/mlst

Michael Levin - Why Intelligence Isn't Limited To Brains.

March 24, 2026



Machine Learning Street Talk

Podcast Transcript

Revision: 47fe8c3a

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

Future evolution of human intelligence and consciousness 3

Science fiction as a faster Gantt chart 5

What we actually mean by humanity 7

Bow-tie architecture and the intelligence ratchet 10

Agency as empirical discovery, not philosophical position 12

Least action principles and free energy as proto-agency 14

Learning in gene regulatory networks 17

Multiscale competency architecture 19

Embodiment as perception-action loop, not physical body 22

Cancer as a shrinking of the cognitive light cone 24

LLMs, diverse intelligence, and epistemic humility 27

References 32

Overview

Michael Levin runs a lab at Tufts that does things most biologists would call impossible ten years ago. He regrows frog limbs, builds synthetic organisms from skin cells, and trains gene regulatory networks the way you'd train a dog – not by editing their DNA, but by exploiting learning interfaces that nobody knew existed until his group started looking.

This conversation goes deep on what Levin calls "diverse intelligence" – the idea that cognition doesn't start at brains, or even at cells. He walks through experiments showing that networks of just four chemical nodes can do Pavlovian conditioning, habituation, and counting. The implications for medicine are concrete: instead of killing cancer cells or performing gene therapy, you try to reconnect them to the electrical network of the body and let the collective intelligence of the tissue do the repair work. His lab has demonstrated this in animal models.

The broader argument is about humility. We built our entire framework for detecting minds around medium-sized creatures moving at human speed in three-dimensional space. Levin thinks that's parochial. He points to least-action principles in physics, problem-solving in transcriptional space, and the creative confabulation that happens whenever a compressed representation gets re-expanded – from embryonic development to human memory – as evidence that agency is a spectrum, not a binary.

Tim pushes back on several fronts: whether compassion is just as accidental as our biology, whether embodiment requires physical space or just a perception-action loop, and whether LLMs might have more going on than we think. Levin's answer to the last one is characteristically careful – he doesn't think current AI is anything like human minds, but he also found cognitive properties in sorting algorithms, so he's not willing to bet the house on his intuitions about what any complex system can or can't do.

Featuring: Prof. Michael Levin (Tufts University) and Tim Scarfe.

Future evolution of human intelligence and consciousness

Michael Levin 00:00:00

Each 1 of us, and also us as the current implementation of humanity, are going to be replaced. Persistence in current form is impossible. It's impossible in biology. Every species will either die out, or it will change and adapt, in which case it is again not the same species. So the next question is once you've given up the idea that you can stay exactly as you are, the next question is what would you like to be replaced by?[1]

Tim Scarfe 00:00:25

Michael Levin is a distinguished professor at Tufts University. His research spans multiple fields, including developmental biology, bioengineering, and cognitive science. Levin is known for his groundbreaking research on bioelectricity, regeneration, and what he calls diverse intelligence.

Michael Levin 00:00:43

What do we actually mean by humanity? What is the important feature about being a human? Do we mean that you have to have modern homo sapiens DNA? Do you mean that they have to have the kind of standard human organs and a human anatomy and they better not, you know, have pieces replaced by some sort of cyborg technology. I do think that it's possible to create beings that are highly intelligent in the sense of problem solving capacity, but that don't necessarily share these things with us. I think probably that is what we will have first. I mean, that's certainly where I see modern efforts at AI going.

Tim Scarfe 00:01:16

Professor Levin, it it's it's such an honor to have you on MLST. Thank you so much for joining us.

Michael Levin 00:01:21

Thank you so much. Thank you for having me.

[1] Michael Levin — person

Tim Scarfe 00:01:22

So, I've just enjoyed reading your paper, A Bridge Towards Diverse Intelligence and Humanity's Future. Can you tell us about that?

Michael Levin 00:01:30

Yeah. You all basically, what I'm working on is considered part of the field of diverse intelligence. And this is the overall the large project of trying to understand what it means to be an intelligent agent, in particular, an an embodied intelligent agent. And my claim is that we, as humans and as scientists, are really only calibrated for a very particular kind of for for detecting only a very particular kind of intelligent agent. We're good at noticing minds that are embodied medium size scale, medium speed moving at 3 you know, in the three-dimensional space. We're okay at noticing those things, then we have some trouble maybe a whale or an octopus or something like that. But I I actually think that there are many other spaces in which minds are embodied. And this is this is present throughout biology and I think wider, even even wider than what we would consider biology. And I think that we need to develop principled theories for how to detect, how to communicate with, and how to rationally relate to these other minds that are out there,

Tim Scarfe 00:02:35

I think it's a very important project. This is maybe a fun a fun question here. I I get the impression from reading some of your work that you're either a fan of science fiction or you've at least, you know, you you you know enough about it because I mean, is that true? Are you would or is it fair to say you're a fan of science fiction?

Michael Levin 00:02:52

Yes. I think that's fair. I mean, I grew up reading all of the, you know, classic science fiction. I was I'm a huge fan. I mean, honestly, I think that

Science fiction as a faster Gantt chart

Michael Levin 00:03:00

what people call science fiction just means you have a a faster Gantt chart than than everybody else. You know? Right. That's really you you you know, people say, well, this is sci fi. But but but to me, anything that isn't forbidden by the laws of physics, and I'm not sure we know that much about what's actually forbidden, is is possible and probably likely assuming that, you know, we we we don't kill ourselves off and so on. So I I don't see, you know, some of some of these things that science fiction has dealt with, I see as very very practical kind of issues and and fodder for for philosophical and scientific discussion.

Tim Scarfe 00:03:34

Yeah. And I mean, the reason why I ask is I I view science fiction as as really writing where you you have the entire palette of possibility, if you will. And so you can explore essentially any any topic you want to. As you said, you know, as long as it doesn't violate the laws of physics, it's at least soft, maybe hard sci fi even. Right? And for me, it's it's been extremely useful. And I get made fun of a lot on on the show for mentioning it. But like you, I find a lot of inspiration and really just thought process that happens in science fiction. Like, let's explore the idea what if, you know, what if there was an intelligence that was the size of a a cloud as large as the solar system? And it showed up and it didn't even really recognize that there was intelligent life on Earth, you know, until after it had done tremendous damage. You know, what happened? I mean so it's just the ability to explore these ideas. Right? And I think there's a lot of thought provoking thinking that happens in science fiction.

Michael Levin 00:04:31

Yeah. Absolutely. And and it also reminds us to be humble about our choice of perspective. Right? So so what science fiction to us today, you know, every every just about everything that you and I did today in our daily course of life would have been utter science fiction not that long ago. And it would have seemed completely, you know, ridiculous except to a few, you know, there there right. There are a few visionaries that that see some of this stuff early. But but, so it's a moving target. You know, these goalposts are constantly moving. And and I think that's critical is to remember that our current perspective is not privileged in any way, that the things that seem sci fi to us now may well be commonplace, and and we need to think about what that means for us.

