say. They're not as convinced as I am that that actually has practical con-

[1] Keith Duggar on Doom Debates — video
[2] Is o1 Reasoning? (MLST Philosophical Steakhouse 1) — video

sequences. Like, I think you're in that same that same camp. Right? Which is that, okay. Fine. There's this technical point in which in which that's correct. However, practically speaking, you know, LLMs can generate generate Python programs that you could run-in a system if you wanted to. They do have, like, a very large, context window. And so if they're trained right, they can do a lot of important things. And so I understand that that's still an open question, and I have more work to do to convince folks, that there is a practical consequence. So that's fine. Yeah. Like, I mean, I'm happy with that.

Tim Scarfe 00:03:11

It's certainly true that if you look at the statistical distribution of programs on GitHub, it will have a power law distribution, which is to say most of the programs will be short, low in complexity, and you distinguished the Linux operating system as being a really complex program. And any complex program could be decomposed into lots of smaller programs. But we should also distinguish the difference between something like a Python program and a Turing machine program. Now there's this famous programming language called brain fuck, and that actually describes Are the we actually, are we allowed to say that on YouTube? I wonder. I think as long as you don't use the f word within the first 7 seconds, then you don't get dinged for advertising. Oh, okay. Okay. Alright. Yeah. And because machine learning models don't really understand, they won't know that we used it legitimately. Oh, I got you. I got you. Yeah. That's that's Yeah. The point. We didn't use it in an offensive way, but no, there there is there is a difference between something like a Python program and a Turing machine. So a Turing machine is this hypothetical physical machine that can, you know, it's got a it's got a program. And of course the program is fixed because when we talk about language models being able to accumulate information in their context, the analogy breaks a little bit with the Turing machine. In a Turing machine, it's a finite state automata, and it's fixed, and then it can write to an infinite memory on on both sides.

Keith Duggar 00:04:44

To a potentially yeah. It's to an unbounded potentially infinite memory, not an actually infinite memory, folks. Like, nobody's saying that that exists. Like no. You're right. But go on there. What distinguish the difference between

Tim Scarfe 00:04:58

potentially infinite and infinite.

Keith Duggar 00:05:02

Well, potentially infinite is is a, let's say, a procedure that can that can keep going in principle for, you know, forever, but it may not ever I mean, you may never get to the end. Right? So for example, time, like the flow of time in the universe may or may not, you know, it may be potentially infinite. Like, the universe will just keep going on and on and on, but it will never reach the end. Right? There will never be an actually infinite, time span. And so that's the difference. Like, for example, if somebody were to tell you that, that they had a register in a computer that was, arbitrary precision, you know, then what that would mean is that this thing could, as necessary, expand to to, have more digits more digits of precision. Whereas if they told you, no. I have a register that's, you know, actually infinite. That's a very different kind of thing. Like, that's that's a thing that in a finite amount of time can do an infinite amount of work, right, versus saying that this is something that over an infinite amount of time, it can do an infinite amount of work. So that's really kind of kind of the point is do you like, in a single time slice, do you have an actual infinity, you know, sitting there? And we don't know how to build that. Like, we don't know how to build an actually infinite, you know, register or but we do know how to build things that we can keep adding to and adding to.

Potentially Infinite vs Actually Infinite Memory

Tim Scarfe 00:06:37

So I felt that 1 confusion coming from Leron was the difference between the neural network weights being a finite state automaton and the input being the finite state automaton. Would you agree with that?

Keith Duggar 00:06:53

Yeah. I mean, there's this and there's this really subtle distinction, because as I've said so many times, and I hate saying it every time I say it because it doesn't do justice to to Turing, and is that a Turing machine is just it's nothing but a finite state automata with an unbounded read write tape. And so you have to think of that unbounded read write tape as like an extension, really, of its state space. Right? Because it can go and modify it and read it again. So it's basically, you know, it's it can grow its state space to as large as necessary to solve the problem. And that's why that's why it can explore, you know, an unbounded set of computations, set of states, like, because it can keep expanding. Right? And I think there's there's all and it's become more apparent to me over these last months, right, debating this issue is that there's always been almost like this

kind of ambiguity or sort of unfortunate choice of language in computer science where where they refer to machines as both a model for a physical thing that you can build as well as a notional mathematical model that maps to that as well as a class of algorithms. Right? And so that leads to some confusion. And I actually hope to clear that up in the future or to present it in a way that makes more sense. But, again, I think people have agreed. Okay. There's a technical sense in which this is right, but doesn't make any practical difference. Right? And I would 100% admit I haven't communicated enough or clearly enough, my view on why it practically matters, And I could be wrong. You know? I don't I don't as yet have I have strong opinions about it, but, at the same time, it's it's not as clear cut as you know, it's not as clear cut whether or not an LLM based system. An LLM plus some Turing machines plus some, you know, hash table it can read, write to or whatever. The the capabilities of that system are much less clear. You know? Okay. But another source of confusion is, you know, Liron asked you about the human brain,

Tim Scarfe 00:09:16

and he said it's a finite machine. And you said, yes, you accepted that the brain is a finite state automaton, but it runs an algorithm, which is a Turing machine controller. We can effectively expand our our memory, maybe even internally, but let's just use the the external version of that. So then you're getting into this kind of quite ambiguous zone where you're saying in the class of Turing machine, sorry, in the class of finite state automaton algorithms, there are 1 kind which can expand its memory, and then there's another kind which can't. And you're saying the types of algorithms which are learnable on, let's say, a language model, a bog standard language model, are not the kinds that can expand their memory.

Keith Duggar 00:10:03

Yeah. And so here's this ambiguity I was talking about because, in computer science, you would not call an algorithm. So if you have an algorithm which is a pushdown automata algorithm or a Turing complete algorithm, you would never call it a finite state automata algorithm even though the controllers for all of those machines are finite state automata. Right? So so it's this funny, you know, it's this funny ambiguity where, like, yeah, at the end of the day, all these machines are a controller, which is a finite state automata, having access to, you know, no memory in the case of a finite state automata or a stack

Algorithms, FSA Controllers, and the Training Problem

Keith Duggar 00:10:50

or, you know, read write tape or whatever. Right? And so, yeah, there there's an ambiguity there. And what I'm saying is in the space of all algorithms, there are algorithms that utilize external memory and there are those that don't. And that currently when we train LLMs, they are trained in an environment where, they don't have potentially infinite read write memory. They oftentimes have potentially infinite write write memory, like output. Right? But not read write memory.

Tim Scarfe 00:11:27

Okay. I I got that. So what you're trying to do is imagine there's a diagram and you have finite state automata and you have Turing machines. And you're saying that rather than partitioning the space of algorithms neatly between those, clearly we can take an example of a Turing machine algorithm, because a Turing machine has this code book, which is the finite state automata. Mhmm. And you're saying in the context Yep. Of a Turing In the context of a Turing machine, we call that algorithm a Turing machine algorithm. But all of the algorithms that run on finite state automatas, so basic language models, you you call those a finite state automata.

Keith Duggar 00:12:08

Yeah. I mean, that's that's my claim. Right? My claim is that they're that they're the algorithms that we end up with when we train, these machines where they're set up to run as finite state automata, we end up training finite state automata, you know, algorithms. Right? And I think and as I've said, I think that there's there's more work I have to do there to convince people of that fact, right, which is that it isn't just it's not just that they're in construction finite state automata because they have these fixed windows. Right? They have these fixed context windows. But what I'm saying is that in addition, in my opinion, the algorithms that we access when we train those things come from the space of finite state automata algorithms rather than from the state of push down automata, you know, Turing machines, etcetera. Okay. And that's what that's kind extra work I have to do. Right?

Tim Scarfe 00:13:20

Okay. But to be clear, you're saying that the substrate, so a a native language model can, in principle it's a similar thing to Eugen Schmidhuber was saying, in principle, an RNN can be, trained to be a Turing machine, but you can't train it with stochastic gradient descent. And he wasn't just talking about the infinite precision thing. I think he was talking about some other examples for solving parity where they can be something more than what they are if they are trained with stochastic gradient descent. Because the thing that confuses me is look at how ChatGPT is trained now. It's trained to do rag natively. In fact, I interviewed Patrick Lewis. He's the guy who coined the term rag, a deep expert in rag. And he was telling me that's how they do it at Cohere. When they train their models, it goes into a special RAG mode, so it kind of tightens up because there's loads of things you need to optimize when you're training RAG. And they are training these models en masse to expand their memory through an external store. So, and when I say expand their memory, let me be clear. These things are not expanding their physical memory, but they are bringing into their context things from the outside, and they are potentially putting things out of their context to the outside. Now rag is a special case because I think it's only bringing stuff in. It's typically not bringing stuff out, although there's no reason why it couldn't do that in principle. So I'm I'm asking you the question, how do you know that training models in this way wouldn't allow them to get into that space of algorithms, which you would call Turing complete.

