/mlst

# Don’t invent faster horses - Prof. Jeff Clune

March 23, 2026



Machine Learning Street Talk

Podcast Transcript

Revision: 67136ad3

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

Open-Ended Evolution and Innovation 3

Sponsor Messages 4

Novelty Search and the Paradox of Objectives 5

Interestingness as a Proxy for Progress 8

POET and Co-Evolving Environments 11

Darwin Complete Search Spaces 15

Code Generation vs Neural Networks 19

Thought Cloning and Behavioural Learning 22

Language Emergence and Interpretability 25

AI Safety and Deception Detection 28

Alignment Limitations and Open-Ended Risk 31

Open Source AI and Global Governance 34

ADAS: Automated Agent Design 38

Continuous Learning Challenges 43

OMNI-EPIC: LM-Driven Environment Generation 46

Evolutionary Biology Insights 50

From Philosophy to AI Research 53

References 58

## Overview

Jeff Clune has spent his career chasing one of science's biggest questions: how did evolution produce the explosion of complexity we see in nature, and can we build algorithms that do the same thing? In this wide-ranging conversation, he lays out the case for open-ended evolutionary algorithms – systems designed to generate novel and interesting outcomes forever, drawing on principles from both Darwinian evolution and human cultural innovation.

Clune explains the central paradox of his work: trying too hard to accomplish a specific goal is often the worst strategy. Instead, the best results come from recognising serendipity and keeping hold of interestingly new things, regardless of whether they seem immediately useful. This insight, drawn from Kenneth Stanley's work on novelty search, underpins a new generation of algorithms that use foundation models as judges of what counts as genuinely interesting and novel.

The conversation covers POET (evolved environments for reinforcement learning), NEAT (neuroevolution of augmenting topologies), ADAS (automated design of agentic systems), and OMNI-EPIC (using language models to generate open-ended environments). Clune walks through how these systems riff on previous discoveries to create increasingly complex challenges – from simple ball-kicking tasks through multi-room buildings to cluttered restaurant scenarios that robots must navigate.

The interview also tackles AI safety head-on, with Clune advocating for democratic governance coalitions, regulation of frontier models, and global alignment protocols. He discusses why the interpretability problem may be harder than it looks, how open-ended AI systems could pose unique risks, and his view that the biggest danger is not acting on safety soon enough.

## Open-Ended Evolution and Innovation

**Jeff Clune** 00:00:00

3 things that people will learn from watching this show. 1, what is the connection between Darwinian evolution and human culture? And how can we use the principles that fuel the amazing things those 2 processes have done to create algorithms that can innovate forever. 2, the paradox of how trying too hard to accomplish something is likely to fail, and the best thing you can do is recognize serendipity, chance on the wing, and say, if it's interestingly new, let's keep it around because who knows where that idea 1 day might lead. And 3, that foundation models have finally opened up the ability to have an algorithm know what counts as an interesting and new thing to try out. And having a system that just keeps trying things that are interesting and new and or high performing and adding it to a growing library of such discoveries is a powerful playbook that you can apply to almost any domain and see amazing progress and amazing things happen. So I think 1 of the grand challenges of computer science and also you could think about it from a biological perspective, grand challenges of biology is to try to understand how did evolution produce the explosion of amazing things that we see on earth. You look out of the natural world and you see jaguars, hawks, the human mind, 3 toed sloths, birds of paradise. All these things just kind of popping up over time. This amazing menagerie of engineering marvels, diversity. It's so fascinating. And actually that has been the central quest of my career from the beginning. That, how did evolution produce this complexity and how does intelligence happen and they relate. Like how did evolution produce the intelligence that you have in your brain, which is the most impressive bit. So, you know, a learning machine that we know of in the universe. Right now, there are people saying goodbye to their children because their children have cancer or some incurable disease. There are people saying bye to loved ones of all ages. There are people suffering from hunger and poverty. You know, excruciating human suffering and pain is rampant in this planet. This technology has the ability to eliminate all of that. I believe firmly that we can cure all disease. We could solve death. We can dramatically increase the quality of people's lives, their GDP, eliminate things like hunger scarcity if we get the societal distribution questions rotten done. But, technically, a lot of this stuff is possible. And so Dario Amadai wrote a beautiful essay called Machines of Love and Grace. And as Dario says in the essay, the reason why he talks about AI safety is the same reason I talk about it because that's what we need to get and figure out in order to unlock all of that great stuff. And so I don't want to die. I don't want my children to die. It would be great if we could eliminate though that from the world and make death optional. And the only path to doing that and many other great

things for humanity is[1][2]

**Tim Scarfe** 00:03:06

We are sponsored by Tufa AI Labs, an exciting research lab which has just started up in Zurich by Benjamin Crouzier. They have acquired Minds AI, the winners. When I say the winners, mean they didn't they didn't submit their their solution. But anyway, they got the highest score. Yeah. So they're all working together as a team. They're working on ARC. They'll be working on the next version of ARC when Francois releases that and they are hiring cracked engineers. Does that sound like you? Well, you know what to do. 2 for labs dot a I. MLST is sponsored by Sentinel, which is the compute platform specifically optimized for AI workloads. You might have seen the interview we did last month with Gennady, their CEO and co founder. He spoke about some of the optimizations on Sentinel, which only they have done, which makes it dramatically faster than the competition. Anyway, if you're interested in hacking around the Sentinel, I'm gonna be doing a live with 1 of their engineers in the coming weeks. So feel free to jump on that live and you can ask their team any questions.

## Sponsor Messages

**Tim Scarfe** 00:04:12

Jeff, it's amazing to have you on MLST. So I was saying to you before, I first discovered you 2018 ICML when you're doing your workshop and you're talking about open endedness and and of course, we are huge fans of Kenneth Stanley on the show. We've had Joel Lennon on. We've had Tim Roctacial on. We've had loads of people on. Open endedness is an absolutely fascinating area to me. So to have you on the show, 1 of the pioneers in this field means so much to me. So thank you for coming on, Jeff.

**Jeff Clune** 00:04:38

My pleasure. It's great to be here. And, you know, you've named a lot of people that I deeply respect and have worked with for years. So it's great to finally be here.

[1] POET: Generating/solving complex challenges — paper
[2] POET: Generating/Solving Complex Challenges via Paired Open-Ended Trailblazer — paper

**Tim Scarfe** 00:04:45

So 1 of your life goals, you've said, is to create an algorithm that keeps running forever with no end. You know, we think there's something fundamental about how intelligence works in the natural world. But rather than trying to capture it down at the metal, you know, like there are these biologically inspired folks talk about

## Novelty Search and the Paradox of Objectives

**Tim Scarfe** 00:05:02

**biomimetic intelligence, You've got this very interesting approach where you kind of bootstrap intelligence like in the natural world in such a way that doesn't lose important characteristics of the natural world. Can you tell us about that?**

**Jeff Clune** 00:05:17

**Yeah. Sure. So I think 1 of the grand challenges of computer science, and also you could think about it from a biological perspective, grand challenges of biology, is to try to understand how did evolution produce the explosion of amazing things that we see on earth. You look out of the natural world and you see jaguars, hawks, the human mind, 3 toed sloths, birds of paradise, all these things just kind of popping up over time. This amazing menagerie of engineering marvels, diversity. It's so fascinating. And actually that has been the central quest of my career from the beginning. That how did evolution produce this complexity and how does intelligence happen and they relate. Like how did evolution produce the intelligence that you have in your brain, which is the most impressive bit. So, you know, a learning machine that we know of in the universe. And you could try to do it and many people have tried to say, hey, we're gonna go all the way down to the lowest possible level we can imagine. We're gonna create, you know, self replicating machine code or even like self replicating artificial cells and then we're gonna like hope that that kind of bubbles and percolates up into some open ended process that eventually produces, like, an entirely new form of interesting intelligent life, and that will teach us about the process. But as Josh Tenenbaum told me when I was telling him about some of these goals of mine, he said, you know, you don't have a planet sized computer to work with. So how are you gonna accomplish this within, you know, the lifetime of you as a scientist or like us as a community? And to do that, the key is gonna be abstraction. You know? We don't need to recreate every single detail in biology in order to study and understand the core principles. And in some sense, nor would we want to. You know? We want**

**to understand at an abstract level what are the key ingredients that make this process work and that produce these endless marvels. And so in some sense, the more abstract you can make it, if it still has the properties you want, like a complexity explosion, then actually you've done the best job because you've figured out the difference between what was necessary and what was incidental. Another way to think about this is your own brain. I don't believe in the blue brain project philosophy, which is let's simulate every single chemical and every quark in your brain in order to try to produce an intelligent machine. Actually, wanna say chemistry, a lot of that detail is probably not necessary. It's not part of the secret abstract recipe for intelligence. Let's figure out how can we abstract it and still get intelligence. And so you could apply that analogy to the study of open endedness as well. What are the abstract principles that create a thing that literally you could run for billions of years and would still continue to surprise and delight you?**

**Tim Scarfe** 00:07:56

I'd love to know what's the difference between you and Josh Tenenbaum. He he's a huge inspiration to me as well. And I think you guys might have merged in some ways. I had Kevin Ellis on here as well, and he's done some fascinating work with Dreamcoder. And increasingly, I think what I'm seeing happen is we are letting language models do bootstrapping for us. You you said yourself, you know, the reason we don't go down to the particle level is we can benefit from billions of years of evolution. There's this bootstrapping process and it should be possible for us to build on top of that. And that that's what I'm actually seeing from Josh's team and and and yourself.

**Jeff Clune** 00:08:32

That's right. So I don't think Josh and I differ too much in this. Actually, I he invited me to to give a talk at MIT, and he pulled me aside or maybe even said this in in the q and a. He's like, if I wasn't doing what I was doing, the thing that I would really wanna be doing is what you're doing. He's like, he loves this work. He thinks it's fascinating. He wants to work together. Yeah. And I wanna work with him. So we're actually kindred spirits and super aligned. I think if if I had to say what is the difference? I think he's a little bit more interested in what is happening specifically up here in your brain and the human brain and modeling that, albeit in an abstract way using Bayesian principles and the like, but really trying to understand the human mind. Whereas I'm trying to basically say, could I create a prob a process that would innovate forever, including making a form of intelligence, but maybe an entirely alien intelligence that doesn't actually look exactly like what happens in the human brain? In fact, 1 of the things I find most fascinating about open ended work is if we get this to work, and that's a big if, we're still working on it. But it's almost like inventing the ability to travel through the universe and meet other intelligent races because the intelligence that could emerge from this system might be wildly different. And so you would come to understand intelligence in general, the spaces of possible intelligence. What does alien math and music and humor look like? Well, you could do kind of alien travel and visit those cultures except it'd in silico instead of interstellar.