Tim Scarfe 00:05:11

Professor Levin, 1 of the reads I got from reading your paper was almost I guess you were railing against human chauvinism. And there were elements of what I don't know whether you would be comfortable with me saying this, but transhumanism. And this idea of a space of possible minds. I mean, Aaron Sloman, of course, wrote about that in 1984, if if I remember correctly. And we spoke with Murray Shanahan about that recently. And it's fascinating. But we as humans, we tend to anthropomorphize. We think that our form of humanity is the only valid 1, and we're constantly making distinctions between us and machines and talking about things like agency and so on from our own lens. And you you spoke to this beautifully. But 1 thing I wanna talk about is losing our humanity. So you said personally, if I and humanity are supplanted by a population of highly intelligent, motivated, creative agents with a huge radii of compassion and meaningful lives who transcend our current limitations in every way, what better outcome could I hope for? What's your take on that?[2]

Michael Levin 00:06:15

Yeah. Well, I I wanted to talk about 2 things there. First, you know, each 1 of us and also us as the current implementation of humanity are going to be replaced. That's there there's no there's no other alternative. Right? Each each 1 of us, we're all going to die. We're going to be replaced by something. So persistence indefinitely. I mean, there are some antiaging things coming down and so on. But but but overall, persistence in current form is impossible. It's impossible in biology. Every species will either die out or it will change and adapt, in which case it is again not the same species. All of us start out life as a single cell. We become an embryo, a child. We learn. We grow. We change. You know, persistence in the current form is impossible. So so the next question is once you've given up the idea that you can stay exactly as you are, the next question is what would you like to be replaced by? And I and I do think that this idea that we should remain as we are is I I I I don't know. May I I I can't wrap my mind around it because if if I think of humanity, you you know, maturing as a species and we come back, let's say, some some 100 years later, are are we still walking around with lower back pain and astigmatism and and and with our you know, with whatever kind of body you happen to have been born into because some cosmic ray hit a, you know, an embryonic cell and all these limitations? It it seems it seems impossible that this is the future that people want to paint. These limitations that that we face today were not given to us by any kind of rational process that had our values or our welfare in mind. This is just where evolution happens to have dropped us off at the point where we can now participate in that creative process. And I think

[2] A Fire Upon the Deep — book

it's inevitable, and I think it's I I think this is what we should want. We should want the future to be better than the past. And we should we should we should be thinking about what is it that we are inevitably going to be replaced by, and what are our values in terms of what kinds of things do we want to populate the universe? So that's that's the first thing. The second thing I would say is about humanity and losing our humanity. So so the first thing to think about is what do we

What we actually mean by humanity

Michael Levin 00:08:15

actually mean by humanity? What is what is the important feature about being a human? I I've heard that there's people are working on proof of humanity certificates in the face of AI generated content and things like this. So that raises an interesting question. When you say some someone is a human, what what do what's actually important about that? And 1 very kind of simple way to visualize it is you're going on a journey to Mars. You're gonna be there for 30 years. Who do you want to have with you? Right? What is so so you'd like a human companion. What does that actually mean? So let's run down some options. Well, we do we mean the DNA? Do we mean that you have to have modern homo sapiens DNA? I don't know. I don't I don't care about DNA that much. Does it does it what do you mean that they have to have the kind of standard human organs and and a human anatomy and they better not be replaced, you know, have pieces replaced by some sort of cyborg technology? I don't know. Don't really care about that either. I think as we as we think carefully about what is it that we do want to have a meaningful humans human level relationship as opposed to a Roomba or a, you know, a a pet snake or something, then then the question is what do you actually want? So I think we want 2 things overall. We want a similar or at least so so when we say human, what I think we mean is at least a human level radius of compassion. That means you can actually in the, you know, sort of in the linear range, you can generate the and respond to a certain level of compassion for others. I think that is a a what what we really mean by humans. And then the other thing I think we mean is being able to share existential the same existential concerns. So so we you know, you want you want to have a relationship with another being that has roughly a kind of impedance match with with with respect to their cognitive light cone. The the the size of goals that they can pursue, right, should be should be roughly similar to yours, and they should be able to understand your existential concerns. So they they, like us, should be made of parts. They should be, you know, what what the what you can think of as a kind of mor-

tal computation, meaning that, you know, they're they're perpetually at risk of disappearing, basically. Right? So so none of us are are invulnerable and so on. So so these kinds of, you know, be open to these questions. What what where did I come from? What does life mean to me? This is what we mean. The the ability to raise those questions and to have those concerns and to have the degree of compassion of a standard human or or up, That is that is what I think we mean.

Tim Scarfe 00:10:46

Yeah. Just a a quick point on that. So this sounds like you're making an argument for goal canalization or it's very similar to Bostrom's instrumental convergence, which is basically that for an intelligent living entity to exist in our physical world, they would probably exhibit things like compassion and existential concerns. And of course, they make the argument they would probably exhibit power seeking. But why do you draw that inference?

Michael Levin 00:11:12

Well, I'm not actually suggesting that they have to. I I I do think that it's possible to create beings that are highly intelligent in the sense of problem solving capacity, but that don't necessarily share these things with us. I I think that's entirely possible. I think probably that that is what we will have first. I mean, that's certainly where I see modern efforts at AI going, which is which is that they're not really bio inspired in the in the relevant way. They are, you know, very good at solving certain kind of problems, but they actually don't have some of the some of the key features that are at at least as far as we know, I think we have to be very careful with with any kind of blanket statement here. But so so I don't think those things necessarily go together. My my claim is simply that the latter is what we should care about. In other words, if you want to maintain humanity, I don't think you should be maintaining physical embodiment. I don't think the anatomical description of a modern human matters for anything. I think what you should be really thinking about is is, you you know, wisdom and compassion, basically. Right? That's that's what you're looking for.

Tim Scarfe 00:12:17

So you've chosen a couple of the properties of what I think if you went around just interviewed humans, you know, what does humanity mean? They they would have a list. Right? And so you've chosen a few, you know, compassion and and wisdom. At the same time, those are, in some ways, just like our biology, they're kind of accidental, right, as to where we are currently in our in our evolutionary state and how we got here. Why should we care about those 2 in particular? Like, you know, maybe maybe the best form of life or future future intelligence has no compassion. Like, why should we care about these 2 in particular that you've selected?

Michael Levin 00:12:53

Yeah. I mean, I think I think we are getting here to very fundamental values that I I you know, it's like they're like axioms. I I don't I don't I I can't give you an argument that those that that the 2 features that I've chosen are better than anybody else's features, but we have to choose something. And I don't think there are any arguments to be made for any of these that are sort of knocked down logical proofs that, okay, this is this is the way that you should go. I think this is where we have to pick pick some pick some axioms and see what the implications are. I mean, the so this is and and and, you know, and again, to be clear, this is my personal view. This isn't I'm not saying that I can prove this or that everybody else should think of it this way or anything else. It's entirely possible that somebody wants to emphasize something else, but I think mine are pretty, you know, pretty reasonable.