Keith Duggar 00:14:54

Okay. So and and and I'm gonna kinda skip by because I I know we don't wanna get bogged down, like, in the whole the whole debate. So I'll try to skip past some things and just say, look. There's a lot of work that needs to be done to kinda clean up, the language here. You know? So I'll I'll just say this, which is, as far as I know, the way in which RAG is trained never has a a closed loop. Like, it doesn't both write data to the database and read data that it's written to the database. It may do a retrieval of some data, you know, Internet search or whatever and kind of run through that. Right? Auto autoregressively sort of just, like, pass through that as part of its context and build up some hidden state and, like, do something, but it doesn't close that loop. So it doesn't both read and write to the same unbounded store. But as I've said, like, so many times, if you if you alter your architecture to where you do have that, like you do have this closed loop of the ability to read and write to an unbounded store, then you have a Turing machine, and then you had the potential of training or finding algorithms, let's say, from that larger space of, Turing machine algorithms. But as far as I know, like anybody who's attempted to train such systems, fails. And I mean, because

you you run into, like, really basic problems. So for example, let's suppose I'm trying to train such a machine. Alright? Like, I've set it up. It can read and write to a database, and I start and I start running the algorithm. Now in order to do training, right, I have to run it until it gives me an answer. And if I want to be a Turing machine, I need to allow for it to run for some unbounded amount of time, some unknown amount of time. Right? Like, it has to generate a stop token. It has to say I'm done now. Here's the answer. Okay. So I go to train it. I start running the thing. And, you know, I'm waiting, and it doesn't stop. And I'm like, alright. It's, you know, been 5 minutes here and it hasn't completed yet on this, this sample, this batch. Right? Like, what do I do? Do I let it keep running, or do I prematurely terminate it and throw that run out? What do I do? So this is the problem is the Turing algorithms can run for an unknown amount of time. They may never halt. So you're immediately faced with a very difficult problem during training. Right?

Tim Scarfe 00:17:23

Yeah. I'm glad you said that. So the problem is during training. So you could, the whole stopping

Can We Train on Turing Machine Programs Directly?

Tim Scarfe 00:17:30

token thing, that's an artifact of RLHF. So we teach them to stop. We teach them to prefer a certain verbosity and length and so on. Certainly when you just train language models just on the the raw distribution of text, you can just you can just run them forever. And they never say stop.

Keith Duggar 00:17:50

Well, I mean, in order for it to halting, stopping is a basic a basic requirement for for algorithms. So for a Turing machine to to to do what's called decide a problem, it has to halt and give an answer. It has to halt and either say yes or no. Right? And so there are many mechanisms by which you can indicate that the computation is done. You can have a special you can have special states that you enter. Right? Like, if I enter the halting state or halting state 6 or whatever, it doesn't matter. Like, then then I'm done. Or you can write some special thing to the tape. You can you can say, like, you know, here's the answer, or the answer is between a and b. You know? Open paren, close paren, whatever. But you gotta have some way of indicating that the computation is is done somehow. And and here's the problem during training is exactly what I said. How like, if it hasn't

halted yet, what do I do?

Tim Scarfe 00:18:53

Yeah. Okay. But I think we're kind of conflating 2 things. So on the 1 hand, we could talk about today's language models that are trained on internet data, and we can hypothetically imagine that there's some kind of transfer between that manifold and the manifold of Turing machine programs. Another thing we might do is let's just throw it all away and start again. Right? So why don't we create a massive dataset of Turing machine algorithms? Some of them will halt and some of them will not. Now, I believe there's some interesting structure of Turing machine algorithms. It's fascinating to think about what that structure is. And what we might do is just heuristically run a whole bunch of those sampled programs, because we could build a generator or maybe we can be a bit smarter about it. We can create these little units of Turing machine components that crop up all over the place, because I think that there is it's it's like Francois Cholet said with the kaleidoscope hypothesis, right? That there's all of these priors of the universe, and then they additively compose together to produce the kaleidoscope, and intelligence is reversing back from that kaleidoscope into the priors, went into it. There must be something similar in Turing space, and this is related to Chomsky's ideas about nativism and so on. There's probably like a nice set of primitives that could compose together in an interesting way, and we could generate a bunch of Turing machines based on those primitives. And some of them would halt and some of them wouldn't, and that'd be a very interesting structure there. And maybe we already run them, so we know if they halt or not, and we build a large language model on it and we get this manifold of Turing machine programs. And then maybe we try and map that manifold to the manifold that already exists. But that I think would be an interesting thing to do.

Keith Duggar 00:20:41

I mean, I agree. It would be it would be very interesting. And this is where this is where this falls in the category of what I've been saying for a long time, which is we don't yet know how to train from a very rich space of Turing complete algorithms. And so maybe the proposal you just made will be a pathway towards that. Right? Maybe that's not even maybe we don't even need to do that much human human engineering. You know, maybe if somebody just finds some clever tricks, some some improvements to SGD or some alternate, you know, hybrids of of that plus neural evolution or or something like that. They'll they'll find ways to search the space. Right? And I totally believe that that's possible. I'm just saying, empirically, so far, and it's kind of an empirical question. Empirically, so far, you know, people have failed, and it's a very fascinating and, you know, interesting result. Like, why? Why is it so hard? There may actually be some deeper truths there, and I don't know. It's it is a fascinating question.

Tim Scarfe 00:21:55

The space of Turing machine programs, of course, is is infinite. Right? Or maybe you would say potentially infinite.

Keith Duggar 00:22:02

Countably countably infinite. Countably infinite. Yeah.

Tim Scarfe 00:22:07

Yeah. Because I guess the broader problem in AI is overcoming ambiguity and searching. Because it seems that we wouldn't need to sample all of those Turing machine programs to come up with some statistically salient algorithm. But maybe that's not the case. What what do you think?

Keith Duggar 00:22:28

I think it's a big mystery.

Hybrid Systems, Scaffolding, and Novel Programming

Keith Duggar 00:22:30

I think it's a big mystery. So and and here's where I may be, like, woefully out of date with a computer science knowledge. So if I am, I apologize. And, actually, if anybody out there can can educate me on on this, I would I would love to know more, which is as far as I know the last time the last time I looked into this, we know so very little about the space of of algorithms. Right? Of of the Turing Turing complete algorithms. We just don't know a lot about them. I think there's sort of a few a few facts that are known like, you know, most most algorithms, never halt or or, you know, or or this this type of thing. We know that it's countably infinite, that type of stuff. But I think in terms of the structure of the space of all algorithms, either a lot a lot isn't known or or I'm not up to date on the lot that is known, but I think it's it's fascinating. I I think there could be Yeah. Lots of interesting structure there, especially if if, like, you say, you know, you mentioned this this this language earlier, b f. I'm just gonna call it, the b f language. Can spell it out if they want to. And it's it's like the simplest kind of mapping from a Turing machine to a programming language. You know, it's got, like, characters for move left, move right, right something, kind of open a a block, close a block, and, like, that's about it. You know, not not much more than that. That's why it's called b f because if you ever tried to write a program in it, you'd you'd wanna quit really soon and go go find something else to do. But maybe maybe and, of course, we have our own. We have our higher level languages. Right? But the curious thing about the vast majority of programming languages that we've developed is that if you just generate strings in that space, most of them are not valid programs. Right? Like, you have to kind of follow some type of a grammar in order to generate valid programs. And even if you follow that grammar, the vast majority of programs are, like, probably pretty useless. They're like bugs or they just, like, don't do anything or they infinitely loop. You know, what if somebody could come up with a and I'd love if if any work's being done on this, please please let me know. But what if somebody could come up with a language where almost every single program you generate is a useful program of some kind?

Tim Scarfe 00:24:59

Yeah. That would be cool. A couple of things on that. Do you remember when we spoke with Daniel Roberts from MIT? And Daniel, I'm sorry we haven't released your show yet. We we will as soon as possible. Whenever there's a lag releasing a show, it's because of love and care on on my part because I wanna do a good job of it. But he was talking about quantum field theory. You are bit of perfectionist sometimes. I am. It's the curse of perfectionism. But we're not gonna do that on Keith and Duggar's philosophical steakhouse. We're just gonna stick them out. But no, he was saying, you know, in quantum field theory that they also have this ridiculous blowup of particle interactions and so on, the way they reduce the complexity are these kind of heuristics like sparsity and locality and, symmetry and whatnot. And, you know, as you say, we could just enumerate all of the possible combinations of the space. Most of that wouldn't be valid Turing machine programs, but, you know, also there are many, many abstractions on top to, you know, write a generating function of useful Turing machine programs. But then I'm asking the question, well, we keep coming back to this on MLST that there's no true Scotsman, that we need to go to the fundamental level of computation. Wolfrum talks about the Rulliad and these kind of graph cellular automatons. We're talking about Turing machine space, but we also are guilty of talking about the same kind of thing with agency. Oh, you need to have the low level particle dynamics. You need to have the the real ontological particles. There's no point having some higher abstraction. Because if you look at what Charley is doing and the folks advocating for discrete program search, they're kind of saying, well, forget about the Turing machines. Why don't we start with something that we believe to be fundamental? And those are Spelke's cognitive priors, this core knowledge. So if you start with the premise that there is core knowledge in the universe, maybe not the universe, but our our minds. I'm I'm not sure if that's a distinction which matters in this case. But if you start from some structure and you write a generating function over that structure, which is what Chole advocates in the ARC challenge, would that be good enough?