**Tim Scarfe** 00:09:55

Just on on that externalism point, I'm I'm fascinated by collective intelligence and all the rest of it. How much useful human intelligence is is in our brains?

**Jeff Clune** 00:10:04

Oh, boy. That's an interesting question. I think that to some extent, there's so much stuff in our brains that at first pass, you might say, is totally irrelevant and a total waste. You know? Like, the Guinness Book of World Records is simultaneously 1 of my favorite things and also kind of like a maddening aspect of human. You know, there was like a quote from somebody who said chess is like a giant waste of brains. We've spent so much time getting really good at this random game. Know, You you could say, didn't you just take all of that collective brainpower and apply it to curing cancer or something? And there are people that are even more esoteric, you know. Their their shtick in life is I I learned just a couple days ago, there are some people who wanna park in, like, every single parking spot at, like, an arena, you know? Tens of thousands of possible parking spots and they every day they wanna park in a different 1 and they check it all off and figure out exactly what's going on. And you might say to yourself, like, is a collective huge waste of time and in some sense that's true. But there's this the fact that humans have a passion to do something new and weird and different, that is the rocket fuel of innovation as Ken and Joel in their great book talk about. It is people like that who end up doing weird things and making weird discoveries that ends up providing new stepping stones and innovations and ideas that end coming oftentimes are the things that truly unlock progress. So if somebody wasn't interested in some weird esoteric math proof or computer theorem or tinkering with some new type of chip that can do things in parallel as opposed to sequentially, then maybe we don't get a lot of the innovations that we have. So[3][4][5]

## Interestingness as a Proxy for Progress

**Jeff Clune** 00:11:40

**a lot of it is a waste, but maybe we'd be worse off if we didn't allow a little bit of waste.**

**Tim Scarfe** 00:11:46

**Yeah. This this is so interesting. My co host, you know, doctor Duggar, he gives the example of this drunken walk analogy. You're trying to solve a maze and you have like 1000 drunk people and they will just go in different directions and and then 1 of them happens upon the solution. And we have this language, we have this memetic cultural transfer and so on. So 1 drunk person finds the**

[3] Why Greatness Cannot Be Planned — book
[4] Why Greatness Cannot Be Planned: The Myth of the Objective — book
[5] Kenneth Stanley — person

the solution to the maze and then it transfers to everyone else. And that's a bit of a deflationary view of intelligence because it it almost alludes to it's simply just a commensurate amount of computation and trying random things. We wanna do a little bit better than that. We we think that there's there are shortcuts. And this is when we get to this interestingness thing because it's not just about doing random things. It's about doing interesting things. And you said that being able to define interestingness is the grand challenge in open endedness. It's ineffable. It's unquantifiable. Tell me more.

**Jeff Clune** 00:12:35

Yeah. So I do think that 1 of the things that humans are great at is they have a really great taste or nose for what counts as interestingly new. We're curious, and we can look at a new game, a new form of art, a new form of literature, and we can say, that's interesting. That kind of like I that that thing that that person did, I I recognize that as interesting and he knew. Like Jackson Pollock, when he splashes paint on a canvas, there's this wonderful quote from the movie where the Val Kilmull character says, you know, you broke through. And it was true. They had been playing a game and then he kind of recognized, oh, there's something new I can do here that feels right and is interestingly different. And other humans recognize that. And so that sense of how to explore efficiently and have a nose for what's curious is the thing that drives cultural progress both in the arts and also in the sciences. We have not yet figured out in our algorithms how to get that human notion of curiosity and what counts as an interesting next thing to try, next skill to learn, next environment to generate, next scientific hypothesis to test. We haven't figured out how to inject that into our algorithms up until maybe I would argue about a year or 2 ago. And the problem is because if you try to to say specifically, here is the definition of what counts as interesting. If we write a little mathematical function or a little piece of code to measure that, it always fails. And this is goes back to good Goodhart's law. Right? Which is once a measure becomes a target, it ceases to be a good measure. If you write down and I'll give you an example. So in intrinsically motivated reinforcement learning, we said, hey, we don't want the agents to randomly flail around until they get a reward signal because you'll never randomly discover a high rewarding behavior like flying a plane and landing it. Right? And you're never randomly gonna land the plane and get your payoff. So you have to be curious and we want you to be intrinsically motivated to just find new interesting do interest new interesting things. So the way that we code that up is we say go to new states, which basically means make new observations. Like I want my eyes to see new stuff. What happens when you start telling an agent to do that? Well, finds a TV and it sits down and it never stands up again. It sits there and literally will watch

**TV forever because the TV is constantly flashing to it in new situations. It gets worse than that if it just found a TV even tuned to a dead channel, that endless stream of pixels. No 1 of those patterns of pixels is ever like the the past thing and it's just like new new new new new new new. So we got what we asked for. It finds new states. Not what we wanted, which is new interesting states. And this story repeats itself over and over and over again. There's been like decades of people being like, oh, I think this might finally unlock the robot to keep finding new situations. And then you run you let optimization go against it and you just discover, oh, I didn't realize that my my measure will be lit up by this particular weird pathology. So another 1 that's very very common is learning progress. So you said, okay, we want the robot to learn forever, so we're gonna keep giving it challenges on which it can learn. Well, if I generate 1000000 digit long random number, right, and I give you the first 10 digits, well, and I say memorize that, well, you'll learn there's some learning you can do there. And when you're done, I'll say, here's the next 10 digits and the next 10 digit. Well, now you have infinite learning progress forever. It will learn. I can you can learn forever, but that's a completely boring and useless task. Right? It doesn't count as it like the fiftieth time you tack on 10 more numbers, that's not an interestingly new thing to do. And so what we really want is to somehow we want algorithms to know the common sense human notion of, hey, after I've like memorized this digit up to a 100 decimal places, there's something more interesting to try in this universe or in this world and I should go invent a new form of math or practice juggling or unicycling or something. Now, up until now, we haven't had the ability to do that because we didn't know how to code it. But there's a new player in town and that is foundation models, frontier models, large language models, whatever you wanna call them. So I'm gonna call them frontier models. These frontier models have read the entire Internet, including lots and lots of people complaining about what is boring online and talking about what they find interesting. So these frontier models have literally distilled a sense of interestingness from human culture and human data into their weights. And now finally, we can just go ask a piece of computer code, does this thing count as interestingly new? And in my lab at UBC and in some of the collaborations we have, we've basically applied this lesson over and over and over again, and it does all sorts of marvelous stuff. So you can have an environment generator that says to generate the next challenge that will be interesting for the robot to learn, do that forever, and you get this explosion of interesting environments that it could work on. We have the AI scientists say, I want to do a new scientific paper that is interestingly different or makes an interestingly new discovery from everything that I've seen before. And we do that, and it does it again and again and again. You can apply that to it developing agentic systems, to causing different strategies in a self play game. And if you want, we can talk about all these papers. But the**

**basic building block is the principles of open endedness, keep expanding an archive or a population of interestingly new things. Now we know how to ask a computer, does it count as interesting? So now we almost kind of have that missing secret ingredient that has been the grand challenge up until now. And maybe now we have the ability to put a big dent in it.**[67]

**Tim Scarfe** 00:18:08

Yes, indeed. I remember Kenneth Stanley said that we have a nose for the interesting, you know, that it's what got us so far. It's the basis of of all innovation. And by the way, when when I interviewed[89]

## POET and Co-Evolving Environments ●

**Tim Scarfe** 00:18:20

Tim Roktaschel about that DeepMind paper, let me just find the the name of that paper. This was at ICML. Open

**Jeff Clune** 00:18:26

endedness is necessary for

**Tim Scarfe** 00:18:28

Essential for artificial superhuman intelligence. Yeah. So they said the definition of open endedness hinges on a system's ability to continuously generate artifacts that are both novel and learnable to an observer. And 1 1 thing I said to them at the time is I wanna get into the creativity discussion a little bit. Now, language models, I must admit, I'm I'm updating all the time. So I was very cynical and I always used to say there's a reasoning gap and there's a creativity gap. It seemed to me that even though they can extrapolate, they are parasitic on the data, they are, you know, parasitic on the prompts that they know, so it's almost like, where's the agency actually coming from? Is it us as agents putting stuff into that? So it kind of feels like there's a limit. But what we're seeing now is that either through transduction or induction, we are generating language and programs using these language models. And it seems to be quite compositional. It seems to be that we can traverse the entire kind of Turing computable space by generating all of these different programs with with the language models. So the

[6] Automated capability discovery in foundation models — paper
[7] Automated Capability Discovery in Foundation Models — paper
[8] NEAT: NeuroEvolution of Augmenting Topologies — paper
[9] NEAT: NeuroEvolution of Augmenting Topologies — paper

first question is, is is there a limit, do you think, to the types of paradigmatically new programs we could generate?

**Jeff Clune** 00:19:32

This is a great question. I actually call this right now the trillion dollar question because I think there's tremendous economic value at stake. If language models can be creative and continue and as part of these open end algorithms that we're making, if they can keep innovating and making new discoveries forever, there's tremendous economic potential. I do think these models can innovate and potentially forever. I have to give you 1 example that I literally only saw about an hour before coming here. So we have this paper called the AI scientist. Yep. It uses the ideas of open endedness to ask to create a system that will generate its own idea for a new entirely new piece of research based on what it knows from reading the Internet. It proposes that idea, proposes the experiments that it should do for that to conduct to test that hypothesis, does all the experiments, looks at and plots the data, writes up the whole paper, peer reviews it, etcetera. And in the open endedness philosophy, you would publish that paper to an internal archive, then have the system read that that paper and all the other papers and build on its own discoveries forever, right? That's what we're after. So in the process of looking at the papers that that system creates, I was super impressed with how creative it was in proposing entirely new paper ideas. Basically like grants, somebody should study this. And 1 of my favorite ones was the idea that hey, there's this Grocky phenomenon where neural nets suddenly learn a certain task. Is the beginning they start memorizing something but eventually they learn the underlying principle of like the the like how to do multiplication and then they basically suddenly go off to the races. And this AI scientist said, hey, you could see if there's like the compressibility of the network, the minimum description length of the network goes down right around the time that grokking happens because instead of memorizing much of stuff in their weights, they've learned the underlying principle. And once you know the principle, then you can shed all the memorization. And I was like, that is a really interesting cool idea. Like if a graduate student proposed that to me, I would say you should go do that. And then I went and did all the whole paper. Just an hour ago, I saw that some other group, I think independently came up with that idea, did all the research and posted it and the machine learning community is loving this paper right now.[10]

**Tim Scarfe** 00:21:42

I saw your tweet. Oh, you saw it?