Tim Scarfe 00:13:42

Sure. I mean, well 1 thing about that which I find kind of interesting is that a lot of things that we perceive as limitations, if you will, of human cognition, human biology, whatever, if you kind of look more deeply and you look at the, you know, the algorithm, if you will, of evolution and the fact that what it cares about is essentially the long term survival of life overall, not even a particular species, but the entire life ecosystem, if you will, then they start to actually be less limitations and or things that that actually have value. So you bring up this concept of

Bow-tie architecture and the intelligence ratchet

Tim Scarfe 00:14:20

bow tie. Right? Which is like the fact that I have to take my thoughts, compress it into a language, communicate them to you, then you sort of decompress those into, you know, thinking in your head. It seems like, you know, maybe a disadvantage. The fact that we take our genetics, compress it into an egg and a sperm and then they combine and reproduce an individual, there are all these bottlenecks everywhere, these kind of bow ties. Right? Very similar to AI, you know, coder decoder architectures. Right? And most people look at those like, this sucks. You know? If only if only I could just have a neural link to transfer, like, data directly to you. Like, if I could just clone myself instead of going through this whole sexual, you know, reproduction kind of thing. Like, they see limitations where actually a lot of the points that you make in your work are no. These actually have, like, tremendous value when you're a life form in an environment of insufficient knowledge and limited resources. You have to survive under variation, variability. Can you kind of talk about that in 2 perspectives? Like, 1, did I get that correct? And then secondly, why do you think it is that people always perceive these things as limitations? Like, frame is all about their own personal limitations rather than seeing the bigger picture.

Michael Levin 00:15:33

Yeah. Yeah. Well, so so first, you know, to talk about evolution and the goals of evolution and so on, guys should have Richard Watson on. He's a computer scientist from the University of Southampton, and he has a he has a very interesting kinda new take on evolution. So so that's you know, I'll let him I'll let him handle that 1. But I think that I I think you're right. I think people you know, I think in general, and and this is totally like armchair psychology. Right? I I don't really know. But but I suspect that people tend to project a lot. And in a lot of what I see as a resistance to some of the things that I and others have written about, you know, the the continuum between humanity and other kinds of forms of mind and so on. I I I think what a lot of people see there is a projection of their own limitations. You know, when people talk about what is impossible for for for for for machines, and of course, they they say machines as though this was some kind of crisp category when we know we have to deal with cyborgs and hybrids and chimeras of of of all kinds. You know, I I I think that that there's a real undercurrent of limitation in the in the in the sense that compassion and and those kinds of things are a 0 sum game. You know, that that if we that if we recognize too many other minds that we then have to care about, there there won't be enough. There won't be enough for us, and and we may not be able to muster sufficient

sufficient care. And so I have a feeling that a lot of this kind of resistance really does come from a deep you know, and and as I pointed out in this in this paper and another 1 that's in review now, are these are fundamental, very, very ancient human issues, you know, being supplanted by the next generation. This is not a new AI thing. This is as old as the hills where where you know you are making high level intelligences that are going to probably be better than you. You have no idea. I mean, you do your best or not, but but you have no idea that they're going to value any of the things that you value. You know, in fact, through their teenage years, they will tell you that they don't. And and so, you know, and so who are you going to get replaced by? This is this is a very old problem. This is not new with AI. So so I think all of these things are bubbling up to the surface now, but but they've been with us all along. And the thing about the[3]

Tim Scarfe 00:17:45

bow That's that's like that's like Socrates, the kids of today. Right? Yeah. Or the the youth of today.

Michael Levin 00:17:50

Yeah. Exactly. Yeah. And and the thing with a bow tie, so so let me talk about that for a second. So I I I I do think that that this this basic architecture is incredibly important because what I think it does in biology is set up the intelligence ratchet that we see in the living world. And and roughly speaking, works like this. As a as a being that is limited in terms of a finite being under under limited resources and limited time, you don't have you you don't have the luxury of being a Laplacian demon. You can't track microstates. You have to compress your experience. You have to learn and and squeeze lots of particular events into some kind of a a compressed representation. You're throwing away all kinds of correlations and and and you're hoping to, you know, capture a pattern, ignore all kinds of things and and capture a pattern, and that's your that that's how you form memories. Well and this is this is true behaviorally, but it is also true on a developmental evolutionary time scale, as you pointed out, you know, squeezing down to a genetic representation and then re expanding. So the thing about that is that there's an algorithmic process on the left side of that bow tie, which takes what what the experiences you've had and squeezes them down into representation. But the other side is is not purely algorithmic. It's creative. And it has to be because by throwing away all that stuff, you now don't really know what the past actually was. You just have a representation that you then have to interpret. The the necessity to interpret that in real time is what gives rise to not only behavioral intelligence, meaning flex you know, flexible adaptive behavior

[3] Richard Watson — person

because you your your allegiance is to the salience of the information, not the fidelity. You actually have no idea what the original memory was. All you have is your memory trace. And in real time, you have to you have to continuously do this kind of, like, storytelling about yourself and the outside world based whatever engrams you happen to have at the time. Right? Because we don't have access to the past. And the same thing so so that's cognitively, I think, what's happening. And developmentally, what's happening is that biology

Agency as empirical discovery, not philosophical position

Michael Levin 00:20:00

unlike our computer technology, biology is dealing with a very unreliable medium. So we work very hard so that every layer is isolated, right, from the layers above. And when you're writing at a high level computer computer code, you you're not worried that that the metal is gonna act weird and that you're, you know, suddenly your your, you know, registers are gonna the numbers are gonna float off and be something else. You you can trust. Right? You have full trust in in the layers underneath. You know? Biology has the exact opposite approach because everything is unreliable. You have no idea exactly how many copies of anything. You're going to have the molecules come and go. Things degrade. Things misfold. It's it's and and you know on an evolutionary time scale, you know things are gonna change, guaranteed. Not only if the environment's gonna change, but your own parts are gonna change because you're gonna mutate. Everything is different and each time. And for this reason, I think that evolution does not overtrain on the past, and this is why, and I've we we could talk about it, but I have in my papers, show many, many examples of this. Biology is incredibly good at solving novel problems and making some kind of coherent being in all kinds of crazy configurations that either nature has done or or that we do in the lab. And the reason is is because even the normal embryo was not a mechanical construction to begin with. It is solving that problem every single time, you know, by I like this idea of beginner's mind. It's like you're given some tools, you're given some hardware that's encoded by the genetics, you have no idea what it used to mean. You have no idea what kind of environment, you know, and how it was meant to be interpreted, and you have no allegiance to that, and you have no ability to know exactly what that is. All you know is you're going to use that information as best you can right now to do whatever the next thing is. And that creativity, that that the requirement to to confabulate, really, to to tell a new story without any allegiance to what the story used to be is is is the is the intelligence ratchet

that starts off in physiological and and metabolic space and proceeds through anatomical morphospace and eventually behavioral space and linguistic space and so on. I I I think that that's that aspect of it having to compress and then creatively figure out what do my memories mean, I think is is is foundational to being a real agent. And I think that's what biology does.