Keith Duggar 00:27:01

I mean, this is there there's so much I have so much, that's so appealing to me. That approach is so appealing. Right? And I think this is pretty similar to what WALL EADSUBBA, you know, tells us too. Right? Which is that there are these universal templates and and, you know, if if we've got enough of those and if they're they're well specified enough, then yeah, you can have these grammars that generate super useful algorithms. So definitely, I think there's a lot of appeal there. I know that I know that it's criticized a lot and people, you know, say, well, it hasn't worked or or this project failed or that project failed. But I I do kind of feel that those may be a very interesting path forward at some point. And, actually, I think, if I were to guess if I were to guess where we'll end up some long time from now, like 50 years from now or more, is we'll have hybrid systems. Right? So there'll be components that we've trained, constructed, generated, whatever it is through those mechanisms. There'll be components that are handcrafted by humans, and then there will be machine learned components that we've learned how to put together in an interoperative way. Right? That'll probably really be the future of of, artificial intelligence if I had to guess. You know, like, the pendulum will swing. Eventually, people will, like, it will kinda plateau in what we can do with stochastic gradient descent and ever larger massive, you know, training data centers. And then, you know, the the sort of DreamCoder on program synthesis and art challenge and SPELCI and and template type approaches. We'll start to produce more and more, and then we'll figure out we still need a little bit of human ingenuity to make sure they're all working together well. So, yeah, I'm I'm big fan of those approaches.

Tim Scarfe 00:29:03

Yeah. And it gets back to the functionalism discussion. We'll we'll touch back on this again when we talk about cell, but there's always this notion of the universe is hideously inefficient. You know, it's created these beautiful kind of recursive minds that we have. And we have all of the priors that help us understand the universe and how this happened. I mean, we can talk about the evolution of language and cognition and nativism and so on, but it's very, very mysterious, but we know it was a very kind of long drawn out inefficient process and, you know, evolution had a lot to do with it. But then there's the question of, okay, well, now maybe we could design something which does the same thing. Right? And and that's kind of what functionalism is talking about. It's saying if if we could design a system which performed the same abstract functions as the thing that we did, and it and it has the same behavior, then it might be isomorphic. Now, of course, someone like John Searle would say, well, there's an ontological difference. There's a difference in the the causal structure. Even actually in functionalism, there's huge debates about what kinds of causal structures you need to

maintain. Like Chalmers, for example, he says that, you know, there are certain counterfactual trajectories and you need to have a very rich causal structure in order to create a good enough simulacrum of the real phenomenon. But the point is though that I I think I believe this, that there are probably uncountably many other configurations of systems which would be as if intelligent. And it's it's it's kind of like the no true Scotsman argument to kind of say, well, that one's not really intelligent or that 1 is intelligent when in fact, in basically every situation, they behave in the same way.

Keith Duggar 00:30:51

Yeah. And and this is kind of what happened in our Discord. Right? Because fresh off the fresh off the the episode that we did on whether or not o 1 could reason, which which led to all kinds of debate and and, abuse on YouTube and, you know, and Leron's response and me going over there. In parallel to that in our Discord, somehow, and I forget how it was. I forget how what the origin was. We pick it back and trace it. We got into this huge, you know, I don't wanna call it a flame war, but this huge internal MLST Discord debate about John Searle and the Chinese room

The Chinese Room Argument

Keith Duggar 00:31:30

argument because I mentioned I mentioned the Chinese room argument. Right? And and 1 thing that happened, like, early on in that debate is is people accusing slash mischaracterizing Searle is saying that he denies that a machine can understand. And in fact, that's the total opposite of what he said. He you know, he he's been saying this whole time, right, that the only thing that can understand are machines. Right? And that we're a machine, and the current and currently, the only examples of things that can understand are biological machines. Okay? And but that doesn't mean you couldn't build a different kind of machine. Like, his only claim is that it has to be a machine. It has to be it has to have a realization. It has to have, like I mean, I I think the correct term too is embodiment. Right? Like, it has to have a physical realized embodiment or not. Maybe not. Let's just call it a realization. We'll get into embodiment later. And that realization has to have some sufficient causal structure. Right? As you as you said, like an internal causal structure. He'd he'd as far as I know, he doesn't exactly specify, like, what that internal causal structure is or should be or or whatever. He's just saying that it has to be there. So you can't just have, like, an ethereal algorithm that's kind of running ethereally on on some substrate that doesn't have that causal structure.

You've gotta have this realized causal structure. And and I agree with that. And I also agree with you and Tim that there can be many different forms, or many different substrates upon which we could build that that causal structure. Yeah. Maybe and and maybe there's more than 1 causal structure. It's like, okay. This causal structure is sufficient for understanding. This alternative causal structure is sufficient for understanding. There may be a whole bunch. Right? And and there may be a whole bunch of different physical ways in which you can realize that causal causal structure, but it does have to be realized.[3][4]

Tim Scarfe 00:33:44

So John Searle is quite possibly the most misunderstood person or philosopher on on the planet. I think the Chinese room argument is probably misunderstood by about 99% of people on online. You've pointed out a few of the the confusions, you know, some people say that he is saying only biology as a machine substrate can give rise to consciousness and therefore understanding. And he doesn't say that, as you said, he says it's the only existence proof. And I'm not a huge fan of the actual Chinese room argument itself. I think it was a product of the time and he was talking about semantics. So the broader theme is he's talking about certain patterns of physical causation. And that's, you know, many theories of semantics are basically talking about certain patterns of physical causation. So there are things that happen in the world and all of these chains of causation end up materializing, well, yeah, not materializing, but end up being in in your mind. And those those links form semantics and and how you understand the world. And he's pointing out this Chinese room argument as an example of the deterioration or or the, you know, how those links can be broken. Now, our friend Jez in the Discord, server, he said, well, Searle's argument is not a theory of semantics. So what's the purpose of this? It's not telling us anything about semantics. I wanna learn about causal theories of semantics. And Searle is just giving an example of why he thinks this particular situation doesn't have semantics. And then I think he pointed to Nagel's Bats, you know, that wonderful essay from Thomas Nagel. And of course, that's not really a theory of consciousness either. That's a very similar thing. It's just talking about in a particular inconceivable situation, how we wouldn't be able to understand what another conscious experience is, but it's not really talking about what consciousness is. The other thing that John Searle is misrepresented on is people call him a dualist, and he's not a dualist. Right. He's a physicalist. He thinks there's 1 ontological substrate, and maybe we'll get onto that in a minute. But it's, yeah. I I think he's misrepresented by a lot of people.

[3] Minds, Brains, and Programs — paper
[4] Chinese Room Argument — concept

Keith Duggar 00:36:01

Yeah. I think, I think I'd rank him up there right next to to Chomsky in terms of being misunderstood and and misrepresented. And, you know, you wonder sometimes about the sociological phenomenon that that caused that that caused that to happen. But I think, yeah. And and Jez, I love Jez. Having him in the in in our Discord is a real, you know, privilege for us because because he's Yeah. Just a precise thinker and and seems to know, like, some ungodly amount of, philosophy knowledge. And and I on the 1 hand, okay, I I I finally was able to to see his point. I don't think it's fair. I see his point, which is his point's kinda like, okay. There's the Chinese room argument, but it has all these, you know, problems and it doesn't really provide a theory and, like, you know, it's it's almost like he's saying it's a teaser, but he and and therefore, it's not really worth that much because it's just a teaser for where, like, where the real work, you know, needs to be done. But but I think you have to understand that a lot of times I mean, that philosophical article and by the way, Tim and I are big fans of, the Stanford Encyclopedia of Philosophy. Right? And even it says, and many philosophers that are that, let's say, oppose, maybe the the resulting intuitions of the Chinese room argument admit that it's 1 of the most important arguments, like, ever made in philosophy. Right? I think there was a joke where where somebody and I'd have to look up the reference, but it was hilarious where he said, you know, basically, the field of cognitive science can be defined as the field which tries to prove the Chinese room argument false. You know? So so I mean Yeah. It's an existential crisis. Yeah. And so I don't I don't think it's fair to just ignore that. Like, it's it's a very good argument because, or there's a sense in which it's a good argument because if you tell it to a person, it allows them to kind of immediately and intuitively grasp that there's a question we need to answer. Right? And a lot of times I've heard it said that that philosophy is really 1 of the biggest tools we have for deciding what questions do we need to ask. Where are the unknowns? Like, where are the holes that we need to poke at? You know, the threads that we need to pull. Right? And so the Chinese room argument makes it viscerally clear that there's something we need to we need to think about here. And, actually, we should we should say what the Chinese room argument is. Do you mind? Because I'm sure we have a lot of listeners who may not may not know it. So just and I'll and I'll just, you know, be brief about it, which is, you know, John Schirle says, imagine imagine there's this, you know, room. Okay? And inside the room is a guy with a bunch of books and and and, you know, rule books and and dictionaries and sort of all kinds of things. Okay? That that allow them to map sentences given to them in Chinese, okay, into answers or other sentences given in Chinese. So for example, a Chinese person can walk up to this room and it's got a little squawk box or a little, you know, thing that can slip a a note under the door with with some statement in Chinese or a question, and he'll do this procedure, this algorithmic

procedure to produce a result, and he'll give it back. In this room, imagine that it's just, you know, it's it passes the Turing test or whatever. Like, people slip notes under there. They get them back, and and they think that these are coming from, you know, from a a person who really understands Chinese. And his point is it's obvious that this room doesn't understand Chinese. It's a guy who has no clue anything about Chinese. He's just performing this sort of mechanical algorithm. If somebody slips under the door or the crack there a joke, okay, he's gonna translate it into, like, you know, some answer like, That's really funny. He's not gonna be laughing at all. He's gonna have no understanding that that was a joke. Right? And so this argument Yeah. Is immediately obvious to people, which is, yes, something's missing here. Something is missing in terms of understanding. And that setup, that Chinese room setup has been used to explore many other, you know, cognitive, cognitive, aspects like consciousness, intention, intentionality, etcetera. Right? I mean, I think it's a great argument for bringing about that intuition that there is a question that needs to be answered.