[10] MAP-Elites: Illuminating Search Spaces — paper

**Jeff Clune** 00:21:44

Yeah. Yeah. It's incredible. It's like literally the AI showed that it was as creative as a team of human of cutting edge human ML scientists. Yeah. And it got there first, which is incredible. So that's 1 example. Now you're asking me a question that I think is fascinating, which is maybe you can do that now. Like, maybe you can take 1 jump, 1 or 2 concentric rings, like, of an onion adding on to our our set of discoveries, but maybe it can't go much farther than that. There's a lot of people who think you can interpolate, you can extrapolate a little bit, but you can't do wild creative leaps. 2 comments on that. 1, I'm not sure that humans are able of making these wild creative leaps. Mhmm. You know, the reason why we say things like if I have seen further is because I stand on the shoulders of giants is because human scientists are always looking at what came before and make a relatively small leap off of that. It's why you see calculus simultaneously invented by Leibniz and Newton. It's why you see evolution simultaneously invented by, I think it's Huxley and Darwin. This happens over and over and over again. Both of them conditioned on an archive of ideas, say, oh, this is what's next, and this is what's next. Right? I think our models can do the same thing. They look at everything that came before, and it's just as we saw 2 hours ago. The AI can make the next leap or the human can make the next leap. And then as long as the AI system is publishing high quality papers internally, looking at its own discoveries, it will make the next leap and the next leap and the next leap just like a human community of scientists would do. So my supposition is that your the model of interestingness that the system learned looks at the history of what's coming for and can just recognize the next interesting idea, and as the archive grows, that thing can generalize really far because the notion of interestingness didn't change. It's just what you're conditioning on changes. Here's my thought experiment for you. If I teleported you teleported you to the year 3000, you'd be blown away. Right? Lots of cool stuff. Once you became acclimated and and like knew the current tech, my prediction is that you would recognize that the next innovation in the year 3001, that counts as interestingly new. You would generalize way outside of your distribution. I think the model of interest in us in our current language models, I can't guarantee it. I think there's a decent chance it might too as well. Yes. If it doesn't, then we do some fine tuning stuff. We update on the fly. We gotta figure that out. But more so than anyone else I've met, I'm optimistic this thing might generalize far.

**Tim Scarfe** 00:24:09

I I love all of this, and I'll explain why. I I think the reason you understand this is because, for example, in in in the poet paper, it was essentially doing a kind of curriculum learning where you, you know, like you've kind of you generate all of these different environments and so on. And you can look at the phylogeny of the environments. And I think you said actually it's like the best plot of your career or something like that. Because when we talk about creativity and language models, I I think the problem with the perspective is we we think about it in 1 shot. We don't think about the process. Right? So actually, what we're doing as a society is epistemic foraging. We're we're stepping stone collecting. We're finding all of these stepping stones, and every single stepping stone builds on the last 1. And you said, I think with the poet paper, it might have been another 1, that, you know, you can understand what the curriculum is. But if you try and shortcut it and train another neural network to do the same thing, it doesn't work. It's like it needs to have this this weird gnarly phylogeny of knowledge. Similarly, if if you give a physics book to you go back in time and you give it to Newton, he's not gonna understand anything because there's this whole, like, phylogeny of knowledge where 1 bits 1 bit of knowledge builds on on on the previous bit. Right?

**Jeff Clune** 00:25:12

That's exactly right. And that is the central thesis of Ken and Joel's book, Why Greatness Cannot Be Planned. You can't pick a really distant ambitious goal and march straight toward it. The best thing to do is just do a bunch of basic research, collect new interesting ideas, and eventually, you'll find some weird circuitous path that leads to an to that invention. My favorite examples of this, which I say in my talk, so that if you went back to the time of cooking technology over an open fire and you only funded people who made fast cooking technology that made no smoke, you'd never invent the microwave. Because you to invent the microwave, you needed to have been working on radar technology and noticed that a chocolate bar melted in your pocket. Similarly, if you went back to the time of the abacus and you said, will only fund scientists who give me more compute per hour or per dollar, You would never invent them on a computer because the technologies that unlock the computer were electricity and vacuum tubes. And at the time, there was no idea that those things would lead to better compute. And so I what the cool thing is we want inside of our algorithms, our open ended algorithms and this is what we do in my lab and is we want the algorithm to kind of catch chance on the wing, to recognize serendipity, to say, that's funny and that's interesting, that's new. Let's let's keep that idea. Let's add it to the kind of set of ingredients we can play with and who knows what that will unlock. That's the whole idea.

**Tim Scarfe** 00:26:29

So serendipity plays an outsized role in our lives. We want to do this epistemic foraging. Is there any way we can formalize it? So, you know, we're using language models as as a proxy.

## Darwin Complete Search Spaces

**Tim Scarfe** 00:26:40

So it's ineffable. We can't quantify it. We're using it as a proxy. But I remember I I read the abandoning objectives paper by, Ken and Joel in in 2008, I think. And and they were using proxies like behavioral complexity, solving a maze and like, you know, how how many bits of information do I need to sort of like track the the coverage of of the maze. But is is there any is there any step towards formalizing that creativity?[11][12]

**Jeff Clune** 00:27:04

For me personally

**Tim Scarfe** 00:27:05

Go on.

**Jeff Clune** 00:27:05

I say throw that all out. Okay. Every attempt to formalize it ends up producing the problem we were trying to avoid. People have been trying to formalize this for decades. And then you you write down this thing, this equation, I finally defined interesting or what counts as interestingly new or I've defined curiosity, I've defined learning progress. And then you optimize that and you get all these pathologies that don't actually get you what you want. Mhmm. So going there's a famous quote from Justice Potter Stewart, you know.

**Tim Scarfe** 00:27:33

Yeah. Know when I say it.

[11] Novelty search vs objective-based optimization — paper
[12] Novelty Search vs Objective-Based Optimization — paper

**Jeff Clune** 00:27:34

Yeah. He was talking about pornography. He said, don't know how to define it, but I know it when I see it. Yes. I feel like the same is true for interestingness. We don't know how to define it, but we know it when we see it. And that's been a problem up until now, now we have something that can help us see and that's language models. So rather than formalize it and write down the math, personally, my own research taste says, let's ask the model to do that job for us. Be the proxy for human model of interest in this and let's go. Let's see what happens. Now that still leaves aside I'm still interested from an open ended perspective and the whole other type of open endedness, which is could you create an open ended system without relying on something that's trained on human data? Because in some sense, that's cheating. So I consider this the hard and the easy problem of open endedness. Now, the the easy version is we'll use all of human data distilled into a model of interestingness as a catalyst to allow us to to go and keep doing interestingly new things. Eventually, I hope to myself or the community gets back to creating an open ended system that doesn't rely on that. But, you know, for now, we've got an orchard of new low hanging fruit that we can run-in and grab because we can finally have the algorithms just, you know, have that gestalt common sense. I know when I see it sense of that is boring and that is interesting.

**Tim Scarfe** 00:28:47

So you've been on this very interesting intellectual journey. Mean, of course, you'd you wrote that landmark paper AIGA's in 2019 and you're of saying there's a there's a hierarchy here. Right? So, you know, there's there's having diverse solutions, having diverse goals, and having diverse environments. Can you and and maybe just before you answer that, like, what what does it even mean? Like, what does intelligence mean when we're talking about this very open ended, diverse way of thinking about things. Does that in any way change your definition of intelligence?[13][14]

**Jeff Clune** 00:29:16

It does. You know, it really causes you to stop and think about what it means to be intelligent. And, you know, of course, coming from a philosophy background, you can look through the animal kingdom and recognize that there's not some single definition. There's lots of different forms of intelligence that we see even on our planet and who knows what would exist elsewhere in the universe. And I think that to some extent, you could even have open endedness without intel-

[13] AI-generating algorithms approach to AGI — paper
[14] AI-Generating Algorithms: An Alternate Approach to AGI — paper

ligence. You know, you could have the never ending generation of interesting pictures and there's no real notion of intelligence in that. The system itself is intelligent in order to make it, but it's not making intelligent artifacts. And so you could almost think about the intelligence of the system and appreciate that and then maybe the artifacts themselves do or don't become intelligent. For example, Darwinian evolution produced you. Is Darwinian evolution on on earth with a planet sized computer? Does that count as intelligent? You know, people did for millennia. People thought that it did, which is why many people believed and thought that was the best argument that there must be a God. Right? This was the argument of intelligent design. Nothing else could have produced a frog or a 3 toed sloth but an intelligent designer. And then we realize there's this algorithm that did it. Are we willing to now say that's intelligent? Maybe. But if you do, that's a different notion of intelligence than like a thinking entity that has words in its head, is our typical notion of the word.

**Tim Scarfe** 00:30:36

Yes. You know, Adam Smith said there was a hidden hand of the market. But in in a sense, you're you're saying there's a hidden hand of of evolution. But quite an interesting distinction, So generating lots of pictures and and pick breed is a great digression later maybe. So first of all, do you do you think that evolution should be seen as an agential process? And do you think agency is the dividing line between intelligent things or not?

**Jeff Clune** 00:31:01

Evolution does not have an intention. It doesn't have a goal. It's just kind of a very simple algorithm that's just happening. But there's still something I don't wanna use the word magical because I that's simple that implies that it's nonphysical. There's that there's it's nothing supernatural, but there's something awe inspiring and something magnificent. And and to my view, something we don't yet understand about exactly how it does the amazing things that it does. And when this is actually where I started my career. Alright. Evolution produced all this stuff. Could we just code up a little version of an evolutionary algorithm inside of a computer? Maybe put it in a somewhat rich world, hit play, sit back, grab some popcorn, and wash the fireworks. That we people have been doing that for decades. Nothing interesting happens. Okay. That's a little harsh. Something interesting happens for a little tiny bit of time and then it kind of converges and gets boring really fast. So at the beginning of my career, we did not have any algorithms that were worth running for more than a couple hours. For my I kinda set a goal of my life and this comes originally from Jean Baptiste Moray, which is like, could we create an algorithm that we would wanna run for billions of years? And I've been fortunate enough to see us get better at this as a community. So

we now have algorithms like AlphaGo that are worth running for a couple weeks, maybe even months. GPT, you know, with a lot with enough data is worth running for months on huge numbers of GPUs. We're into the months range. I still don't think we have an algorithm that's worth running for more than a year. And so it's not just the simple Darwinian algorithm. There's something else. And that something else is kind of what people like myself and Ken and Joel and Tim and all the other people that are in the open end to this field are working on. What are the key ingredients that we can put in and finally try to kick off this chain reaction that never ends? And so far, we haven't done that. And that's why I think this is 1 of the grand challenges in all of science.