Tim Scarfe 00:22:07

Yeah. Reading through your work, and of course, we can debate what the definition of intelligence is, but you framed it as some kind of general problem solving ability. We've been debating on the channel for years, so we'll skip that 1 just for the time being. But when you spoke about the metamorphosis from a caterpillar to a butterfly, and also in your work in xenobots and anthrobot, you seem to be making the argument that evolution is bestowing the capacity for intelligence rather than solving specific problems or having a specific morphology. And you demonstrated in your work that you could repurpose low level, you know, cells, if you like, into completely different spaces and they could exhibit intelligent behavior. And to me, that suggests that you're making an argument which is kind of leaning towards primordial agency at the lowest level. I mean, we've spoken to panpsychists like Philip Goff. You're not making an ontological argument. You're just saying that the locus or the lens of agency and intelligence should be at the very lowest level and it's you know, it has incredible plasticity and generalization.[4]

Michael Levin 00:23:07

Yeah. I mean, I I I would like to make 2 arguments. 1 is that I I am not sure there's any truth of the matter in the sense that some kind of objective unique view from nowhere that where you can say, this is this is what what has mind and this is what doesn't. I think that all of this is observer relative. Now for for for complex agents, that observer is the system itself. So there are just the strange loop that can reifies itself in certain organisms. But, basically, all of this is from the perspective of an observer. And so what you can say is from from from my perspective, I look at very basal systems, and we've now done this not only in cells and tissues, but we've done this in molecular pathways. We've done it in very simple computer algorithms, like sorting algorithms. We can talk about that. But but the idea is when you look at a system like this, what set of tools are appropriate? Are they tools of mechanical rewiring? Are they the tools of cybernetics? Are they the tools of behavior science, of psychoanalysis, and and and so on? Right? And I I you know, I'm I'm a panpsychist in the sense that I think that the utility of these kinds of approaches becomes clear, varied low down. In other words,

[4] Philip Goff - Galileo Error — book

you don't need anything remotely like a cell or anything that complex before you have to start thinking about from the perspective of the system itself, what does it learn, what does it measure, what does it remember, what are the valences. I I think all of that stuff starts very low down. But I don't think this is a philosophical position. I think this is an empirical claim. So in other words, someone might say, I don't believe any of that. I I'm looking at the system, and I'm going to use good old physical rewiring. Somebody else says, no. I think this this system has memory and preferences and goals, and they use a certain other set of of tools. And then we all get to find out who, you know, who had more better research programs emerging from that. You know? So I'm not talking about the hard problem of consciousness here or or anything like that. That's that's a little bit of a different kettle of fish. But but in terms but in terms of, you know, cognition and intelligence, which I I'm not defining intelligence that way, but for the purpose of our work, that is the the slice of intelligence that I'm interested in, which is problem solving. I do think there are other aspects to it, but I I don't know of an experimental way to get at it, so I'm sticking with problem solving.

Tim Scarfe 00:25:25

Yeah. I think it might have been better if I used the term panagentialism. I mean, of course, you're you're you're not an ontological pan you know, you don't think that the the bits of the universe have phenomenal experience. But I think you do think that at the lowest level in the biosphere that there is the either the facility or the capacity for agency. And if I understand you correctly, it's this diffusion or collective agency which is the most important lens to understand systems. And another interesting thing about this is, I know we talked about transhumanism earlier, but I do feel that the lens of analysis determines the viewpoint.

Least action principles and free energy as proto-agency

Tim Scarfe 00:26:00

So a lot of social scientists, for example, they look in the non physical world. They they don't pay attention to individual human physical humans. They they look at culture and they look at language, and they diminish the individual agency that we have. Whereas, yourself, for example, you're you're looking at the individual microorganisms in in the biosphere. And there are some intermediate views like Dawkins' view or even Joscha Bach, we spoke with him a couple of weeks ago. He's talking about, cyber animism, which is the embodiment of of information in in into our physical beings. So do you think that lens of analysis changes the

perspective?

Michael Levin 00:26:33

Yeah. Yeah. I I I really think it does. And and I think that I yes. I I think that as a as a matter of empirical discovery, we are finding that using agential lenses on on systems starts very early, you know, way be way below what most people so so I I think intelligence and cognition and agency are a bigger circle in terms of this Venn diagram kind of thing. It's a bigger circle than than life, than what we would call life today. When we estimate the the agency, the intelligence, well, Venni system, we're basically taking an IQ test ourselves. What we're saying is this is what I am smart enough to notice about this system. And I think we're just at the beginning of getting at all good at noticing that. We're just beginning to develop tools to recognize these other minds.

Tim Scarfe 00:27:20

Yes. I mean, I I think, you know, a physicist might kind of challenge that, right, and say, I I have no need for the assumption of agency to explain, like, celestial mechanics or, like, anything else, actually. Like, you know, nowhere in the sort of standard model of physics will you find anything that has to do with what a human being would normally associate with agency. Now you can, like, reduce your concept of agency all the way down to know if, like, a photon and electron interact and they wanted to do that or they had some but, I mean, physicists don't do that. So can you give an example where in physics, like we benefit from taking this perspective of agency to better explain anything that cosmologists, physicists, solid state guys, any bread and butter physicists would care about?

Michael Levin 00:28:06

A couple of things and I'll give you a very specific example. So 1 thing I think that there there needs to be, again, an impedance match between the tools you're using and what you hope to see. So so if you insist on using low agency tools, so rulers, voltmeters, that kind of stuff, you're guaranteed to see that kind of thing. And and and the field of folks who look at that stuff, we call those physicists, and they see certain certain things. Now we have to remember so so my claim is not that human level cognition is is usefully painted onto electrons. That's not what I'm saying. What I'm saying is it's a continuum, and what we can do is we can ask a very simple question. What would the most basic version of agency have to look like? Not the human level version, but, like, stepping it all the way down because, you know, we were unfertilized oocytes once, and so clearly, somehow that gets scaled up during development and during evolution. There's a there's a continuum. So so what does the basement version of it look like? What do you have to have in order to to be on the spectrum at all? And and that's and that's important because if we look at if we look for human sized versions of this at at other levels, of course, we're we're not going to find it, but we have to reduce our expectations concomitantly to going down on that on that continuum. So I actually think if you do that, you find something that physicists absolutely deal with, which is least action principles. So the very basic version of having goals and having some problem solving capacity to meet those goals is the kind of thing you see, for example, when you find out that light chooses paths that are the least action. It doesn't look of course, it doesn't look like human agency. The question is, is there a story of scaling by which you can go from something that's simple like that to concepts that we need in neuroscience? And and so so 1 example of that scaling is free energy minimization. And there are, you know, that that's a that's story in which, for example, Carl Friston and Chris Fields and and colleagues have used the same conceptual framework to talk about what particles do and eventually talk about what what humans and, in fact, ecosystems do. So so what we're looking for here is is stories of of of scaling that allow us to say, okay. This is the very basic, the most primitive version of what through various dynamics of life and and maybe other versions will become eventually the things that we recognize as as human level agency. But I'll give you a very a very specific example.[5]