Tim Scarfe 00:40:37

We'll get back to the, decomposition of kind of cognitive functions in a bit, because that's very interesting. Because in the original paper, he was talking mostly about intentionality, but when you Understanding. Read his later work The original pen was all about understanding. Yeah. Yeah. Oh, I think it it was intentionality and understanding, if I remember correctly. But we can we can go back and have a look in a minute. But but my point is is that you can actually, you can swap those words out for things like agency and consciousness. Certainly, there's a wonderful talk which we'll reference now on the screen, which is John Sole talking at Google about maybe 12 or 15 years ago. And I do recommend that you folks at home watch that talk if you actually wanna understand what John Searle is talking about because he's he's super duper clear in it. And 1 thing that struck me in that talk actually was is very, very instructive. He said he knew he knew that his dog was conscious, not because of its behavior, but because of its mechanism. It had eyes and it had a mouth. Right? And and it was a mechanism much like his mechanism. And, of course, the broader point of is that it's physical mechanisms which give rise to patterns of physical causation, which give rise to our consciousness. Because to Keith's 0.1 minute ago, the reason why Searle in the room didn't understand was because there's this phenomenal, which is to say there's this conscious experience possibly of of laughter. And if you break the, you know, that causal patterning, so if if the conscious component is is not there, then of course, you wouldn't have the complete understanding. Therefore, you you wouldn't understand. So it's very interesting to kind of distinguish this kind of function and behavior from the mechanism. Because when we spoke with Murray Shanahan, he said, well, I can feel

Searle on Physical Realization and Causal Structure

Tim Scarfe 00:42:30

if something is conscious if I share, a physical environment with them and the behavior is kind of intelligible in a certain way, then I might intuit that that thing has consciousness. And of course, this is quite slippery because then you could start talking about a large language model as if it had consciousness based on based on its behavior. So professor J. Mark Bishop, he wrote a very interesting I I think he was the editor of of a a review piece on the Chinese room argument. So he's been studying the Chinese room argument for about 30 years. And we had him on the show a few times. I must admit, I think the first time I had him on the show, I didn't fully grasp this thing because I think this is a thing that takes years of kind of reflection to actually understand. And, certainly, I think by the second or third time I had him on the show, I I was starting to grok what he was talking about. But his opening line, I think, the second or third time he came on was this. He said, indeed, it's my belief that computation cannot realize understanding deep mathematical insight or consciousness, and hence, that there exists a gap, a humanity gap between that which can be achieved by computational processes alone and that which can be achieved by humans potentially using computational tools. 3 a priory philosophical arguments, John Searle's Chinese Room, Lucas Penrose's argument, and his own minor contribution to the debate, the dancing with pixies reductio, and we'll talk about that in a second, underpin this claim. He said that if any of these arguments hold, then an unbridgeable ontological gap between computation and mind is firmly established. Well, aside from a diet of too much dodgy sci fi, he said he believed a deep confusion between epistemic concerns, for example, how we as humans might establish if a sister or brother had understood our words or was feeling pleasure or pain, etcetera. And ontological, for example, how we might actually instantiate consciousness, pleasure, or pain in a machine, lie at the roots of this widespread quasi religious belief that computation can actually realize cognition. He said he suggested that any commitment to what the American philosopher John Sell famously termed strong AI, the view that suitably programmed computers can actually understand natural language or generally instantiate any other mental capabilities of the humans whose behavior they mimic is quasi religious. And he went on to talk about, you know, computational processes can never make us milk or produce energy from sunlight or sadly make us rich. So, yeah. Beautiful stuff.

Keith Duggar 00:45:14

And I I agree with you, by the way, that I think the first time we spoke to, professor Jay Mark Bishop, maybe I wasn't prepared either to really fully ingest, what he had to say. So it's nice to go back and read it with a new a new lens and understand things understand things better. I think, and this is a this is a problem with with the language too because, Searle and I know this is gonna sound weird. I know this is gonna sound really weird to us as, like, computer people. Okay? But he distinguishes between compute and machine. Alright? So when he's talking about a computer, what he's talking about is a kind of machine that doesn't have that sufficient causal structure, okay, to produce understanding or consciousness or or whatever. Okay? He is absolutely not saying that there isn't a machine that we could build from some nonbiological stuff that could have that causal structure. In fact, like, Jay Mark Bishop's statement is maybe even slightly stronger because he's he is really focusing there on the on the human. Right? And saying and you could almost read into that that we'll never be able to, produce that in some non nonhuman, you know, machine. I think we're definitely far away from being able to do that. But but Searle is saying, no. We can. Like, we can we can build a machine that can do that. It just has to have this necessary causal structure. It has to have this realization with this necessary causal structure. And, both of the you know, both Searle and and I think even almost kinda, let's say, more beautifully there, J. Mark Bishop are getting at something which is really fascinating to me, which is, you know, on the 1 hand, we have these beautiful equations and abstractions which did which do really well at mapping to and explaining and predicting stuff that happens in the physical world. Right? This is the the remarkable effectiveness or the unreasonable effectiveness of mathematics, if you will. And yet, on the other hand, we know, like, we know in principle and in fact that these abstractions and equations do not capture the whole story. You know, a mathematical model of a black hole is not a black hole. A weather simulation is not weather. Okay? A computational process simulating the human mind is not the human mind. There's this gap. There's this gap between our abstractions, between our computations, between our math and physical realization. And it's fascinating. And I think this is where it's really largely unknown, and there's so much work to do there to better understand that. But I think you, you know, you have to accept that there's this gap. I mean, the arguments for that are just too strong.

Tim Scarfe 00:48:15

Yes. So the 2 major components to and by the way, I agree with you that Bishop probably is actually taking a stronger position than than Searle did, but we'll we'll come back to Bishop in a second. But yeah. So Searle is saying that we need to have physical realization in order to maintain With the the necessary causal structure. With the necessary causal structure. Yep. Precisely that. So the the the core of it is there needs to be a necessary causal structure. And he's saying, I believe that some kind of physical realization is necessary for that. There's another component to his discussion, has led to accusations of him being a dualist. And that is basically this, that he thinks that the correct physical instantiation and therefore the correct physical cause or structure gives rise to consciousness, and he's a materialist. So he only thinks there's 1 physical substrate, but he he makes the case that, this this consciousness is subjective. Now, of course, we can refer back to Murray Shanahan's view on this. This is capital s subjective, but not in a dualistic way. So it simply means that you, as an observer, you do not have privileged access to what is in my consciousness. And that's part of the reason why he thinks that when we kind of simulate, cognitive processes in machines, because they wouldn't have access to this special subjective experience, they would not have the full fidelity of understanding. Now, don't think Saul explicitly said this, but I think he's basically making the argument that our consciousness strongly emerges from a certain physical instantiation and causal structure. Maybe you can explain what that means.

Keith Duggar 00:50:08

Yeah. And this is this is a super controversial super controversial topic here, you know, strong strong emergence. You know, is there such a thing? Does it exist? So I'll put my cards on the table and say that, I believe that both weak emergence and strong emergence, are are real phenomenon. So let's let's talk about what those mean. So in general, what emergent means is that, you know, if if you have a bunch of parts, okay, like a bunch of molecules, for example, and let's suppose these molecules follow, you know, laws, like the laws of statistical mechanics, for example, and you and you put them together that it's possible for the whole, the sum of enough parts to exhibit behavior dynamics,

Emergence, Wolfram, and Computational Subjectivity

Keith Duggar 00:51:00

if you will, that are just fundamentally different, from what the individual parts do. And strong emergence would say it's possible for them to exhibit behavior that is not even derivable from the behavior of the parts. Okay? Now, like, physicists, I I I don't know what I haven't done a survey, so I don't know. But, you know, if you you talk to most of them, they're they're really strongly or the ones who are vocal, let's say online, are very strongly reductionist in the sense that they're like, nope. You know, the standard model, all, like, all I need is a standard model, and I'll be able to, like, compute everything and derive everything. Except they're really they really can't. They can't do it in any remotely practical sense whatsoever. Okay? Like, poor quantum mechanics guys, like, they can't even really compute past what? Helium? You know, when they when they try to go beyond, like, too many too many particles, they just get totally slammed by computational problems. And they'll come in and say, yeah. Yeah. Okay. Fine. Like, none of my computers can do it. But in principle, I could. You know, in principle, could. And the fact is, even in principle, they're like all the theories we have, all the effective theories I've ever seen, they at least have points where they break down. Okay? They have singularities that crop up when you try to apply, like, you know, the integrations or the effective theories. And as soon as you have a singularity, what comes out the other side is is undefined. So in my view, I think it's true that there are both computational barriers as well as in principle analytical barriers that we cannot cross. We can't get from this side of parts to that side of of the emergent behavior. Right? I think that this is possible at least. So if you take something like, you know, a a good example is, you know, fluid mechanics, like the the behaviors of vortices and waves and the and the the giant, you know, red spot on Jupiter and and things like that, and then you keep going up and up to life itself, you know, life as an emergent phenomenon, or gliders in the in the game of life, these type of, you know, little things that that stay stable and move around. And it's like what Friston talks about, things that have these Markov boundaries and these these sort of stable, dynamic attractors that that form like a thing, you know, that can kind of persist persist over time. All those things are definitely weakly emergent. Some of them may even be strongly emergent. And a lot of people who are even opposed to to strong emergence in general when faced with consciousness, like, you know, maybe accept that, okay. Fine. That's the only strongly emergent behavior. Like, I think we Chalmers said this to us. Right? He said, like, yeah. I'm not really sure about all these other emergencies, but but consciousness may be may be the only example of a strongly emergent behavior.