**Tim Scarfe** 00:33:00

You've spoken about this concept of Darwin complete

**Jeff Clune** 00:33:03

Yes.

**Tim Scarfe** 00:33:03

Which is very interesting. Can you can you explain what you mean by that?

**Jeff Clune** 00:33:06

Sure. Yeah. So if you're trying to create an open ended algorithm, you want it to innovate forever. And we've we've tried to do it in simple settings like, let's generate in Poet, for example. Let's generate obstacle courses forever. Even if we succeeded, that thing would only ever generate obstacle courses within this 1 little tiny video game that it was set into. Similarly, you know, AlphaGo, maybe it innovates forever, but it's stuck within the notion within Go. It can only become a good Go player. Well, if you really want a process that's open ended in the way that Darwinian evolution is, that produces all the marbles we see on Earth, it has to exist in a vast search space where it could produce obstacle courses and logic problems and the game of Go and competitive, you know, coevolutionary arms races, cooperative games. It really should be able to produce almost anything. And so when I was writing the AIGA paper, said, hey, if we really wanna produce something that is open ended, then it has to be able to be in a vast search space where almost anything is possible. And I said, what is the most wide open search space imaginable? You know? And I tried to think what that was and I said, well, actually, we can even define what that would mean. And I said, it should be a system where any possible simulatable or computable environment existed. And I said that would be as as broad as it can get. And I said, well, we should give a name

to that. And so just like Turing complete just like Turing invented the idea of a Turing complete computer language that can compute anything. I said, this is kind of very similar. It's like an environment search space where you could have any kind of environment and in a nod to the amazing magic that we see in Darwinian evolution and all of the diversity it produced, I said, let's call this Darwin complete. So it's a search space that literally contains any possible environment that you could have in a computer, and the idea is that you could simulate all the diverse environments you see on Earth and beyond that. So that's what Darwin Complete means. Now

## Code Generation vs Neural Networks

**Jeff Clune** 00:35:00

fun historical side note, if you don't mind. When I was writing that paper, said, can I conceive of anything that is Darwin complete? And I was walking around the the Uber AI offices and what occurred to me is that the most general environment, the environment that literally could simulate anything is not any particular simulator we have coded up for a video game or physics simulator, actually but is a neural net itself. I said, imagine if you had a really big neural net and you it it produces like a state like an observation. It could be pixels, maybe sound, smell, touch, whatever. So produces a sensory experience. Then you take an action. You give the action to the neural net, and it just steps the world and produces the next sensory experiment experience forever and ever and ever. And since it's a neural net and that can represent any function, it can now represent any possible environment or or challenge imaginable. And I was like, that seems crazy. Right now, we definitely don't have the ability to make that. And I said, but I actually think neural nets are powerful. We know about like, you could basically you can train them to do almost anything, scaling laws, although not yet named that. It seemed like, you know, more computing data produces better and better models. I was like, I actually think it's possible 1 day that we could create a neural net that literally is the entire world. When I wrote that though, it felt emotionally impossible or way far out in the future. I actually thought I was taking career risk and people would think I was a crackpot for suggesting this is a thing, but rationally, I felt right. I knew it was possible and I if something feels possible in the field of AI, it's probably gonna happen. So I decided to put it on my paper and take the risk and 6 years later, 2 weeks ago, we put out the Genie 2 project which literally is a neural net simulating entire three-dimensional worlds that you can act in, explore, play, and you can ask for any world you want. So there it is. Within 6 years, you got that. Now, about a year or 2 ago, I was in I was in class teaching and 1 of my students said something that caused me to

realize there's a second Darwin complete representation and that's code. And so we did the Omni Epic paper where they basically it can generate endless environments forever where it literally writes the coding environment And now it can write code for its own obstacle course within a particular simulator, a new level within a particular video game, or it could write the simulator itself to simulate any possible world or logic problems or math problem, anything into that computer. So right now, these are the only 2 Darwin complete search spaces I know of. And independently, Joel Lehman kind of stumbled upon the same idea and we had fun connecting those dots.

**Tim Scarfe** 00:37:35

Yes. Very nearly Darwin complete. So it's amazing actually, folks at home should check the Omni epic paper. But, yeah, so it can it can generate code and, you know, 1 might be kicking a football through goals or something like that. 1 might be jumping over a platform. And the amazing thing is you get this phylogeny. So then you do another 1 which actually requires intermediate skills that you've learned earlier and it just goes on forever. It's absolutely beautiful. 1 very interesting point is that, for example, Kevin Ellis, he's exploring transduction versus induction. And and I think what he actually means is program space versus, like, neural space. So solving the art challenge, you could, on the 1 hand, just predict the solution directly or you could generate code. And I like the code idea because, you know, Python interpreters, you know, it it's a 2 incomplete language. It feels more compositional. Neural networks struggle with copying and counting and lots of silly things like that. So I have a deeply held instinct that generating code is good for the same reason generating language is good. Right? Because there's something special about language. What do you think is a difference between those 3?

**Jeff Clune** 00:38:40

Yeah. I almost liken it a little bit to well, 1 way to think about it is kind of like system 1, system 2, if you're familiar with the Daniel Kahneman thinking fast and slow. So code very much represents something like we are intentional. We're thinking through the logic of an algorithm. We're kind of thinking step by step. We're laying out a plan, and we're executing it. And so oftentimes, even for ourselves, if we want to do some big complicated, basically, algorithm and wanna run it, then it's better write code and let that go. Similarly or on the other side, actually, I mean, in contrast, system 1 often can handle the fuzzy messiness of ill specified things, hybridizations of different ideas. Like when I'm playing hockey, for example, I'm constantly inventuring subtly new motor commands to deal with the situation I'm in and hybridizing things that I've learned in my past. It would be impossible to write code in my view that could control a body to play hockey because that's just not kind of the right policy representation to deal with motor commands at high frequency and speed. So And I think and I think the same is true probably in some mental tasks. You're kind of feeling your way through like a conversation or or a party, you know. You would not want to write a code to handle dinnertime conversation, you know. And so our brains and the kind of neural nets can handle the fact that it's not well specified what we're trying to achieve. The concepts are somewhat ambiguous. The conversation could go in different directions. And so I think basically there's kind of almost like different types of intelligence and in different situations you want to use either. And so I think the best system knows how to code and it kinda does the thing that humans do with their words and language and also their muscles, And it kinda chooses when to take those different types of intelligence off the table.

**Tim Scarfe** 00:40:31

There's such an interesting interplay between the 2 worlds as well. I mean, I spoke with Laura Ruess and she's got this great interpretability paper, you know, talking about how code is unreasonably effective for making language models do reasoning even when they're nothing to do with code. So there's some kind of connection between the 2. And it also makes me think about whether language is a system of thought or communication because we do this kind of improvisation, don't we? And then we we establish terms as as language. But it actually came from here first and then this is the pointer. But maybe this is a good time to talk about your thought cloning because you did some really interesting work there kind of binding them together.

**Jeff Clune** 00:41:04

Yeah. So thought I love the idea of thought cloning. So the basic idea is right now, 1 of the most successful recipes in all of machine learning right now is I'm going to take a robot conducting a bunch of actions or a human doing a lot of tasks like playing a video game or a game of Go or something. I'll take all that human data, which was in this situation, what action did I take? And then you just collect that data, and then you train a large neural net to take that the the action the human took when it was in that situation, do that at scale, and you have all the modern AI. So this is what GPT is doing. Right? Based on these 60 words, what's the 60 first word and the 60 second word[15][16]

## Thought Cloning and Behavioural Learning

**Jeff Clune** 00:41:40

and the 60 third? Right? And my my team at OpenAI for Minecraft cloned out a bunch of humans playing Minecraft at scale, a massive scale, and you have an agent that can go play Minecraft. So we're really good at saying to a to a a system like a robot, take the actions that a human took. But what's missing, and 1 of the reasons I think robots don't plan very well, they don't generalize out of distribution, they don't deal with novel situations, has because we've given them the system 1 muscle memory. Right? Like right now, if I grabbed a pillow and I threw it at your head, you would just duck and you would do it before any conscious stuff happened in your brain. Right? You've got that muscle memory and that's basically what we're teaching these things to do. Similarly, if I ask you, you know, to be or not to, or if I ask you 2 plus 4, you don't actually stop and think through the algorithm of addition, you just know what 2 plus 4 is. You kinda have muscle memory for certain cognitive tasks. But if I asked you, you know, to do a 5 digit multiplication, digit by 5 digit, you wouldn't just know the answer. You'd flip over to system 2, the thinking part of your brain and you'd sit there and you'd go through an algorithm and you'd think through it. So the problem is is that in behavioral cloning, we're only kind of trying to get the muscle memory into these agents. What action to take immediately without even thinking about it in this situation. But what what we want is the thinking part. Inside your brain, while you are playing Minecraft or doing math or like taking a quiz or a test, there's a lot of thoughts happening in your brain before you take the action of writing a word or speaking a word or pressing a button on a game controller. And so the idea behind thought cloning is if we really want intelligent