[5] Free Energy Principle — paper

Learning in gene regulatory networks

Michael Levin 00:30:30

Consider gene regulatory networks. So what you have there is never mind the nucleus, you know, all that machinery, just just the the very simple thing you got. Let's say half a dozen nodes, these are chemicals, that turn other nodes on and off. That's it. You don't have anything else. You don't have a cell. You don't have any you know, all all all all it is is is is half a dozen or or more chemicals that turn other chemicals in that set on and off. Now when you look at that and so let's say we're gonna model it, and then this is this is what we've done. People model them with ordinary differential equations, and you can see exactly what's going to happen when when when these things interact. Now now 1 way to look at this is to say, okay. It's completely deterministic. It's fully transparent. There isn't any any magic here. So this is going to be a dumb mechanical system, and I've got, you know, a dynamical systems theory by which I'm going to just make predictions about what this thing is going to do. Okay. So that's that's 1 set of tools where where where you say, I I don't see anything here that would warrant any kind of, you know, any kind of agency talk or anything like that. And that's fine. And if you do that, you get certain kinds of capabilities. But you're also precluded from other kinds of capabilities. For example, these gene regulatory networks are very important in health and disease, and if you wanted to manipulate their behavior, that viewpoint makes a specific prediction. What it means is that if you're going to change the behavior of this thing, what you have to do is rewire it. You have to introduce new nodes or cut connections between certain nodes and other nodes. And in biomedicine, that's gene therapy. The way you would do that is through gene therapy. That philosophical outlook has very specific implications for your roadmap for regenerative medicine, for example. Okay, and so we looked at that and we said, well, yeah, transparent and it's relatively simple and there isn't any magic. But of course, that's also true if we look inside of human brains. There's no magic there that we're ever gonna find either. So let's do this. Let's not assume we know what this thing is and let's just try things on that spectrum of agency to see what other tools are appropriate. So 1 other tool that you might try is something that's used for brainy systems, and that's training. So taking advantage of a learning interface. How might you train a gene regulatory network? Well, what you might do is you might choose some of the nodes. So let's look at, for example, in the case of Pavlovian conditioning, just as an example. We actually found 6 different kind just to kind of spoil the whole story, we found 6 different kinds of learning in these gene regulatory networks. So let's take Pavlovian conditioning. So you've got a dog, you

normally you ring a bell, the dog responds not at all, you show him some meat, he salivates, then eventually you pair the bell and the meat, and then eventually he'll salivate at the sound of the bell, for example. So okay. So so that so so we have 3 3 things here. We have the con the unconditioned stimulus, that being the meat. We have the response, that being the salivation, and you have something that's initially a neutral stimulus, the bell, that eventually becomes the conditioned stimulus after you've, you know, you've exposed the thing. So so you take your gene regulatory network and you pick some nodes and you say, okay, this node is gonna be my response. I'm just gonna watch what this node does. This node over here, I'm gonna find a node that's an unconditioned stimulus, meaning when I when I trigger this chemical, the response gets gets upregulated. And then I'm gonna find some other node, and I'm gonna try I'm gonna treat that as a neutral stimulus, meaning when you tickle that node, nothing happens to the response. Now I'm going to do paired presentations, and I'm going to see what happens. So if you do so this is a very standard kind of behaviorist approach. It black boxes, you know, you don't have to know is it an animal, is it a gene regulatory network, is it a you know, you don't have to know what it is, you can just kind of do it and see what happens. If you do that, what you find out is that biological networks, but not so much random networks, so that's interesting. I think evolution actually likes this feature. So random networks don't do this, but biological networks have 6 different kinds of memory. They can do habituation. They can do sensitization, they can do associative learning, they can do some other things. They can count to low numbers and things like that. None of that is obvious or visible by taking the sort of chemistry and dynamical systems theory approach, and you wouldn't find it unless you tried it. Why? Because we are not good at guessing these things. Most people's intuitions, when they look at this thing, they would not say that this thing is smart to and and I'm not saying it's, you know, I'm not saying it's human level. I'm not saying it's sitting there having hopes and dreams. What I'm saying is it is on the continuum. It is on the continuum of agency because there are tools that you would apply to learning systems that get you to new capabilities. And so what are those new capabilities? Well, now that you know you can train the thing, you have all sorts of possibilities in terms of drug habituation and and all kinds of biomedical, you know, drug association, all kinds of cool biomedical applications which we are now pursuing in the lab that don't require you to use gene therapy. Because because now you understand that you have a system that has priors, it stores those priors, and future behavior is a is a is an is a trainable outcome of past experiences. So my point is is is is is simply this. We have to be humble about the fact that we don't know how to just, you know, notice these things and these are not you can't have philosophical feelings about it. You have to do experiments and see what that does for you.

Tim Scarfe 00:35:37

Okay. So it's almost like a more sort of holistic approach to to, let's say, the the gene regulatory network, a more holistic approach to trying to solve solve the issue. And the reason why I say that is because no matter how many ODs or PDEs I wrote down with all the, you know, gene regulation, it's still only because of how insanely complex, you know,

Multiscale competency architecture

Tim Scarfe 00:36:00

biology is. It's probably still only a projection of what's actually happening. And so, for example, you may find that, yeah, every time we introduce too much sugar and we brush your hair, like, seems there should be no connection there. But somehow brushing your hair, we've learned, can correlate with, you know, proper insulin, you know, functioning or something. It's entirely possible, right, just because of the insane complexity of the feedback of of your biology, you know, tactile influences, things like that. So it sounds like it could lead to more holistic or at least exploratory rather than very analytic and reductionary kinds of approaches to to treatments. Treatments.

Michael Levin 00:36:39

Well, it's definitely not reductionist, but it's my claim is not that things are so complicated that we should try random stuff because we'll be surprised. Mean, that's true too. No no doubt that's true too. But that isn't my my claim. I mean, the the this this learning capacity shows up in a network that has the minimal network that we found has 4 nodes. And so you really don't need to be that complex for some some of these things to happen. But but also to be clear, the things that we found were not random things that we were, you know, poking and and we just happen to notice it. We took some very specific protocols from behavioral science. And this is what we've been doing for 30 years is is stealing stuff from neuroscience and applying it other places because I think that this is not just the random vagaries of complex systems. I think these are common threads that run through systems which don't even have to be that complex. We found things

Tim Scarfe 00:37:29

Okay. Interesting. So it's kind of a it's it's an interesting middle ground. It's sort of like on the 1 hand, you kind of, like, have some holistic kind of a a woo approach where we don't even have it. We just try different things and we found that crystals correlate with something or other. On the other hand, you have to break it down, the very extremely analytical find the gene, you know, to cut. This is an interesting middle ground closer actually to that traditional, you know, sort of scientific approach, but but guided by a set of principles that came from, like, let's say, a higher higher level of of emergence. So I don't know. I mean, it seems pretty interesting. Yeah. 1 1 1 way to think

Michael Levin 00:38:06

about it is look, humans have been training dogs and horses for thousands of years knowing 0 neuroscience. Right. Exactly. We knew absolutely nothing what was between their ears and and the reason that we we were successful at it was because living systems control they don't control themselves by micromanagement. They encapsulate, right, trainability and and behavior shaping all all throughout. We're just hijacking that that normal interface. So that that learning interface that the system is using encapsulates all that crazy complexity. So I I actually think there's this kind of weird uncanny valley with complexity where, for very simple systems, we can kinda micromanage them and and and we can tell what's going on and we can control them. Then things get really complex, and we don't have a good way of handling it. But when things get not just complex enough, but but the kind of thing that life is doing, which is this multiscale competency architecture, then even though the complexity has gone through the roof, the controllability actually gets better again. Because now they're offering this top down goal directed behavior, and you can those are now your control knobs, not the materials, not the physical properties, not what each part is doing. You're not trying to micromanage symptoms, which mostly what our biomedicine does. You are trying to convince literally, I'm using these words quite literally. You're trying to convince the system to be on the same page with you about what the set points of that homeostatic process should be. You're trying to get buy in so that it's not going to resist whatever medical intervention you've just done. This is this is I I think another way to think about this is the way you engineer with agential materials. So engineering with passive matter, even with active matter or computational matter, is different than engineering with an agential material. Let's say you use a completely different set of tools because you're working with a material that has competencies and and agendas of its own, and you just have to engineer differently in those cases.