Tim Scarfe 00:54:08

Yes. Quite an interesting analogy with with Searle. I mean, of course, they have a completely different ontology. I mean, Chalmers has a has has an information ontology. I mean, the the weird thing about Searle is he's actually quite no nonsense. I mean, if you actually take the time to understand what the hell he's talking about, it's it's surprisingly no nonsense. And you said something very interesting there, which I think we should say for our agency discussion, but I believe that agency is also physically realized. And there's a distinct difference between the type of agency we have in in the real world and and simulated, but maybe we can save that. But the other thing you said, is there are many boundaries of subjectivity, and this is very important for building AI systems. Wolfram is hinting at it when he talks about, how we are computationally limited as observers. This argument from Searle is another great example of that. And this is all just pointing to how it might not be possible in principle to simulate many, many things, in a computer. I wanted to touch on, Bishop's dancing with Pixies reductio before we move on. So that was that was his main contribution. So[5][6][7][8]

Keith Duggar 00:55:17

that's 1. Go go on. Can jump in with 2 can I just jump in with 2 things? So in or actually not in preparation for this, but as a result of that debate in our Discord, I watched that that Google talk, Searle's Google, you know, Google talk, which I hadn't watched before. And I agree with you. There there's this this element of no nonsense pragmatism to him because I didn't know the story he told at the beginning where how he got down this whole rabbit hole in the first place is, you know, I guess he was maybe leading a research program at some point long ago, and he kinda walked over to the, like, let's say the, you know, the cognitive science team or whatever. And he's like, hey, guys. What's up with this consciousness thing? How come you guys haven't solved this problem already? Right? And they're like, yeah. No. No. No. No. We're not really working on that because it's not science. And and that kinda set them off on this path. Like, wait a second. Why, you know, why isn't this science? Like, what are you talking about? And and I want to also comment on, Stephen Wolfram. I really and what's the phrase he uses? It's a computationally bound. Right? Isn't that the Yeah. The terminology? It's absurd theory. Computationally limited computation. I think it's such a cool, framing of it. It's it's really interesting, you know, saying how any time obviously, let's suppose for the sake of argument that well, certainly, the universe is manifesting a computation. You

[5] J. Mark Bishop on MLST — video
[6] Dancing with Pixies — paper
[7] Artificial Intelligence Is Stupid and Causal Reasoning Will Not Fix It — paper
[8] Searle Google Talk — video

know, it's unfolding as a computation. I even think it's a hypercomputation. But let's suppose it's not a hypercomputation. Let's just say it's some Planck scale or sub Planck scale, you know, pixel simulation, like, The point is it's this I mean, you can't even fathom how complex that computation is. It's obviously so many orders of magnitude. I'm not even sure higher than the computation that's happening in your brain. And and it's fascinating to me this point that Wolfram makes that so many of the things that we the abstractions we create, the the laws that we find, the science that we do, the behavior that we observe is is due to this fact that we're we have this much, much less computation than the phenomenon that's unfolding, you know, before us. There's it's like they're compression artifacts. Like, in a sense in a sense, all our scientific theories are compression artifacts. I mean, to me, it's just fascinating. And and, you know, that's why I'm always excited to see what he puts out. Though every time every time Stephen Wolfram publishes something, like some article or blog post, by the time I get to the end of it, because it's really massive, I get to the end and I and I'm reminded of, like, how little I've accomplished compared to him. Like, how does he produce all this stuff? I don't know.[9][10]

Tim Scarfe 00:58:04

Yeah. That reminds me. There's a really good article by, Stephen Wolfram on recursion. And the last third of the article, he's, citing, Godeleschebach and and Hofstadter. And I think there was actually a formulation of this in Godeleschbach, like a simple recursive function and studying how it could give rise to these structures. It's actually kind of similar in principle to how a cellular automaton works. And then, Wolfram did a whole load of different examples of this and he kind of saw there's this branching factor and sometimes you get cycles back and sometimes you get these amazingly beautiful patterns and then they weirdly stop for no apparent reason. And there's this whole space of computation, which is fascinating. I don't want to spoil the surprise on that because Wolfram is actually going to come on and we're going to discuss that with him. But it also, has echoes of our fascination with recursivity and the recursive mind. I think there's a lot to be said for that. And also Keith, what do you think about this concept of computational subjectivity, given that we are Turing machines? If you think we are Turing machines, then why in principle would there be computational subjectivity?

Keith Duggar 00:59:18

Well, at the moment, I have a very simplistic view on that. And and again, for anybody listening, you should always take my simplistic views as as potentially

[9] Nestedly Recursive Functions — paper
[10] Godel, Escher, Bach: An Eternal Golden Braid — book

just ignorance. Right? Like, maybe I haven't thought deeply enough about something, and so I've reduced it. You know, I've done the reduction thing that I sometimes rail against. I've reduced it to something simple for now. It's it's almost like a placeholder for me. It's like, okay. I got this simple thing for now. Maybe I'll dive into it dive into it later. Okay? So so for for me at the moment, I kind of agree with at least what I think is a fair representation of something Carl Friston has said in, you know, a a blog post or or somewhere, which is that, let's suppose you're a comp let's suppose you're a machine. Okay? Let's suppose you're a machine and you're running a computation. Alright? And you start off without any sensors or any any data whatsoever on the computation you're performing itself. Like, you don't know how much memory you're consuming or or what the temperature of your CPU is or whether or not a particular loop has been running for a long time. Like, you just don't have any kind of telemetry on your own computation. If you start hooking that up, okay, like you start having data and knowledge about yourself and the computation that you are performing right now, right, then when you get to a certain what Friston calls temporal and counterfactual depth, okay, once you have enough once you have data from your own internal activity processing causation that spans enough time, okay, both in the past and the future, like your ability to kind of forecast out what you're going to do. And once you have enough counterfactual depth, once you're able to think about alternate paths and alternate paths that you could have done in the past, that that's you know, once you get far enough along that depth, you become self aware, you you gain, you know, consciousness, and maybe some other, you know, cognitive aspects.

Tim Scarfe 01:01:28

Yeah. We should probably save some of this for our agency discussion, but we did have a bit of a slightly different interpretation on that I remember because I felt that he was talking about the depth of nesting of the Markov blankets, which give rise to a kind of counterfactual depth of planning. So his thing is that you are more of an agent when it's as if you are planning more steps into the future, so your planning horizon increases.

Keith Duggar 01:01:55

And when that happens Can we ask him? Could we cause you know, so fortunate to have, yeah, I mean, we're so fortunate to have him as a We should email him. We should maybe email him, see if we can record like 5, 10 minutes of him explaining this to us. That would be amazing. Yes.

Tim Scarfe 01:02:10

But even that, I guess, is a form, if you think of it computationally, is is some form of of recursivity. But he's saying that there's some kind of goalie of saying depth.

Keith Duggar 01:02:20

Yeah. Well, well, I mean, the the depth so the depth can be nested. So, I mean, I see what you're saying that there so I don't know if the nesting if he thinks the nesting is a necessary component, but, certainly, it could be it could be nested. You know, if you have enough if you have enough depth and and and I guess in some sense too, it it almost has to be because, you know, since since your mind you know, 1 of the most efficient ways in which you can do depth of computation is recursion. And so it probably is nested, but it would be really interesting to get his take on that and the degree to which he thinks nesting is a necessary component, versus just an optional, you know, an optional or sufficient component.

Tim Scarfe 01:03:05

We need to get Friston back on. I think what we need to do, Keith, is I want you here. We we need to get you in a room with Friston. Are you okay? Oh, I would love to meet We've got a we've got a the thing is that, I mean, maybe I I can just go in and film at his place

Friston, Self-Awareness, and the Path to Consciousness

Tim Scarfe 01:03:20

and we can do it over the internet or something. But I would love for us all just to be in a room sipping a as you say, and discussing the meaning Yes. Of the So but anyway, we'll we'll we'll save that for our agency discussion. But there was 1 objection from Cog Champ on our Discord, seem to recall, and this was quite an interesting objection. He was saying, yeah. Yeah. We need the physical causal structure. And he was saying it's like an oxymoron to say that computers don't have causal structure. So, you know, we have all of these algorithm. I mean, there are books written about this, you know, like that the algorithms are affecting us in in many ways. So they are somewhat, you know, maybe not agential, but but they do have causal power on us. So if there is a space of causal power on us, why could we not shape that space with the right algorithm?