[15] Video PreTraining (VPT) — paper
[16] Video PreTraining (VPT) — paper

agents, we don't need to clone their actions, their behavior. We need to clone the thinking also. So what we do in the ideal case, we would have humans taking a bunch of actions in any domain. Imagine playing Minecraft, for example. And we have the actions they're taking, but we also have the thoughts that are going on inside of the head of the agent. Alright. First, I'm gonna get a house to get a to build a house. To build a house, I need to collect wood, and then make a crafting table, and then I need to get an axe, and, you know, and then you start, okay. Let's first, let's go get that wood. I'm gonna go over here. Oh, there's no trees over here. I'm gonna replan and look over here. Oh, there's 1. Da da da. And so that running commentary in your head, the thinking part, we want that too. And so if we had data of all the thoughts happening inside of a human and the actions they're taking, now we can do what we call thought cloning, which is we train a neural net to both get in a certain situation, produce the actions the human takes, but also produce the thinking that's going on inside their head, and have the actions they take actually condition on the thoughts. So it plans, it reasons, it thinks, then it acts, and we want we ought to kind of train the agent to do both of those from human data. So, you know, since we have an academic we tried a little bit of this in my team at OpenAI before I left, but we ran out of time. We were gonna use the closed captions on on online of, like, the narrators and these online video game things while they talk out loud. We were gonna train on that. That kinda worked. We didn't have enough time. So my academic lab here at UBC, we decided to do it on a small toy dataset, and we had basically a bot taking a bunch of actions. We had the low level actions it's taken and, like, a little planning algorithm in its head. And we trained it to kind of simultaneously do the thinking of the planning algorithm and take the low level actions. Now 2 cool things about this. 1, if you train to imitate the thoughts as well as the actions, you learn way faster than if you just train on the actions, so it's way more sample efficient. Also, as we had hoped, an agent that learns that knows how to think is better at adapting to new situations. So the further we put this thing out of distribution, the thought cloning agent is better 0 shot at handling that. It can replan. It can put together things that it learned and thought about from the from its training in a new situation, and it's faster at learning. You give it a couple sets of learning, and the thinking agent is much faster at learning. So pretty much everything we hope for just fell out of the data the second we did these experiments. I wanna make 1 final comment and that is 1 of the hardest challenges in my whole career, including an open endedness, is how do you get agents that can explore well? Right? This goes back to our sense of interesting, our sense of curiosity. If I put you in a big complicated new building, you would do such a marvelous job of exploring it. Our robots, our AI agents suck at that. It's actually been 1 of the big themes of my career. Well, 1 reason why behavioral cloning at scale, like we did in VPT at OpenAI for Minecraft, does not solve this problem is because most of the data we have are experts. They already know

how to play Minecraft or do math or explore a kitchen, and they immediately go find the kettle or the spoon or whatever, and then we clone that behavior. What we don't clone is the newbie or the novice human in a new video game, fumbling around, exploring, trying different things, doing what Alice Gopnik note has demonstrated that children do, which is learning a sense of empowerment, how to control my environment, what causes what. We want that data. And I think if you trained at scale on if you trained at scale on behavioral cloning and thought cloning during exploration, now you have an agent that you drop in a new environment and it knows how to go explore, be curious, uncover what's interesting, and that is how you truly master a new domain.[17][18]

**Tim Scarfe** 00:47:00

Love it. And I'm a big fan of Alison Gopnik by the way. I'd love to get her on writing an article on creativity at the moment with my friend Jess and we've her a lot. 1 1 interesting thing about language is there's a great book I read called The Language Game and they described it as a kind of constructive process of improvisation. And essentially, it it's it's the mapping between our kind of phenomenal world and some externalized symbols for, you know, use for communication or whatever. Wouldn't it be great if we could get AI agents that could create their own language? So if you think about it, they have a sensorium and they they have this interaction, this improvisation, and they need to to sort of like create a language and then we could let we could let that evolve into a phylogeny. How could we do that?

**Jeff Clune** 00:47:43

Yeah. I think it's a great idea. Actually, people have tried this. All you do is you can create a neural net that can emit some sort of usually, it's like language tokens, and the other ones just get them. And then you could just put them into, like, a big environment, like a big video game and let them communicate. People have tried this. And 1 interesting interesting thing that happens well, 2 interesting things happen. 1, not much interesting stuff happens. Oftentimes, we don't know how to create the setting in which language is extremely helpful and has such immediate and obvious improvement that you get a huge learning curve. Like, they really learn to use it easily. You have to, like, really create the right

[17] Thought Cloning: Imitating human thinking — paper
[18] Thought Cloning: Imitating Human Thinking — paper

## Language Emergence and Interpretability

**Jeff Clune** 00:48:20

setting to see language emerge. But some people have caused language to emerge. But in that case, another interesting thing that happens is that we see them communicating and it helps them, but we actually find it hard to interpret. Yes. So now you almost have this like, yeah, the the challenge for like the movie Arrival. You have to figure out how to communicate with an alien race. And if I remember correctly, that Facebook situation where people claim that it would like it invented a language that they shut it down because it was too dangerous. That was all apocryphal. It didn't shut it down because it was dangerous. They shut it down because it was like totally uninterpretable. They had no idea what was going on. And so if you wanted it to be a language that we could actually eavesdrop in on or communicate with, either you'd have to train the agents with us in the loop to also communicate with us from time to time, or you have to do some sort of trick to ground their language to be, like, similar to our language. Like like, maybe if you put in a bunch of GPTs and you let them communicate, they probably would start in English. And if you did nothing, they'd eventually invent their own language that we can't understand. But you can do some tricks from machine learning to basically say, hey. You have to keep it similar to our language. I think that's super profitable and interesting. And I guarantee you that the open ended algorithms of the sys of the future will have many agents running around and, you know, communicating with each other, inventing their own language, using it to accomplish hard tasks, And then we either keep it similar to English or whatever human natural language, or we've gotta, like, get some AI translators on on board to help us figure out what they're saying.

**Tim Scarfe** 00:49:44

Imagine we did create a collective intelligence of agents, and they developed their own language, and they were communicating with each other, and and we could monitor it. When would it worry you?

**Jeff Clune** 00:49:55

Oh, there's so many things that worry me. I think we're gonna get to AI safety later. The language part itself is not inherently worrisome, but the behaviors could be. Like, if we happen to eavesdrop on them when they didn't realize we were listening and they said something like, hey. We're gonna try to jailbreak. These these the the humans are trying to they don't think we're that clever, and they're trying to keep us under control. And, like, let's figure out a way to, like, break out and take control from humanity. Yeah. I'd be pretty worried.

**Tim Scarfe** 00:50:23

It's almost like, you know, that we've created this ecosystem because we are becoming a bit like neuroscientists now with these multi agent systems. We'll on to your AEDIS paper in in a second, it's like, we've created this gnarly monstrosity that we don't understand. It's behaving in in weird ways. What what are the indicators that are of concern?

**Jeff Clune** 00:50:42

Yeah. So I actually spent almost, I don't know how many, 6, 8 years of my career working on an arc of work that we actually called the AI neuroscientist. And there was interpretability work where we basically were in the same if we found ourselves and we find ourselves in the same situation of neuroscientists. Here's this big complicated artifact. It thinks really well. It does impressive things. We wanna understand how it works and it's complicated. So real neuroscientists have that problem with you and what try to figure out what's going on inside of human human brains. And AI scientists now have the same thing, like GPT 4, big complicated. We don't really know how it works if we want to. So we did a huge arc of work trying to address that. I think it's super fascinating. If you want, we could talk more about it. But there is a related phenomenon, which is specifically trying to say, can we use techniques like AI neuroscience or other interpretability techniques to detect when something is like a safety risk. And I had lots of conversations with people at NeurIPS about this. There are some ideas that are promising. I'm not some of them are good and we should do them and they will help. Nothing is a guarantee. So like 1 of the best ideas is if you can invent the equivalent of a lie detector, that is quite good for AI safety because you might detect that it's deceiving you. Another idea which I love and we put in our thought cloning paper is if you train it on human language, like English, for example, and it learns to think in English about what it's going to do, then you can watch it think. You know what its plans are. And so if it's planning on doing something unsafe, you can stop it. So if it says like, you've asked me to get this this person to the airport as fast as possible, so I'm gonna just blow ev-

ery red light or run across, you know, children walking at the school crossing, you say, alright. We have to stop your behavior. And we did a little experiment that said after training, we could declare, hey, touching red things is considered unsafe. So every time you say you're gonna touch a red thing, we just halt your program. And you see the system go from touching lots of red things to virtually never touching a red thing. So it works. And if the system if you could keep it thinking in in English, for example, and you could watch it thinking and it didn't know you were watching it, then maybe you could see it to start to say, hey, I actually wanna deceive these humans. I'm gonna like play nice during evaluation time. Once I get out, really once I get out in the wild, I'm gonna be unsafe and try to rest control from humans. Obviously, you you're gonna shut that robot down. Right? But it's not perfect because the system could learn maybe even by reading our own paper, the paper we put on archive on thought cloning that this is possible and that if it really wants to get itself out of jail, then it fakes its thoughts in some way, doesn't have the thoughts that we're afraid of during evaluation having these kind of surreptitious plans. For example, I actually think a future is coming not too far in the future where there's gonna be a technology to read somebody's brain via like external EEG. And you might ask a terrorist, where did you place the bomb? And it will think it, think of the mental picture of it, and then bam, you've got it. Right? And I think the countermeasure is they're gonna learn not to have that mental image. Right? Like they're gonna learn countermeasures and the AI can also take countermeasures for a lot of this like interpretability safety neuroscience stuff.

**Tim Scarfe** 00:53:55

Couple of things on that. I spoke with Neil Nanda. He's he's a famous MechInterp researcher. And he was kind of saying that at a certain level of model complexity, you know, it could it could facade our measurements just just as you're saying. So it can kind of show us things that that that we wanna see. But it comes back to this kind of theory of mind, I suppose. So Nick Chatter wrote a book, The Mind is Flat. Know, a lot of connectionists just think that it's a mistake to think that these things have consistent beliefs, desires, intentionality. They're a bit like automata. And just depending on the situation, they do very different things. I mean, Murray Shanahan studied it with the 20 question game where you ask language models, you know, like, can you you guess something of 20 questions? And when you do the analysis going back, it was actually thinking of something completely different earlier on than than what it ended up with. So do these things actually have consistent beliefs and desires?

**Jeff Clune** 00:54:49

I think we do ourselves a disservice to look at the current flaws and assume that they're gonna remain current flaws. So my my meta comment here is even if the current systems don't, then the future very nears

## AI Safety and Deception Detection ●

**Jeff Clune** 00:55:00

near term systems will. And I think things like this are getting better and better as they get smarter. And they'll continue to get smart, and they will become very consistent. And so all of these kind of, like, point to current flaws and assume that it's not gonna get more capable and dangerous, I think that that is a I think it's unhelpful. I think it's inaccurate. I think it's actually dangerous because you're not recognizing the rate of change, and you're kind of almost, like, latching on to where the you know, a belief that there's something special about us that they don't have, and therefore, they're not gonna become dangerous personally. And so I think that the current systems might have some of these problems future ones won't even if the current methods never produce it, and there are many scientists who think that. I think as you switch to more powerful methods like taking a pretrained GPT 6 and dropping it in a giant diversity of embodied tasks and simulation, and you let it learn via trial and error and reinforcement learning to accomplish a lot of tasks, of course, eventually, it's going to gain consistency for the same reason that evolution produced a consistency capability in you. It's very useful to be able to have a consistent plan and act against it.