Tim Scarfe 00:39:58

Yeah. I'm I'm a big fan of of much of this. I mean, you know, with mortal computation and biomimetic intelligence, you know, we're moving away from a paradigm of of abstractions and instructions, and we're moving more towards training, as as you said. But you you were saying some interesting things about agential material, and your definition of agency in in the paper was basically a kind of cybernetic loop. And philosophers talk about it in terms of self causation and intentionality. Of course, they're not talking about micro causation. They're talking about it as a lens. Many of the FEP people, mean, I've interviewed quite a few of them at this point. They talk about agency as an instrumental fiction or, you know, basically an as if property or or a model. And and even the free energy principle itself, it doesn't talk about the micro physical processes which leads to the emergence of things and and agents. It's just a property that those things have after they coalesce. We've got this abstraction called agency. And and then we use this abstraction to design AI systems and and to and to reason about new forms of intelligence and and so on. Is it is it a lossy abstraction?

Michael Levin 00:41:04

Well, yeah. Okay. First of all, my sort of metaphysical position is that everything is an abstraction and everything is a lossy abstraction. In other words, you know, people sometimes people will say, well, that's just a metaphor, you know, when you're talking about intelligence and memory and and tissue and whatever. That's just a metaphor. Well, yeah. And and what are pathways and what are genes and what are I mean, all all of it. I I think all all we have is metaphor. That's I I think that. And and the only question remains is is what does your metaphor let you do? So I don't believe that metaphors are exclusive. There are multiple different metaphors that you can have about the exact same thing. This is the kind of the nature of the polycomputing paradigm that Josh Bongard and I have been working on. So I think it's all metaphor. And but but what we want are enabling metaphors. We want metaphors that help us reuse concepts across fields to develop new concepts to drive new research programs. So is it is it a metaphor? Yep. Is it a construction? Yep.[6]

[6] Polycomputing paradigm — paper

Embodiment as perception-action loop, not physical body

Michael Levin 00:42:00

I don't know what else we have besides those things. And so now the only thing that we have to work on is to to to improve the ones we have. Now, you know, in terms of are are we going to miss things most most certainly. And and so that's, you know, that's as as scientists and philosophers, that's I think what we're in the business of is is trying to open up our metaphors so that we see more interesting stuff.

Tim Scarfe 00:42:23

Yeah. So just 1 point on that then because you said that you don't support a deflationary reductionist account of biological agency, you know, which is to say, we're just chemical machines because you said that framing is utterly insufficient for progress in science and engineering. But I did get the read though earlier on that that you were making that argument. And 1 of the reasons for that is I think when you were talking about embodiment, for example, you said, yeah, embodiment is crucial. And and then by the way, the reason why a lot of, cognitive scientists love embodiment is from this externalist tradition, you know, that they think that the intelligence is actually outside in the physical world. And when we interact with the physical world, we can kind of, you know, basically carry a lot of that intelligence in in some indirect way. But you said, you know, this goes beyond robotics, which gives an AI an actual body to cohabit an environment alongside us and beyond the coming revolution in virtual reality where many of us will spend a considerable amount of time in interesting worlds. So so so which is it? Presumably, you you do think that we can have very rich intelligences completely in silico away from the biological world.

Michael Levin 00:43:26

That's yeah. And that's because I I see embodiment slightly differently. So so typically, when people say embodiment, they mean the 3 d physical world. And you're moving around in the 3 d physical world, and that's what they you know, they say that organoids let's say brain organoids in addition are not embodied and, you know, certain kinds of AI is not embodied and so on. What I think is important about embodiment is the necessity of doing this perception prediction action loop. And that loop can happen in lots of different spaces that are not obvious to us at all. So biology, and this is not just for for robotics and AI, for example, biology. So so you could have and in fact, we do have cells and tissues

that may be going nowhere in the physical world, nothing much is happening as far as we can tell. But what they're doing is they're solving problems in physiological space, in gene transcription space. Once you get a multicellular kind of thing going, you can do solve problems in the anatomical morphospace. All of this taking measurements, judging them against your preferences, your valence, and your set points, and then figuring out what do I do next to improve my situation. All of this happens in all kinds of spaces, only 1 or 2 of which we know how to how to visualize at this point. And so I think we have to be very careful about what we call not embodied, and also I think the choice of spaces is very much up to the observer. You know, as as a simple example, imagine you have a bacterium and the bacterium is trying to go up a sugar gradient. Right? So what what we see is this loop that says take a measurement, and depending on what you see and what your memory is of of the previous values, can either go up the gradient or tumble and go somewhere else. You know, that's what we see. But actually, 1 other thing the bacteria might do is turn on a different gene that lets it metabolize a different sugar that's at a much higher local concentration than what we were looking at. So for the bacterium, taking a step in transcriptional space, is is does it represent that differently than taking a space in in in physical space? I'm not sure. Right? For us as humans, those are completely different things. And when I talk about problem solving and movement in transcriptional spaces, lot of people are like, well, that's not real. You know, that's not real space. But I don't know that the that the bacterium sees it that way. This is a this is a, you know, a formalization that we've made. And so I think we have to be very careful, and I think that kind of embodiment is critical, but it's not the physical space or the physical body that's critical. It's the loop. It's this having to be an agent to say, what do I do next? And having a set of options and a set of pressures and constraints and and and enough of a of a of a cognitive system which can be very minimal that that lets you do things where the future is gonna be different from the past. That's that's what I mean by embodiment.

Tim Scarfe 00:46:09

I guess maybe I wanna throw in 1 little tiny perhaps, caveat in there. And when when Stephen Wolfram impressed me on giving a definition for an agent, I said almost the exact same thing except I added 1 additional thing which is the actions I take need to affect the environment I'm perceiving. Yeah. Otherwise, like, I don't actually have this loop. I'm perceiving something and then I'm acting in some other space that has no influence on the the space I'm perceiving. The like, in your in your definition, does that also have to be true? Or can you be embodied by perceiving 1 space and acting in a completely different 1 and there's never any any loop?

Michael Levin 00:46:46

Yeah. No. No. I I like the emphasis on that. You know, it's a kind of niche construction. Right? It's it's a it's a it's the, know, you can think about niche construction and and stigmergy, you know, leaving use use it writing to the outside world as memory and so on. I I like it. I I think it's a continuum. So I think some spaces are much more easily bent than other spaces. And so, you know, in transcriptional space, yes, you've got the possible all these possible dimensions of gene expression. You can bend that space by changing your own transcriptional machinery, your own translational machinery, so you can act on the space. And then what you do next is now gonna perceive as somewhat of a different space. So I completely agree with you, but I but it I think there are degrees. Right? Some some spaces are more easily deformed than others, but, yeah, I I think that's an important part of agency. Yeah.

Tim Scarfe 00:47:34

Do you think there are any useful physical bounds to intelligence? I mean, I I know we've just been saying that the future eludes us, we always have people who are they just they can't conceive of the future, and they think what we have now is is very good. But 1 of the things you've said, of course, is that potentially what supersedes us could be a being with, you know, much, much higher intelligence. What are your thoughts on that? How could it be potentially limited by the physical world?