Keith Duggar 01:04:14

Well, I think it I mean, I think maybe their cog champ and and Searle and possibly us were talking past each other because Searle is saying it has to have a certain causal structure, not just that any old causal structure will will do the job. Right? And so, sure, computers, you know, a computer is a kind of machine. It does have a causal structure. When it's performing a computation, there's obviously a causal structure there. Otherwise, the computation wouldn't work. Right? Like, you know, the the transistor charges cause this, the clock cycle causes that. I mean, so there's a causal structure. I think, what Stroll is saying is that you have to have the right kind of causal structure. There's a particular causal structure, and he doesn't specify this, by the way. And this is kind of the weak point, or as far as I know, he doesn't specify this. This was the weak point that Jeremy and Cogchamp, you know, Jez and Cogchamp point out. Right? Which is like, okay. There's a causal structure there. Like, why isn't you know? Or he hasn't told us what the causal structure is. Totally fair. I mean, it's Searle's claim that there's a certain kind of causal structure. And so I think it would be highly unlikely, highly improbable that the causal structure that we build into a computer into a digital computer for purposes of, you know, carrying out a calculation. I find it, like, pretty pretty astronomical that that would just happen to also coincide to the causal structure necessary for understanding, intentionality, consciousness, etcetera. And this is part of what is brought out in the Chinese room argument. Like, people intuitively know that a bunch of gears that you're sitting there, like, just churning to to punch out, like, you know, the next conclusion to the sentence. Like intuitively, we think it's missing something. That's the whole point. Right? It's missing the right causal structure.

Tim Scarfe 01:06:08

Okay. So I think a fair interpretation then is so isn't completely ruling it out. He he thinks that there is something quite special about physical, real mechanisms. Right? And that is likely to give this rich causal graph that couldn't, know, that's not likely to be replicated with any algorithm on on the internet or or anything like that. Because it does bring us to to the next adjacent point, which is, there are, you know, linguists like Steven Piantadosi, and Steven's written many papers saying that Chomsky was wrong and that language models do learn semantics. And he's got this theory, which is called concept role semantics, and Jez is very interested in this. And, you know, essentially, it's saying that there is some kind of causal structure, but it doesn't actually need to be as physically materialized as we think. And there are internalist and externalist versions of this because there's always the problem with having an externalist causal semantics, which is to say, well, I can imagine abstract concepts. I can imagine unicorns. I can imagine things that don't necessarily have any analog in the real world. And of course you could go the next step and you could say, well, maybe they are composed of primitives, which do exist in the causal world. But there is a whole school of thought that in a computer, if we just represent this causal semantics graph to high enough fidelity, then the understanding threshold will be met.

Keith Duggar 01:07:42

Honestly, I'd at this moment, I'm not prepared to rule that out. Like, I I don't know that I can say that's not possible for understanding. Okay? For for understanding. Yeah. I can't, I mean, I can't rule that out at at this moment. I think, you know, there has to be a certain causal structure. Whether or not we can get to that causal structure by virtue of just a special kind of program, like a complex enough kind of program could be possible. And then if and then if it even goes as far as to, consciousness, well, then we're in a, you know, interesting situation. Right? Because it's like, okay. I've got this computation now, which has somehow met all the causal criteria for for, consciousness, and I suspend the calculation. And so it's just sitting there stored, you know, on a on a on a disk drive now. There's a bunch of, like, bits represent maybe I've been g zipped it. So I've g zipped this file. Like, that file definitely no longer retains the causal structure. Right? Because it's been compressed, and it's not running. Right? But I can rehydrate it as as now a conscious, you know, conscious being. So, that'll definitely open up a bunch of interesting territory. I mean, I think clearly, definitely, Bishop, I think, is arguing against this possibility if I understand him, you know, correctly. I think Searle too, you know, probably thinks this is not not the case. Right? I don't know. I think it's Okay. That's a tough 1.

Tim Scarfe 01:09:41

Another thing is I don't think Searle explicitly has a position on internalism or externalism, but I think people assume that he's an internalist because he talks about the brain. I think he believes that the brain alone, you know, so you have the brain in a vat, and because of the special physical mechanism in the brain,

Causal Structures, Concept Role Semantics, and Understanding in AI

Tim Scarfe 01:10:00

that brain would be conscious, which means there'd be some subjective component. But based on what he said in the Google talk, I think he is, I guess this would be technically embodiment, but there are many different brands of, or, you know, flavors of externalism, are pointing to there being cognizing elements outside of the brain and maybe the body and so on. He said that, you know, his dog has eyes and a mouth, and that contributed to this mechanism of it being a conscious being. So I think it's reasonable to assume that he's at least open to the idea of there being like an external component to cognition.

Keith Duggar 01:10:37

Yeah. I mean, maybe That would be an interesting question as well. I think I think my reading of it, and this is just my reading of it, is that he does require a certain internal causal structure. Like, without that internal causal structure, you're not gonna have consciousness or understanding. And then, like you said, I think, maybe he doesn't require externalities, external, you know, causation, but that but that there can be a part of it and or they can be a part in of of our ability to know or to infer that something has, you know, consciousness. But for example, I would bet just from having read a lot of them and looked at, you know, his videos and whatnot, that if you were to put the question to him, does a person, let's suppose a person isn't necessarily in a coma, but they're somehow totally shut shut off from the outside world, but their brain activity is still is still normal. You know, it's like a dream type of state. Right? I think you would still argue that that person understands whatever language they learned natively, that they're that they're even conscious, that, you know, despite the fact that they currently have no external connections really, whatsoever.

Tim Scarfe 01:11:58

So J. Mark Bishop, his contribution was this dancing with Pixies reductio. And, and this I'm just reading a summary of it here, but it's a critique of computational theories of consciousness that builds on Hilary Putnam's ideas. So the key points are, 1, it targets claims that a suitably programmed computer can instantiate general phenomenal consciousness or qualia purely in virtue of executing a program. 2, it argues that for any computational system claimed to have phenomenal states due to its program execution, we can map those computational states onto, state transitions in any open physical system, like a simple counting machine or even a rock. And 3, this mapping can be done for a specific execution trace of the program with defined inputs over a finite time interval. 4, therefore, if the original computational system has phenomenal experiences due to its computational states, then so must the mapped physical system, which leads to an absurd, he says, form of panpsychism where consciousness would be ubiquitous in the physical system, little pixies dancing everywhere. The argument aims to show that computational functional organization alone is not sufficient for phenomenal consciousness. Now, is it such a reductio ad absurdum to say that if we map a computational process to an open physical system, then the open physical system would be consciousness. And that's an absurd thing. Do you find that absurd? Yeah.

Keith Duggar 01:13:32

So, I mean, Putnam's Rock is sort of you know, the the original version of this. And, yeah, I think it is absurd to say that a rock is conscious. But I think so there's a couple little, I don't know, loopholes

Tim Scarfe 01:13:47

or, you know Not not not the rock. The the the overall,

Keith Duggar 01:13:52

you know, organization, would that be conscious? Well, but this is specifically about, I mean, the rock or in the case of dancing with pixies, an open an open system. Right?

Tim Scarfe 01:14:05

Well, so so he's saying take a cognitive process in your brain and represent that as a series of, like, pipes and water and paper cups and and whatnot. You know? Just just sort of map it to some open physical system. So it's performing the same computation. Then let's imagine that that thing in the physical world that we've just mapped it to is also conscious. Does that seem absurd to you?

Keith Duggar 01:14:36

Well, I just wanna clarify here. You can do that same mapping to a rock. That was the whole point of, you know, Putnam's Rock is that you could even map it. Just forget about, you know, nice pipes and and, hoses and gears and things that, you know, maybe are cool. Like, you can even just map it to, like, the this this sort of atomic dynamics of a sufficiently large rock, right, to some subset of its dynamics. And, yeah, I think that's absurd. Like, it's absurd to think that that the rock has consciousness, and I think it's, like, somewhat, absurd absurd to think too as well that, you know, a bunch of gears and, like, whatever of of an arbitrary structure that we mapped it to necessarily have consciousness. Right? We are Fristonians.

Tim Scarfe 01:15:24

Right? We're we're Fristonians. So so Friston says that there exists certain dynamics and functional organization, which give rise to consciousness. So if we if we say yes to that, then why do we say no to mapping a cognitive process to an open physical system?

Keith Duggar 01:15:45

Well, I mean, so I'll tell you why. Like, I'll tell you where the I'll tell you where kind of the the, problem boundaries are in in that argument, which is that 1 1 thing I really like about the dances with dances with pixies, reductio ad absurdum, is that it does lay out very precisely the argument. And in fact, lays it out so precisely that its own weak points are explicitly put in there. Okay? Which is it says, if we if we pay careful attention to it, that for any finite period of time, you can do this mapping. Okay? So, like, you can choose some finite amount of time, a day, a year, or like a whatever.