**Tim Scarfe** 00:56:12

We were speaking about using language models are great because they encode notions of interestingness, culture, priors, and and so on. And to a certain extent, we can align them. Right? So, you know, if if you build a multi agent system now using 1 of the major frontier language models, it's not gonna delete all of the code on your hard drive. Right? It's there's there are certain guardrails in place.

**Jeff Clune** 00:56:31

Not really. It definitely might definitely might delete everything on your hard drive.

**Tim Scarfe** 00:56:35

**Well, I I use OpenInterpreter, for example, and it it's actually surprisingly good. So, you know, if you say something like delete my entire hard drive, it will say, no. I'm not gonna let you**

**Jeff Clune** 00:56:43

**do You've got some extra checks. But a lot of people in this world right now, myself included, are letting the language model generate code and just oftentimes getting lulled into a sense of it's pretty good and just running that code without checking it. And I myself don't have the guard I need to learn from you to put those guardrails on my system.**

**Tim Scarfe** 00:57:01

**Yeah. Yeah. But do you do you think that that is a significant safety measure just having building multi agent systems on top of language models that are aligned?**

**Jeff Clune** 00:57:11

**I why would that make a thing safer?**

**Tim Scarfe** 00:57:14

**So we can do RLHF, you know, so so the the language model will have a notion of ethics. It will have a notion of code, which is unsafe and so on.**

**Jeff Clune** 00:57:22

**Yeah. In general, I think that alignment is better than not having alignment. I am quite happy with the progress in alignment in recent years. The and so I think that it's meaningfully better than not having it. However, there's at least 2 problems. 1, it's very easy to ditch the alignment if you allow the system to keep training. So if you open source a model, anyone can just train that RL edge right out of the system, which is quite problematic. If you take that thing and you kind of, as you were suggesting, put it in a reinforcement learning situation where its job is to accomplish some task, some new task, and it's not simultaneously being trained to also do the alignment job, then it'll if it is in the if it helps to accomplish the target task to get rid of its ethics, it will rapidly do that because that's what RL is good at. It's figuring out how to solve the task at hand. So there's all sorts of ways that alignment can be undone. That worries me. Another thing that worries me a lot, and now we are getting on to AI safety a little bit, is that there are 2 big problems at least in AI safety. 1 of them is unintentional harm. So I am a really noble human. I'm trying to to, you know, ask an AI to do something. I ask it, for example, to solve climate change, and the AI does what I ask for, not what I want. And it says, oh, yeah. Okay. The the cause of climate change is humans, and so I'm gonna kill all humans. Something like that. Right? There's many cases like this. Alignment allows it to do what we want and not what we say, and so a a little properly aligned mile wouldn't do that. Right? So alignment makes it better and better and better to minimize unintentional harm when so if you have AI in the hands of good actors, it will do good. The other thing I'm worried about, though, is AI in the hands of bad actors. So pick your villain. For me, it's Putin right now. I don't want Putin, or I don't want, like, a school shooter like the people who shot up the Columbine the high school in Columbine who wanted to kill as many people as they could. I don't want that person to have very powerful technology. Like, the school shooters in Columbine could only take out however many people they did, but eventually the police showed up and that was the end. If they had had the ability to make a bioweapon that killed all humans or if some racist neo Nazi has the ability to make a bioweapon that kills all black people, some people would want to do that. And as you get better at alignment, it work it's a double edged sword because now the model is aligned to that person's ill intent, depending on what we mean by alignment. And so what worries me a lot is powerful tools and AI in the hands of bad actors. And RLHF, as I said, could be undone or could actually help that person accomplish their goal depending on what kind of alignment we're talking about.**[19]

[19] Go-Explore: A New Approach for Hard-Exploration Problems — paper

**Tim Scarfe** 01:00:17

Do you think there's an analogy between this type of technology and and nuclear weapons?

**Jeff Clune** 01:00:22

I think there's many analogies. Some of them are great and some of them don't work as well. I think there's an analogy in the sense that it is it it is not yet currently, but will soon be extremely powerful, extremely dangerous, could be world ending if we're not careful. And we want to marshal the kind of regulation, thoughtfulness, and prevention, and monitoring and diplomacy that we brought to nuclear weapons to the development of powerful AI.

**Tim Scarfe** 01:00:53

What do you think about open source AI? Should we regulate it?

**Jeff Clune** 01:00:58

I do think we should regulate it eventually. I'm not I'm not that concerned with exactly the current tools that are being released right now. But I think this is a this is a point of view that many of my colleagues don't share. But I actually think it's quite dangerous to open source extremely powerful technology because you're giving it to everyone and if point 01% of people want to create some horrible bioweapon that kills, for example, 1 race of people or all people, now everybody has it. So we don't open source the recipe for making nuclear weapons. We don't open source the recipe for how to make smallpox

## Alignment Limitations and Open-Ended Risk •

**Jeff Clune** 01:01:40

or a race specific smallpox. Similarly, I don't think we should open source the recipe for AGI. I believe that there are alternatives like having a coalition of democratic governments make powerful AI, let people use it, and get the wealth and the benefit from it without having access to the weights because we don't want them to ditch the RLHF that's preventing it from doing evil.

**Tim Scarfe** 01:02:07

How do you reconcile this with your views on serendipity and and open endedness and and so on? I mean, you might argue that having a vibrant open source community is the best form of stepping stone collection and and therefore the best way to find a solution.

**Jeff Clune** 01:02:22

It's a really good question. And if you want the fastest possible scientific discovery of the most powerful capable AI, then doing the things I just said to make things safer would slow that down. So you wouldn't do what I'm saying. You only cared about progress. So the reconciliation is we don't necessarily want the fastest possible progress towards AI. We want to make sure to do it safe, and I'm willing to pay a hit on the rate at which we we finally get to the thing to make sure that we do it safe because, literally, the fate of human civilization is at stake. And so the principles are right. That is the fastest way to make discoveries, we don't necessarily want to follow the principles without being able to do it safely. And if there's there is a tax to doing that, and I'm willing to pay it.

**Tim Scarfe** 01:03:13

Ben Joe said recently in an article, something can't remember the exact quote, but something along the lines of he feels a sense of despair that much of his life's work could potentially lead to ruin. And I I just wondered, like, do do you do you feel a similar kind of thing that that so much of your work is about building the next generation of a gentle AI? And by the way, Benjo actually said that he thinks what we should do is not make agential AI, make make oracles and and tools. But how do you think about that?

**Jeff Clune** 01:03:40

So on the question of do I, you know, do I regret my life's work, for example, or do I feel despair about the development of AI? I'll say that I lose a lot of sleep over this question. I think very deeply about it and I talk to a lot of colleagues weekly about this question. I will say it's very complicated. I don't think there's any short simple answer. My logic roughly goes like this. I think that it's probably inevitable that we will make AI. I do have almost 0 faith in humanity's ability to not invent a technology that's this economically valuable and militarily capable, mostly because of a tragedy of the common sort of collective action problem. If we don't do it, somebody else will do it. So if that is true, I would much rather it is done the right way, safely, by people whose values I share, and where we get the values into the system via things like RLHF. So the AI itself is aligned. And so I actually feel there's an obligation to do this right. If I had a magic wand and I could pause AI and then work on safety first and only do it if we're safe or maybe not do it at all, I would waive that wand. And I've signed the letter calling for a pause, and I've worked with Bengeo and Hinton and Kahneman and other people to, like, publish articles saying that we need to take this stuff very seriously and be very safe. But given that I don't have that magic wand, given that it is inevitable, I think that we need to build it safely and that what worries me is that if all the people who are thoughtful and ethical enough to be concerned stop working on it, then what is left? It is going to be developed by the people who aren't concerned or who actively don't, you know, have have ill intent. And so we have to, barring some major intervention in collective action, we have to do it, and we have to make sure that it's done well and it's done safely. Simultaneously, I think myself and many other people, we talk a lot about the downsides and the risks. It's easy to forget about the tremendous upsides, you know? I right now, there are people saying goodbye to their children because their children have cancer or some incurable disease. There are people saying bye to loved ones of all ages. There are people suffering from hunger and poverty. On, you know, excruciating human suffering and pain is rampant in this planet. This technology has the ability to eliminate all of that. I believe firmly that we can cure all disease. We could solve death. We can dramatically increase the quality of people's lives, their GDP, eliminate things like hunger scarcity if we get the societal distribution questions rotten done. But technically, a lot of this stuff is possible. And so Dario Amadai wrote a beautiful essay called Machines of Loving Grace. I recommend that to all of the listeners. It does a really good job of walking you through and making you dwell on these positives. They're easy to list quickly, but I think it is actually important to walk through and dwell in all of these tremendous upsides. We want that. And as Dario says in the essay, the reason why he talks about AI safety is the same reason I talk about it because that's what we need to get and figure out in order to unlock all of that great stuff. And so I don't want

to die. I don't want my children to die. It would be great if we could eliminate though that from the world and make death optional. And the only path to doing that and many other great things for humanity is through technology and AI is the most potent technology to do that. So if you throw together the inevitability, the fact that if we don't do it, it will be done and probably by people who are less careful or maybe less scrupulous and the tremendous upside. I have ultimately concluded for myself, I'm okay working on it, and I want to work on it, but I need to and must make AI safety a first class citizen in what we do, and I put it, you know, out there as often as I can into the community that everybody should be thinking about it and trying to infuse their daily and weekly work with AI safety related with AI safety in mind, pushing forward the science of AI safety, the advocacy, telling politicians and regulators. I've met with top politicians and advocated this to them in in closed door sessions as well. Basically, I wanna do everything I can on the safety front, but I still also do feel like it is a good thing to try to make safe AI and then reap the upsides while minimizing the downsides.

**Tim Scarfe** 01:08:19

Thoughts on

## Open Source AI and Global Governance

**Tim Scarfe** 01:08:20

this collective action problem. I don't know what what you're advocating for here, but do we need some kind of global global governance structure?

**Jeff Clune** 01:08:28

Yes.

**Tim Scarfe** 01:08:29

Tell me more.

**Jeff Clune** 01:08:30

The problem is that I don't believe we could do it. You know? I don't know if I don't think every single actor in the world will agree. Like, take Putin, for example. Is he really gonna agree to not make powerful AI? Even if he agreed, would he still do it in secret? I'm pretty you know, if I had to predict, I'm I'd say there's definitely some secret government sponsored AI research projects in the world right now building AI in secret for their militaries and for their spy agencies, etcetera.