Michael Levin 00:47:59

I think it would be

Cancer as a shrinking of the cognitive light cone

Michael Levin 00:48:00

extraordinarily foolish of me to try to put limitations on on on intelligences that by definition are smarter than than I am. So I I don't I don't know. You know? I I I would not I I think we need to be very careful with limitations, period. I mean, again, just to go back, I don't think that intelligence is all there is, and there may be some kind of a filter whereby if intelligence proceeds too far ahead of things like compassion and, you know, I'm not even sure we have the vocabulary for, but, you know, some some kind of, you know, life positive features, it may spontaneously self destruct. So there may be this kind of feature of the world where just by itself, raw intelligence does hit a ceiling because because then there's

some kind of destructive, you know, outcomes that that prevent it from going further. But but but aside from that, if you if you, you know, if you manage that, if if the intelligence manages that aspect of it to keep up to keep up the keep up the growth in wisdom with the growth of intelligence, I wouldn't dare put a put a limit on it. I don't even know how you could put a limit on it. It's like it's like giving an IQ test to somebody way smarter than you. I just don't know how we would do it.

Tim Scarfe 00:49:08

Yeah. I mean, the the evolution of our biosphere certainly seems like almost a a flash compared to the history of of the universe. And and people do get worried about this rapid expansion. So of course, the AI risk people, they say we could have an intelligence system which is 10,000 times smarter than us, what whatever whatever that means. And of course, externalists think that we're all part of a a big collective intelligence anyway and there's already some kind of homeostasis or or some kind of limit. But you seem to be saying a little bit in your paper that we, think of ourselves as as, a distinct entity. You know, we we're we're this we're this human you know, humanity. We're we're humans. And and and I guess you're arguing that it's a category error to think of ourselves as distinct in the first place.

Michael Levin 00:49:51

I, you know, I I don't love fixed categories at all. And so it might be an error from some perspectives. It's probably not an error from others. I I I wanna emphasize, I guess, 1 aspect of it, which is that just because there just because we're a part of a higher collective intelligence, which we may well be, I I am not sure, That doesn't mean that everything is great and our interests are taken care of. So we actually are not very good at predicting or knowing about the goals that that the composite systems will will have and that that the emergent to go you know, goal seeking systems. We're not we're not very good at knowing what the goals will be. And if we just look at the the, you know, 1 clear example where we see this happening is is in multicellular bodies. So, you know, you could you could obviously make the argument that a lot of new and interesting things happened when you became an emergent human and not a pile of individual cells. But, you know, when you go I don't know. This is for some reason. This this is example I always I always go to. When you go rock climbing, you know, you you're gonna have a great day, and you're gonna you're gonna achieve some personal goals. You're gonna achieve some social goals. You're gonna have a great time. You're gonna leave a whole bunch of your skin cells on that rock face, And they didn't get to ask you know, they were not asked whether this was a good deal for them. You know, people do all kinds of stuff to their bodies for various reasons.

Just because you're a collective something is, you know, is part of a collective intelligence does not necessarily mean that that collective intelligence is looking out for you at all. So maybe, but maybe not. And so we need to you know, I I don't think we can just assume that because we are part of a greater system that there's some kind of benevolent homeostasis. I think we have to be very careful with what kind of emergent systems we create, and this is the reason why this is so important. I I think that developing this field of diverse intelligence is an existential level task for humanity because we are constantly building large scale emergent cognitive systems where we have no idea you know, we we're we're just very bad at predicting when it's going to happen, knowing or shaping their goals, communicating with them in any way. You know, we make social structures, financial structures, Internet of things, robotics. We we we make these things in various spaces, and we still have no idea what what we're really making because we assume that it's either going to be mechanical and controllable or random and chaotic, but we don't understand that that many of these things are in fact goal driven systems.

Tim Scarfe 00:52:21

I'm glad you brought up sort of, you know, leaving the skin cells on the, on the rock face because it's the same point I made earlier about people being so myopically focused on their own individual existence and and opining about how all these imperfections that the human cognition, brain, body, etcetera, has where they don't understand they're part of an an evolutionary algorithm that has goals, like, that transcend their their individual goals. So, like, a skin cell could sit there and say, like, why is it that the you know, I'm I'm in epithelial cells. Why are my fibrinogen bonds so weak? You know, how come I just it's so easy to get sloughed off? Why is my lifespan so short? Well, because because you've been optimized for something that actually has nothing to do with you as an individual. It's you as a component of a larger system. Right?

Michael Levin 00:53:07

Yeah. And and and, you know, and and there are so there are defections from that from from that system. So, for example, Tasmanian devils have a have a weird kind of mouth cancer that propagates from animal to animal when they bite each other you know, when they when they fight. And so so cancer in general is is is is a shifting of the cognitive light cone from from from, let's say, organ level down to individual cell levels. So cells roll back in their sort of amoebas. And typically, that goes very poorly for them because, you know, they they they have kind of a short term vision where they're like, okay. I'm I'm I'm leaving this collective. I'm going out on my own. That that doesn't typically work out very well. But for I I I think the 1 in if I'm if I'm not mistaken, I think the the the lineage of mouth cancer in Tasmanian devils is, like, 11,000 years old or something like that. So and and there's an there's another example with clams. So clams cancer

LLMs, diverse intelligence, and epistemic humility

Michael Levin 00:54:00

cells and clams get released into the water. They hang out as single cells for a while. You know? No problem. Eventually, they'll find another clam, settle in, and, you know, there they go. So, typically, that kind of defection from the body plane is a dead end, but not always. And, yeah. I mean, you know, biology has has observers and goal seeking agents at every level competing and cooperating with each other. And just because you're part of a larger system doesn't mean that that it's looking out for you.

Tim Scarfe 00:54:26

On the cancer thing, you you've done some talks about this where you've got a new way of thinking about oncology almost as a subdivision of of the self. So it's it's like a new part of yourself which has its its own goals. But that lens, I think, actually could lead to new forms of treatment. Right?

Michael Levin 00:54:41

Yeah. Yeah. I mean, I think 1 1 way to think about it is to ask why why do we ever have anything but cancer? In other words, we start life started as individual cells. They then then they started cooperating towards these grandiose construction projects, building, you know, livers and hearts and and things like that. Why why why do they do that? So I think 1 thing that we've been working on is this notion of a cognitive light cone, which is the size of the biggest

goal that a system can pursue. And we've characterized the number of mechanisms that allow individual cells to bind together, and the network has bigger cognitive light cone than the pieces. So So suddenly when when you're an individual cell, you care about your your own metabolic state and and and things like that. When you're a part of a network, there's a little bit of a mind meld that happens with these other cells, which we characterize molecularly. And and now you're working on very large scale goals in the anatomical morphospace. Well, that that process that that goal scaling process has failure modes, and we know how to how to induce it. And then, of course, in nature, it happens from time to time. And then what happens is that cognitive light cone, basically, the cells disconnect from the network. That cognitive light cone shrinks to now being the size of a single cell. And so now you're back to it's, you know, it's not anything crazy or novel that's happened. You're back to an amoeba. As far as you're concerned, the rest of the body is just external environment to you. Everything you care about is around a very tiny little little radius, both in terms of space and time. And that suggests, thinking about it that way, that that basically these cells are not any more selfish than other cells. Right? Like in game theory modeling of cancer, would you you could say they're less cooperative, whatever. I don't think they're any more selfish than any other cell. I think their cells are smaller. That's all. They're they're they're still they're as selfish as the collective. It's just their self is tiny now. And so that that way of thinking about it leads to a very specific road map, which which we've been following, which is let let's just try to reinflate their cognitive light cone. Let's let's reconnect them to the electrical and physiological network of the other cells. We're not gonna try to kill it. We're not necessarily gonna try to look for hardware defects. Some cells have it, some cells don't. You don't need any kind of genetic defect to to to have this kind of carcinogenic transformation. Let's reconnect them. And and so we've done that in animal model systems, and it works really well. And so now the question is, can we, you know, move it to human medicine? But this is that that's that's another example where taking these different perspectives, which almost see like almost seem like philosophical perspectives, they lead directly to actionable policies in in experimental work.[7]