Consciousness, Computation, and Philosophical Puzzles

Keith Duggar 01:16:30

Right? In in principle, can do this mapping. Here's the problem is that in whatever thing you choose and you do this mapping, suppose it's an hour. Okay? An hour plus Epsilon. Okay? An hour plus 1 second. Those 2 systems will have diverged unless they have exactly the same causal structure. And this is kind of the key of what, you know, Searle's pointing out is like, sure. I can take you know, let's suppose let's let's boil it down to, the absolute vacuum state of this problem, which is a Boltzmann brain in a vacuum and a Putnam, a large Putnam rock in a vacuum. Okay? And so if we all say, yeah, the Boltzmann brain has consciousness, you know, it's conscious, it understands, it's, you know, whatever else. Okay? And then there's the Putnam rock in a vacuum over here. And if the Putnam rock's large enough, I can map all the dynamics of the Boltzmann brain for any finite period of time to some subset of dynamics in the rock, the rock is, you know, conscious. No. Because outside of that domain, epsilon time beyond that, they diverge. Okay? So the computations are not are not the same. The dynamics are not the same. The causal structure is not the same. And that's the whole core of of Searle's argument. Right? And it's and it's the whole it's the sword that cuts through this idea that, like, just the just the observable functional functional, you know, functional behavior is enough. Like, you have to have the causal structure, built into the realization. But yeah, I think it is absurd to think the rock is conscious.

Tim Scarfe 01:18:18

Yeah. I mean, I was just gonna say that, yeah, I intuitively agree that, you know, it is reductio ad absurdum because you could just take a trace of cognitive system and then it would diverge at Epsilon 1 or like 0 And 0 then there are all of these intellectual exercises of the fidelity of maintaining the causal structure. So what if we have a, Chalmers did this with the combinatorial state automata, you know, so we model the counterfactuals and we model some underlying process and eventually maybe there'd be some threshold level where we would capture the phenomenon. And that's very interesting philosophical territory. But certainly, if you just capture a trace, then then, of course, it wouldn't it wouldn't be the same thing.

Keith Duggar 01:19:07

Yeah. I mean, I agree. I think it's fascinating. It's interesting. I think these are very deep questions. I think they're and that was the power of the Chinese room argument is that at least laid out this this in question in a very visceral way. You know, you you hear the problem and you're like, yeah. There's something we need to think about here. And so it's it's definitely helped to spawn a great amount of research. Yeah. All fascinating questions. You know, I wonder how many we'll see answered in our lifetime. Yeah.

Tim Scarfe 01:19:40

I don't know. Final thought. What do you what do you think about Hinton winning the Nobel Prize? Look. I

Keith Duggar 01:19:48

Drawing for breath after I asked that question. Yeah. Because I I don't I'm not I don't think I'm qualified to to have an opinion, honestly. I don't think I'm I don't really think I'm qualified to have an opinion because I haven't I don't know, you know, I haven't surveyed enough the whole literature and everybody's contributions to it and and and whatever. So I just don't think I'm qualified actually to have an opinion. I did see I did see a lot of, let's say, some of the controversy around that. And and the 1 thing I do know is true is that academia is plagued by politics. Okay? It's plagued by politics and corruption and pettiness and gatekeeping and and, you know, power struggles and turf fights and bad behavior and people not siding when they should and stealing work. I mean, so it's absolutely a case that academia has a lot of problems. Okay? Whether or not any of those are relevant at all for for this this choice of noble, I couldn't say. I can't say. And so therefore, like, I just I take I take the award at face value, which is congratulations to them. I mean, they've obviously had huge contributions to to the field of of machine learning, and, you know, I I would love to see any of the the folks that we talked to on the show or interacted with. It's kinda like, let's say, the the godfathers of of AI. I think they're all massively deserving of of accolades. So it's all like I can't disagree with that.

Tim Scarfe 01:21:39

His his

The Nobel Prize for Deep Learning

Tim Scarfe 01:21:40

contributions are undeniable. But as for the credit assignment thing, no need to comment on that because we're gonna release part 2 of Jurgen Schmidhuber. And he was pretty scathing, actually. So you can only imagine what he said. So I'd I'd I I won't comment. I'll just let let you folks wait until we release that. But the other controversy was it was a Nobel Prize for Physics. And this is kind of related to the whole Chomsky comments about, well, language models aren't a theory of, you know, they're not a scientific theory. They don't tell you about what the thing is and what it isn't. Yeah. And to to what extent is deep learning, like what what has it got to do with physics, if if anything?[11]

Keith Duggar 01:22:24

So, again, I don't I don't feel that I'm qualified to decide what should and shouldn't count as a, a physics Nobel Prize. I certainly know so we interviewed Daniel Roberts and Sho Iida. Right? And they have this beautiful book that I really liked a lot, the principles of deep learning theory, which in that in that case, it's the reverse. They're applying physics, you know, ideas to the study of deep learning. And it struck me as totally legitimate, useful, something we should do more of. So I can't say it's impossible that that machine learning work can't can't actually be physics or or consider physics or contributing to physics. I just I don't think I'm qualified for this. These are these are way above my pay grade to to decide what is and isn't, you know, counted as physics, honestly.

Tim Scarfe 01:23:21

Yeah. But do you accept the analogy to Chomsky's comments on language models, which is that deep learning models are not a scientific theory in the sense that they don't provide a powerful explanation? Well, but Chomsky was talking about, you know, a specific I mean, what Chomsky was saying is that if you use all these techniques, deep learning or whatever, and then you train a language model, it doesn't tell you anything about

[11] Principles of Deep Learning Theory — book

Keith Duggar 01:23:49

language. It may tell you something about physics though, like, because because the physics was in the in the training process, was in the math, was in the engineering, and the science that went into the process that trained it. So it's not the it's not like somebody it's not like Hinton went and trained a, a large physics model. Right? Like a large quantum mechanics model, and then he got the award for physics. It was it was more that the award came for the work that went into the the techniques and the math and the science behind creating these models and training them.

Tim Scarfe 01:24:24

Yeah. But the the naysayers have said that the the techniques that go into deep learning from a physics point of view are quite, don't I wanna say trivial, but, you know, they're they're they're not they're not that complicated. And indeed, people physicists have applied their knowledge to deep learning. But does deep learning give us knowledge about how the universe works? Because I I think that's what the whole pursuit of physics is about.

Keith Duggar 01:24:51

Well, on that particular score, like that question, like, I'd have to say not in my opinion right now. I mean, I don't I'm not aware of any, you know, fantastic insights that I I guess the only 1 I can think of is is I think there's something very interesting about the fact that hyper overparameterized models are able to generalize it all and that they don't that they don't just overfit into absurdity. Right? I think that's really interesting. And maybe that says something deep about, I don't know, the structure of maybe that helps us explain the unreasonable effectiveness of mathematics, for example. But, again, I don't feel qualified to I don't I don't feel qualified to to comment on this specific, award. I think I think there are too few new there are too few Nobel prizes to go around. Every year, there are many more people deserving of a Nobel Prize than there are Nobel Prizes to give out, and so very difficult decisions have to be made. And I just don't feel qualified to second guess them.

Tim Scarfe 01:26:11

Another couple of angles. And, yeah, I mean, like, we should just both say that we are that, you know, we're allowed to have opinions even if we're not an authority on every opinion. But, another comment was made that Nobel Prizes are given out almost speculatively based on the potential of a discovery, because of course, Demis Hassabis and a couple of guys got the Nobel prize for the, I think it was the drug discovery thing. And this is certainly something which has huge potential. But the naysayers are saying that this is kind of validating a lot of the bad things. I mean, apparently Google in a lawsuit, even this week have started citing the Nobel Prize. They say 1 of our VPs and top scientists, he just got the Nobel Prize. So what are you talking about that there's copyright problems? And, I think Hinton is also starting to kind of double down and say, well, look, I told you so. In 10 years' time, these things are gonna be doing everything. And that anything you think that humans are needed for now, they're gonna be doing it. And it's lending credence to this idea that there's just no end in sight for this technology. Let me say as well that we all agree that this technology is revolutionary. It's changing the world. It's changing everything. People just disagree on how that is. Some people think it might recursively self improve and its capabilities will just keep going up and it's really intelligent. And and other people just worry about the the social impacts of it and reducing our agency and exacerbating inequality and and so on and so forth. But no one's gonna deny that this is a a a real revolutionary technology.

Keith Duggar 01:27:49

Yeah. And and the stuff you mentioned, I'll come back to that. But what you mentioned earlier, like I said, there's politics aplenty politics aplenty in in academia, and and I cannot say in this specific instance that politics and corruption and other bad behavior played a role. Okay? But it certainly exists in academia. But as I've said before, that's part of the reason why I'm not in academia because I really had a very difficult you know, unfortunately, like, I'm kind of genetically a boy scout or something, and I, I just had difficulty dealing with corruption that was wasting public money. Okay? Like, public money grants and things like that. So I'm out of there. Right? I mean so, yeah, I can totally believe that it's being abused certainly by corporations to to to get more money and power. Absolutely. As for, like, the disagreements about AI, yeah, you and I, we've obviously, we know that these are powerful technologies. A lot of what we talk about though is that, 1, there are limitations to these technologies, and we're suggesting ways in which they can be expanded and further improved, and powerful tools can be misused. And and I was happy, you know and and earlier, I

AI Harm, Ethics, and Public Policy

Keith Duggar 01:29:05

didn't thank, Leron too. I I would like to because it was a great opportunity to go on doom debates. And I thought the the feedback was tremendously positive. We got a lot of positive feedback of you know, funny enough, it's it seems like it helped to humanize Duggar. So I'm not quite as, you know, evil evil as I used to be. But but 1 concept that we came up within that podcast is that we should be talking about a lot more. We should be talking about a lot more AI harm. Okay? Which is the harm that AI does. It brings lots of great benefits. It's also bringing harm, and it's being misused to cause harm. And we have to think about that, and we have to wrestle that, and we have to reduce the harm that we that we accrue from from AI and and increase the good.