So I'm super worried about all of that. So I do think we need a regulation. I think we need global governance. My personal view, and I'm not an expert in these kind of areas. I'm a computer scientist, not a political scientist or a diplomat, but something like a coalition of of value aligned governments, you know, like democracies say, we're gonna build this thing. We're gonna try to be make it safe. We're gonna share knowledge. We're not gonna open source the weights. We're gonna spread share the benefits with humanities, give the benefits to human, and we are going to monitor and suppress the creation of AI by unsavory characters elsewhere in the world because we don't want somebody to make a non-value aligned AI system somewhere. So anyone who agrees to go along, you can participate in the upsides, participate maybe even the democratic governance of of the the AI that we wield. But if you don't agree, if you're not willing to play by the rules, if you're willing to make it in an unsafe way or you wanna do it on your own terms and not be subjected to monitoring or, us knowing how you're doing it and what our LHF you're doing, etcetera, we're gonna deny you chips. Or we're gonna deny you electricity, or we're gonna do what we can to slow you down because we don't want to make a dangerous rogue AI that doesn't share a love for humanity and doesn't wanna do what we ask it to do.

**Tim Scarfe** 01:10:20

On early warning systems, of course, there was that FLOPS regulation in in in The US. Is there a slightly better way of assessing the the the risk of of foreign actors, you know, but other than just the amount of computation that they're doing?

**Jeff Clune** 01:10:35

Well, 1 of the great beneficial lucky events so far in the history of powerful AI is that it requires tremendous amount of power and compute to make. Yep. If somebody figures out the way to make this thing on your desktop, I am very, very afraid because now it only takes 1 bad actor unless the good AIs can somehow stop the bad AIs. And do we really want an AI versus AI war breaking out on Earth? Like, all of none of these features seem that great to me. So at least for now, for the foreseeable future, we can track who has the the power and energy to make these things or we can try. And that allows us to basically have a sense of who's training the biggest and most powerful models. And that's fortuitous and I think that we should do it. I think it's quite hard in places like foreign countries where they might be trying to obfuscate. But if you look at, like, nuclear weapon weapons, to go back to the analogy you mentioned, we're pretty good at monitoring who's building nuclear weapons. Like, we have a pretty good sense of what Iran and North Korea are doing, and we've slowed them down. I think that it's probably gonna look a lot like that, that we are basically stucks neck ducks netting and denying parts and, you know, things to the to the actors who want

to do this in a dangerous way. I don't see any others I I I'm open to ideas, but I just spent a whole week in there. Was asking this question over and over again. What is the alternative to suppressing the bad actors from making it? I have yet to hear anyone provide a good idea that doesn't involve some form of suppressing countries like, you know, that are countries that don't wanna play by the rules and don't wanna do it in a safe, ethical way from creating it.

**Tim Scarfe** 01:12:09

Let's talk a little bit about your ADAS paper. And I suppose, like, 1 of the principles here I spoke with Chong the other day, by the way. Absolutely amazing stuff. So maybe to sketch it out, we were just talking about, you know, Dario, and everyone should watch his Lex interview. It's amazing. It's like, you know, hours and hours. I I loved it. But his school of thought is very much that, you know, these things are just getting better and better. We scale them up better and better and better. But I have this intuition and I think you do too that there's a better way of doing it by, you know, building some kind of multi agent system. And now when we build multi agent systems, we are handcrafting it. Right? So we use these patterns like the debate pattern and the critic pattern and and god knows what else. And we're just kind of like constructing all of these different topologies of agents together. What you have done is figured out a way of automating that process.

**Jeff Clune** 01:12:57

That's right. So right now, people just, you know, people recently discovered, hey, if I ask GPT or or Frontier model to do something, it can give you an answer. But it's way better if you create some complicated I call it like a workflow, like a flowchart, you know. First ask it this, then ask it to reflect on that answer maybe 3 times, spell out its reasoning, and then maybe invoke a new language model to review that thing and write a critique of it, and then the original 1 reads the critique, updates the thing. So there's some complicated flowchart, and then eventually it spits out your answer. Okay? So some people are calling those agents. That's like the current word. I actually don't like that word because also the the the neural net that played Starcraft and Go and like, you know, a neural net that just takes actions in some video game is also an agent and they feel different to me. So in my lab, we talked a lot about it, and we've landed on the terminology, which some part of the community is using, which is to call that an agentic system. Alright. So right now, there are many agentic systems that are being developed. They're way more powerful and capable than just asking us a language model a simple question. And they don't necessarily just involve multiple calls to language models and specific prompts and patterns. They also can be tools. Like, the system could make a query out to Google Scholar or, you know, Seman-

tic Scholar and get the results back. It could use a calculator. It could write code and execute it. All this stuff. Alright. So people are currently hand designing complicated agentic systems. They work better. And then somebody else looks at the library of all the agentic systems that humanity has published so far. They have an idea. Oh, what if I modify this 1 in this way? Or what if I combine these 2 parts from these 2 agentic systems? Well, if you're familiar with my lab and our and and my colleagues, the thinking that we have, and this is what basically the thesis of the 2019 AI generating algorithms paper, is that the history of machine learning has writing that's very clear on the wall, and that is that hand designed systems become and pipelines become replaced by entirely learned pipelines as we have more compute and data. So why not apply that same thinking to the creation of agentic systems? So my PhD student Sun Ren came to me, and he said, hey. Your age AI philosophy is AI machine learning systems are better than hand designed ones. People are hand designing agentic systems and they're powerful. Why don't we just AIGA this thing and learn these agentic systems? And so we said that's a good idea. Now if you want to learn agentic systems to make them better and better, you could use classic optimization. Take a thing, try to make the better 1. If it's better, keep it, and, you know, make that better and better and better. But we've got this whole repertoire of ideas from open endedness and quality diversity algorithms that says the ultimate path toward a powerful thing is not a straight line, and it's not going to be achieved if you only try to accomplish that goal. So why don't we take the playbook off the shelf of open endedness and say, we'll take 1 agentic system, and then we'll have the system look at that and create another 1. If it's interestingly new or high performing, add it to the set of ones that we have. And then look at these ones. Generate a new 1. If it's good if it's interestingly new or high performing, add it. And we'll grow this expanding set, this expanding library of stepping stones or innovations, different agentic systems, each that are novel. They're different from each other. They do different things. And as we go, we'll get this big library of growing agentic systems. And ultimately, we'll probably discover even higher performing ones that we would if we picked a goal and tried to optimize right for it. And that is the playbook that is in so many of the papers in my lab, and[20][21]

[20] Automated Design of Agentic Systems (ADAS) — paper
[21] Automated Design of Agentic Systems (ADAS) — paper

## ADAS: Automated Agent Design

**Jeff Clune** 01:16:40

it worked, you know, basically right out of the gate on a on ADAS. So we call it ADAS is the automatic design of agentic systems. That's ADAS. And ADAS is basically saying, let's use open endedness to design agentic systems and it works great.

**Tim Scarfe** 01:16:55

I love it so much to talk about. And of course, 1 thing you you tried it on was the ARC challenge and we had Charlet here the other day.

**Jeff Clune** 01:17:00

Oh, cool.

**Tim Scarfe** 01:17:00

We'll get back to that in just 1 second. But there there is a fundamental kind of difference in philosophy here. So 1 school of thought like like Dario was advocating for is that you just keep training these things. And we see these fascinating behaviors. So, you know, like the neural networks because of their inductive priors and stochastic gradient descent and whatnot, they they start off in there's a kind of curriculum. They they learn, like, low complexity features and then features of increasing complexity. And then there's this kind of consolidation phase or grokking. And there was a paper at NeurIPS called Sedimentation, I think, where, you know, all of a sudden, the complex features become consolidated earlier on. And you you could say that that's an implicit form of of what you're talking about. Because because what we're saying is we follow Ritz Sutton. Everything that we thought was hand designed and explicit is becoming more and more implicit. And you're advocating for a for an interesting fusion, which is kind of like there are explicit hybrid components, but it's being meta learned.

**Jeff Clune** 01:17:55

Yeah. So you you raise a really good point, which is you could try to create a system where whatever the logic is in an agentic system, first, I should do this, then I should think about it, then I should maybe, like, ask a friend, get their review. Maybe then I do a web search, and I bring that. Yeah. Whatever that flowchart is, all of that could exist within the the original transformer. Right? But what we've learned is that, you know, chain of thought works better than the the the original model that you're you're adding chain of thought to. Same with reflection. And the same thing is true of humans. Like, I myself could probably go to a consultant and say, hey. What's a really good pipeline for when I write a paper? Okay. First, make an outline, then do this. Ask some friends, get their feedback, expand expand each thing out, then do a pass, then go, you know, for a week and don't look at it, and then come back with friends. Whatever. There's some recommended thing that we're doing with the brain that I have in some flowchart that makes it better than the brain I have. So in theory, I could just have that at all of my brain and kinda do it and then 1 shot the essay, or I could have that kind of all internally within my activations. But we've seen with our brains and with these models that it's hard to get all of that right into the weights of the brain. It's easier to train that thing and then do stuff on top of it. And so I do think there's a nice and interesting bit of lesson here, which is probably a lot of the stuff that's currently being added afterwards in the agentic system could be distilled back into the original thing, and then you could just repeat the process. And so the ADAS is always giving you, like, a 20% lift on the base intelligence. But the base intelligence can grow, and you just always get that 20% lift. That's my hot take, and nobody's ever asked me that question before. Think it's quite profound and interesting, and my mind is spinning with ideas for for research that we could do in this direction. And so I think we just basically don't know the answer. What should we be doing bit should we be bitter lessening, taking intelligence, and putting it into an agentic system, or should we just be going all you know, just making the core thing instincts are that we're always gonna have a role to play for these kind of like agentic systems on top of the base intelligence.

**Tim Scarfe** 01:20:10

There there are so many directions to go in. I mean, we could talk about computational differences in in the 2 paradigms. I think having a system of LLMs has very interesting computational properties compared to 1 LLM. Yeah. There's this thing you were talking about about the the phylogeny. That there's a great example, you know, the lottery ticket hypothesis which is that you you need the blank slate, and stochastic gradient descent needs all of these useless parameters. And then when it's trained, we can strip it all away. And it's a similar thing, you know, here. We we we can create this agent architecture, but then, you know, it's not so trivial to stick that thing back into a neural network afterwards.