Tim Scarfe 00:57:09

I guess it'd be good to get some of your comments on large language models and what they're good at and what they're bad at. I mean, I I certainly think they are lacking in creativity and and agency, but they are surprisingly good. Right? They they have somehow inferred all of this information. We have phenomenal sensations and we understand each other with this, you know, physically embodied,

[7] Cognitive light cone — paper

informed, you know, setting. And then the language models act as if they have a similar kind of understanding, but but they don't. But of of course, they they don't actually have agency. But it is that a useful distinction?

Michael Levin 00:57:39

I think it's a reasonable distinction, and I share your intuitions about it. But I want to say that I think we need to be very, very careful about feeling certain about these things. And the reason is is that we have by by exploring extremely minimal systems, like sorting algorithms, like gene regulatory networks, we have found massive surprises, not just emergent complexity. Emergent complexity is easy. Emergent unpredictability is, you know, it's cheap. It's it's emergent cognition. Basal, but emergent cognition. And so, you know, what if if if I had to if I had to bet money, do I think that today's language models have high agency? No. I don't. Am I certain of it? Absolutely not. Because because my the lesson I've learned is if I don't even know what bubble sort does, then I I probably don't know what these things are really doing. We we we are very it's I I I think it's very early in the field of diverse intelligence to be able to be certain about any of these things. I you know, ostensibly, I don't see when I when I look at the at the structure of these of of of today, you know, language models and so on. I I don't see the explicit kinds of processes that I that I think are important for agency, but I've also learned that just because we don't see them and we don't know how to look for them, that they're not there. So I think we should be we should be very careful. I don't think they're anything like human minds. I think that's very unlikely. I think that, you know, for the first time, we've so so it used to be that if you see something that speaks, you you're pretty sure that it's like you in in a lot of ways. So I think we've now we've now, you know, got we've dissociated those 2 things, and we now have things that talk and put up a really good, you know, a really good performance in in in having a conversation. I don't think they're like us at all, but I think we need to be really careful about what we, you know, what what we are sure of that they do and don't do because our intuitions fail at at such a low at such a simple minimal level that you know, I I've met all sorts of people who say, well, you know, language models don't have this and that. I know I built them. Like, I I I write the code. I know. They don't have and I'm like, well, you don't even know what bubble sort does. I don't think we can say what what this thing does. So we just I I just think we have to be careful.[8]

[8] Cognitive properties in sorting algorithms — paper

Tim Scarfe 00:59:55

Yeah. It's weird you say they're not like us, but they they kind of are like us. And then there's the the the realization that when it becomes indistinguishable, you know, that does it really matter? But I think it's instructive though because in in the in the virtual world, there seems to be a couple of differences. Right? So agency, of course, it's about, you know, having goals and intentionality. But it's also about causation and power and the ability to realize those goals, the ability to change the world to meet your goals. And what would it mean to have a collective, agential AI running out there on on the Internet? You know, do you think it could actually does it make sense to call it agency in the same way we do in the physical, biological world?

Michael Levin 01:00:36

Again, I I am not I don't have any evidence that today's architectures have what's necessary to do it. And and so so but but but in but in general, I think it absolutely makes sense because you could have a system that's doing the kinds of things that we're talking about in a in a in a in a nonphysical world. Now having said that, I I don't believe that our algorithmic model so so having algorithms, having a Turing machine metaphor, don't I think any of that is sufficient for this. So this is not a computationalist argument, that says that, you know, any any kind of algorithm will automatically can automatically do it. I I I don't think that's necessarily the case. But I neither think that you actually need the physical world for it. And by the way, these things these things are embodied to an extent. We are their body. I I I'm not even sure that that these kinds of things are truly disembodied. In in an important sense, we are their body. We move massive amounts of energy, you know, money, resources, whatever based on the things these things do and don't say. And, I I don't have any evidence that today's language models have anything like a a core goal driven agential component, but I think it's completely possible. And I think we need to be very, very humble about, feeling certain about what they what they do or don't, don't have.

Tim Scarfe 01:01:57

Just a quick a quick point on that. There's this interesting kind of agency sharing between the physical and virtual worlds. Right? So you can think of a meme on Twitter as having agency because it's expressed by physical actors in in the real world. So in a sense, there's a transfer. But maybe purely in the virtual world, there could be no no agency.

Michael Levin 01:02:17

Well, look, you know, in an important sense, you and I are living in a virtual world. We we don't have access to and I'm not talking about a simulation, you know, with an alien zoo or something like that. I'm just I'm just pointing out that, you know, you and I are conducting our behavior under extremely limited transduction from an outside reality that we don't really have any access to. You know, there's a there's a there's a big hole in our, you know, in our in our visual field that we don't see. There's 1000000 other things that are very different to how we perceive them. There are massive amounts of information that are inaccessible to us. Everything that you and I do are is is within a construction that we have pulled together ourselves as our the best story of an outside world that we could tell. And, you know, yes, the physics of it are cool and they give rise to all kinds of morphological computations, but there are properties like that in other virtual spaces. And I don't think we should be, you know, too too tied to this this this what we see here as some kind of privileged reality that is the only way that agency can be done.

Tim Scarfe 01:03:21

Professor Levin, it's been an honor. Thank you so much for joining us today.

Michael Levin 01:03:24

Thank you so much. Yeah. Had a good time. Thank you.

References

[1] **Michael Levin**
https://as.tufts.edu/biology/people/faculty/michael-levin
person

[2] **A Fire Upon the Deep**
https://en.wikipedia.org/wiki/A_Fire_Upon_the_Deep
book

[3] **Bow-tie architecture in biological systems**
https://doi.org/10.1126/science.1089072
paper

[4] **Richard Watson**
https://www.southampton.ac.uk/people/5x2l5c/professor-richard-watson
person

[5] **Philip Goff - Galileo Error**
https://en.wikipedia.org/wiki/Galileo%27s_Error
book

[6] **Free Energy Principle**
https://doi.org/10.1038/nrn2787
paper

[7] **Polycomputing paradigm**
https://doi.org/10.1098/rsif.2022.0810
paper

[8] **Cognitive light cone**
https://doi.org/10.3389/fpsyg.2019.02688
paper

[9] **Cognitive properties in sorting algorithms**
https://doi.org/10.1016/j.biosystems.2023.104973
paper