Tim Scarfe 01:29:58

I agree with that. I mean, we didn't speak much more about Leron's podcast, but at the beginning, I was surprised that you said your PDOM was 25%. And Yeah. I've always listened to the Sam Harris podcast, and there are lots of folks out there who are really interested in existential risk, not necessarily AI, but pandemics and nuclear war and stuff like that. It's it's very legitimate. I mean, Noam Chomsky talks about it all the time. He always says we're, you know, like 1 tick away from doomsday talking about the threat of nuclear rum again. This is all absolutely true. And and I I guess I maybe I I should be worried. I accept it to be a truth. It's just not something that I'm personally, you know, that preoccupied with. So that's undeniable. And then there's the sort of the intermediate thing of harm from AI, non existential harm, but just social harm. I'm completely on board of that. And then the, you know, if you go all the way down that line, you get to this existential harm from AI. And for me that would require, AI to be agential. And I, for many of the reasons we've spoken about today, I don't believe that that could happen even in well, maybe maybe I'll say in practice just to be safe for now, but certainly not something that we should be worrying about anytime soon.

Keith Duggar 01:31:15

Well, the the the nice thing from my perspective is if we could just get people to really take seriously AI harm. Okay? Just take it seriously and look into it and try to mitigate it and try to prevent it, try to reduce it. A lot of the steps, a lot of the policy steps that we would take to reduce AI harm, would actually reduce AI risk as well. And if we started to take AI risk seriously, it would reduce the chance of AI doom. So these are, like, steps in the right direction. And the reason why I wanna frame it in terms of or reframe it in terms of AI harm is that something that you could get, I don't know, like, the vast majority of people, like, plus percent, you know, to to resonate with? Like, yeah. I mean, absolutely, AI harm is it's it's reducing quality of life by by leading to unemployment or people losing their jobs or, you know, no no longer being able to get paid for the work they wanna do. It's causing massive harm to our children, you know, through, like, the social dilemma type stuff. It's it's sucking up attention into, like, completely pointless exercises like infinite scroll of, you know, content with no mental nutritive value whatsoever. I mean, if we could just get people to almost all of them would resonate with some aspect of AI harm, and then then we could start to take it seriously for policy,

Tim Scarfe 01:32:49

for public policy. The thing that I don't like is the conflation of AI safety and AI ethics. I'm trying to find the paper that Penny shared the other day. Here we go. Yeah. There there was a paper that I read by Dagmar Monet called Deconstructing the AI Myth, Fallacies and Harms of Algorithmification. We'll And put a link to that on the screen. And 1 thing that struck me actually is, and I think it's fair to say that, Dagmar is is on the left. A lot of discussion is around the reduction of our agency as a function of AI. And many different philosophers talk about this. When I interviewed Luciano Floridi, he was talking about reontologization as we migrate to an information ontology. So we're basically living our lives in the infosphere and our digital identity is in the info sphere and we need technology to survive and so on. Perhaps that's more of a consciousness argument, there are all of these arguments that are talking about various different components of our cognition, be it our agency or consciousness or whatever it is, they're being eroded as we acquiesce to technology. And, you know, so that that that's I I see that as a bit of a leftist argument, but I I'm completely on board with that. That really makes sense to me.[12][13]

[12] Deconstructing the AI Myth: Fallacies and Harms of Algorithmification — paper
[13] What Is the Philosophy of Information — paper

Keith Duggar 01:34:07

Well, I I mean, I think there are so many dimensions of of AI harm as well as AI good, that the whole point is we need to be having these conversations. And now a lot of times the conversations become really, like, quite nutty. You know? So I don't I don't resonate with any philosophy whatsoever that doesn't correspond to physical reality. Okay? So, like, I start with physical reality as kind of my my bedrock, and I think that no matter how much we don't like truths, we have to accept them as as truths and then deal with them, you know, head on. Okay? So but I think the nuttiness in conversations would kind of get worn down if if there was a more robust conversation. So we need Yeah. More people having the conversation. We need more voices involved, not fewer. We need real data, real science, real consideration. We can't leave we can't leave the steering of AI to just the fringes of people. We need to find a way to talk about it with with everybody.

Tim Scarfe 01:35:17

I I agree with that. And I I feel that Luciano Floridi in principle is saying that. He's saying as we become increasingly divorced from the physical reality, even from a legal perspective, the territorial reality, these these systems abstract so many of the physical boundaries that define how our legal system works, how how we live our lives and so on. But I I did wanna comment on I wasn't trying to say I wasn't trying to say he was nutty. That's what I'm trying to say. Oh, no. No. I know I I know what you were saying. Voices. Yeah. Yeah. I know. I'm I'm getting I'm getting on to to to the nutty voices you were talking about. So so this this is why I don't like the conflation of AI risk and ethics. You kind of pointed out why on your Leron discussion when you started talking about orthogonality and instrumental convergence. Right? So there are all of these philosophical thought experiments, which are used as justification for AI risk. And I think that they are extremely speculative. Right? It's fascinating I think it's legitimate philosophy, but Yep. I don't think they mix particularly well, certainly in, if you wanna be doing policy, don't think that those arguments should be factoring in any way the same level of valence as talking about the other stuff we're talking about, you know, like the the effects of social media and bias and fairness and reontologization and all of this kind of stuff. I I I I don't wanna use the word science fiction, but I I think it's completely on the speculative side of things.

Keith Duggar 01:36:49

Well, I mean, I completely agree with you there, which is that and I agree with every point that you said. So, yes, there is a whole host of speculative, you know, things. It's really fascinating. Definitely philosophers and and and maybe, you know, chumps like us should be talking about it, okay, and should be engaging with it, okay, a 100%. And you're absolutely right. It's not what should be factoring into the policy. You know? But what I'm trying to get across to to folks who are concerned about AI harm, AI risk, AI doom, is you don't need science fiction and diamond nanobots and, like, whatever else in order to convince vast majorities of people to do things that you would like to have done that would help with those concerns. Right? You you know, there's always this conflict. People think that that, that public policy should just be run by science. They believe in kind of like a technocracy. Okay? And that's not how the world works, and it's and it's actually contrary to a lot of economic principles and and, you know, political principles and whatever else. You've gotta find you know, because you're dealing with this massive evolutionary algorithm, which is the 8,000,000,000 people plus every other life form on Earth in an ecosystem. And what you have to do if you wanna steer that ecosystem in the directions that you want is you have to be able to find arguments that resonate with large fractions of the population. Okay? And so you need to focus on the core and the simpler arguments that will resonate with more people. Let's do that. And if we do that, we'll be able to get better policies that not only will reduce harm, help to mitigate risk, they'll make us more prosperous, healthier, happier, overall, have more resources to figure out how to solve the alignment problems, etcetera, etcetera. So, you know, we've gotta focus on human flourishing, I guess, at the end of the day, which I know, Sam Harris would agree with. Right? Like, we if we focus on human flourishing and reducing harm, we'll get a lot of the policies we need to handle these other concerns.

Tim Scarfe 01:39:10

Easier said than done. On on that note Oh, yeah. Doctor Duggar, we will round off this philosophical say steakhouse episode. Let us know at at home, folks, if you like this kind of thing. Me and Keith just rambling.

Keith Duggar 01:39:24

Makes an interesting false. I love it. Yes. Anyway, cheers. Cheers.

References

[1] **Keith Duggar on Doom Debates**
https://www.youtube.com/watch?v=4v-Qh3JQ4Jc
video

[2] **Is o1 Reasoning? (MLST Philosophical Steakhouse 1)**
https://www.youtube.com/watch?v=nO6sDk6vO0g
video

[3] **On the Measure of Intelligence**
https://arxiv.org/abs/1911.01547
paper

[4] **Minds, Brains, and Programs**
https://home.csulb.edu/~cwallis/382/readings/482/searle.minds.brains.programs.bbs.1980.pdf
paper

[5] **Chinese Room Argument**
https://plato.stanford.edu/entries/chinese-room/
concept

[6] **J. Mark Bishop on MLST**
https://www.youtube.com/watch?v=_KVAzAzO5HU
video

[7] **Dancing with Pixies**
https://philarchive.org/rec/BISDWP-2
paper

[8] **Artificial Intelligence Is Stupid and Causal Reasoning Will Not Fix It**
https://www.frontiersin.org/journals/psychology/articles/10.3389/fpsyg.2020.513474/full
paper

[9] **Searle Google Talk**
https://www.youtube.com/watch?v=rHKwIYsPXLg
video

[10] **Nestedly Recursive Functions**
https://writings.stephenwolfram.com/2024/09/nestedly-recursive-functions/
paper

[11] **Godel, Escher, Bach: An Eternal Golden Braid**
https://www.amazon.co.uk/Godel-Escher-Bach-Eternal-Golden/dp/0465026567
book

[12] **Principles of Deep Learning Theory**
https://www.amazon.com/Principles-Deep-Learning-Theory-Science/dp/1316519333
book

[13] **Deconstructing the AI Myth: Fallacies and Harms of Algorithmification**
https://www.researchgate.net/publication/382802495_Deconstructing_the_AI_Myth_Fallacies_and_Harms_of_Algorithmification
paper

[14] **What Is the Philosophy of Information**
https://www.researchgate.net/publication/225070243_What_Is_the_Philosophy_of_Information
paper