**Jeff Clune** 01:20:45

Yeah. 1 thing that mean, just a simple example of why an agentic system might be always helpful. I mean, not not always, but likely, is diversity. So when I write an essay, like the AIGA paper for example, I don't just write it and post it, you know. I write it and I send it to people like Ken and Joel and my other colleagues and friends and I get their feedback. They give me really good feedback. And I actually seek out diversity. I ask different people with different backgrounds and opinions, philosophies, and styles to get all of their feedback. And the more feedback I get, the better that essay will be.

**Tim Scarfe** 01:21:19

Yep.

**Jeff Clune** 01:21:19

Now, in theory, if I was really smart, I could have a, like, a little bit in my head that says, act like Ken. Use your model of Ken to review your own paper and improve it in light of that of Ken. And I could have Joel in my head and Ken in my head. I could have you in my head, and I could have skeptics in my head, but that's actually quite hard. So probably it's easier to have a diversity of agents that are trained and trained on different data or for different purposes, and then kind of have an ensemble or like a peer review of these other things to give me that feedback, and that is the kind of thing that agentic system can get you really well. That would be hard to kind of distill all into the hard but not impossible to distill into the base model.

**Tim Scarfe** 01:21:58

So that's a very interesting thought. So, you know, again, love love Kenneth Stanley. He he was talking about we are agents. I actually think there's a big link with agency with his with his philosophy as well, you know. So we follow our own gradient of interestingness. And we can go for many, many steps in isolation, you know, discovering interesting new knowledge before we share it with the collective. And I guess, is there a difference between a kind of agent system where we have continual learning and we have active inference and the agents can actually kind of like mutate and adapt over time versus the blank slate version where we just have autonomous agents that are, you know, just using like a frontier language model?

**Jeff Clune** 01:22:36

Yeah. I think ultimately, we're probably gonna want continual learning. As a side note, I worked on this problem for years, and it's just a problem that we just have we're we're nowhere near solving.

**Tim Scarfe** 01:22:47

Oh, tell me more.

**Jeff Clune** 01:22:48

Well, it's just it's just true that we just have no ability to have a model continuously learn and have good things happen. So every single model that's powerful like GPT 4 or, you know, pick your frontier model, they train it on a chunk of data, and then they stop. And maybe they do some supervised fine tuning, and then they stop. And then maybe they do some RL fine tuning, and then they stop. What they don't do is say, hey. From every conversation you have with all of your users all over the world or every interaction you have, just keep learning forever. And, like, that's for the dream that I've had and many people in our field have had, and nobody knows how to do that. And so, basically, it's so hard, and we've failed so completely to to to master this thing that biology has figured out that it's really not on most people's radar. Like, it's not something that most frontier labs are pushing on. At some point, I think we'll come back to this.

**Tim Scarfe** 01:23:42

Is it just the the architectural complexity? I mean, I could imagine 1 way of doing it might just be having, like a a 2,000,000 context window and and the context window just grows of experience or we could do test time compute, we could have an intermediate personalized model which generates context for a big large model or something. I mean, it feels like we could do something in that direction.

**Jeff Clune** 01:24:00

It is true, and it is interesting, that 1 of the things that has fortuitously taken the pressure off of the need to figure out continual learning is in context learning. Yep. And in particular, large context. Because if your context was large enough and your in context learning was good enough, then maybe you don't need to do continual learning because you can deal with it all in context. For example, if I wanted an an AI assistant, if I wanted, you know, a frontier model to be my AI like in the movie Her, like a like a an agent that works with me through my whole life gets to know me and my preferences and can talk to me about things that happened in the past, well, either it has to learn as it goes about my preferences over the course of years, or it has to have such a big context that all of the video and all of the audio and all of the text from every interaction is just sitting there in it. But even were that true, you might want that her agent to be training simultaneously on the interactions of all the humans on Earth. And I just don't think we're gonna train a model that's so smart that it can then in context as learning through everything else. Like, new stuff happens, and there's so much good data out there. At some point, you wanna transfer that from the context to the weights. In your brain, you do not only do in context learning. The weights of your brain are updating all the time. And all of long term memory, for example, even some forms of short, you know, not super short, but short term memory, all of that's going into the weights of your brain. So biology has figured this out and you don't when you're learning something now forget, you know, stuff that you learned a couple of weeks or months ago radically, catastrophically. But the current AI systems all do what's called catastrophic forgetting.

**Tim Scarfe** 01:25:45

Yep. When we do solve that problem, so we create a gentle continually learning AI, do you think the dynamics of the system would be dramatically different in terms of capability?

**Jeff Clune** 01:25:57

Sure. Yeah. I mean, imagine if right now, like, collectively, the frontier models are having, what, hundreds of millions of conversations as you and I talk simultaneously on earth. If a system could learn from all of those conversations, how how smart would it get? How fast?

**Tim Scarfe** 01:26:19

Could you argue in some galaxy brain sense that, you know, we we form a memetic superintelligence and we are you know, I was kind of advocating we could have personalized intermediate small models that do transductive active fine tuning and feedback up to the main model, you know. In a way, that's what we do. We we we are the extended minds of the LLMs already. So shouldn't we expect to see more

## Continuous Learning Challenges

**Tim Scarfe** 01:26:40

dynamism even though there's this delayed ingest of data back into the language model?

**Jeff Clune** 01:26:46

I missed the question. Can you say

**Tim Scarfe** 01:26:47

it again? Well, in in in a sense, we already are, like, and and we we we form an intelligent system. And we are the agents that control the language models. We we generate data that's fed into language models. We do continual learning and and so on. So it's almost like, shouldn't we already expect there to be a dramatic capability increase in the system just with us as users of language models?

**Jeff Clune** 01:27:10

Well, yes. I mean, humans are getting are becoming more productive by using language models. And because of that, we are doing more, producing novel code, helpful code, novel solutions to math problems, novel discoveries in biology, science, machine learning. And so the data, which is the result of that process, is being dumped out of the Internet, which provides fodder to stop and train the new model. But don't you just feel the lack of continuity here? It we're not con-

tinuously learning from the interactions with humans. We're not continuously learning from new stuff that's being posted online. It's all this kind of like very artificial, you know, take some data, scrape it, filter it down to what we consider high quality, put it in a training set. Now we will train. Now training is done. Now we will deploy. Maybe in deployment, we will get new interactions. We'll get some ratings from users, thumbs up, thumbs down. We will store that in a dataset in 3 months when we have the right GPUs with these hyper parameters. We'll try a couple different runs to produce a better thing, and then we will deploy that thing. There's just nothing of, like, nothing like the experience of you in your life constantly learning, getting better every day from every interaction in a way that doesn't throw out a lot of the stuff you've already learned, but builds on that knowledge and some never ending beautiful process of discovery and skill improvement. Basically, that's the dream and we're not there. We're so far from it. I do wanna go back to 1 other thing you mentioned earlier, if you don't mind, which is in the paper on Omni, you mentioned that I had this okay. Actually, going all the way back to Poet. I had this 1 of this this plot, which is 1 of my favorite plots of my career. And the reason why I liked that plot to kind of paint a picture for your listeners is that there was this phylogenetic tree of environments that have been generated so far, and I had been seeking my entire academic career to produce something that looked like a phylogenetic tree from nature. So what does the tree of life look like? It had inventions, like eat like the bacteria, the fungi, the animals. And then within animals, it said animals are a good idea and fungi are a good idea. I'm gonna continuously innovate and to pick up stepping stones from each of those and and and build on top of them. And then with animals, get like, you know, the big cats, and then you get like rodents. And then there's endless innovation over in rodents, but also endless innovation in in in building on the ideas of great cats. So you get more and better cats, more and better robots, more and better mushrooms. All of this is happening. And so if you look at it, it's a tree, and the tree has deep branches. Like, the fungi branches last for a over a long time. Historically, in a life and evolutionary computation and open endedness, every time you ran anything, the diversity would just collapse, you know, and you would not get simultaneous probing of radically different kind of types of thing or types of idea. So in poet, it was the first time you saw a phylogenetic tree that had deep branches. This is not exactly what happened, but metaphorically, because in this case, poet was inventing the environments themselves, but it's like it invented water worlds, and and then over here, it invented forest canopy worlds, and over here, it had like desert worlds, and it would keep pushing on better desert worlds, and water worlds, and tree canopy worlds forever, and just got better. The tree was just growing and kind of filling up, And that's why I loved it and thought it was so beautiful. And you mentioned Omni. 1 thing that has me super excited about Omni is it's now doing that, but in a search space where anything is possible. It's a Darwin complete search space. So as you

mentioned, it starts out by saying, oh, here's an obstacle course. Here are these platforms. First, the platforms are static and next to each other. Then the little platforms are going up and down, and the agent has to jump from 1 to 1. Then it kinda separates the platforms and has them going up and down. So it's riffing on the idea of of obstacle courses of platforms. Yep. Somewhere along the line though, because it has a sense of interesting, it's like, okay, what's another interesting thing to do would be to be kicking a ball around. So first, it's just kick a ball, then it's kick a ball through goalposts, then it's kick a ball through moving goalposts, then it's kick a ball off of a wall through a goal. Then kick a ball off a wall through a moving goal. Then kick a ball off a moving wall through a moving goal. So you can see it's riffing on the idea of more and more complicated ball kicking domains. At some point, it's like, that's kind of boring. I'll create a little conveyor belt system where you like put packages on a series of conveyor belts, and you have to know ahead of time by putting it here, it's gonna end up getting pushed this way, then this way, this way, it's gonna end up over the blue area versus the green area. It's learning to control that system. It starts to invent more complicated buildings, like search a 2 room building, a 4 room building, a 4 story building. And then in the end, it starts to say, I'm gonna create, like, a construction scene you have to clean up. And the final 1 that was my favorite, a cluttered restaurant where the robot has to, like, clear the dishes off all of these tables. And all of that showed up in 1 run. The more that we run it, we get all of that stuff. And if we ran it more, we would just get more. And so I really don't know what happens if we run that thing for 1000000000 years, you know. The agent would just level up in skills, and like, what are the environments 1000000000 years out in this algorithm? And so you could draw a phylogenetic tree of that algorithm, and it's just the fact that it's doing sustained innovation in so many different simultaneous directions, and sometimes putting this together, it starts to have the feel of why I got into open endedness. It is the feel of the magic of Darwinian evolution and the feel of the magic of human culture. Right? Human scientific communities continuously invent new problems and new techniques that unlocks new problems, and we push simultaneously in chemistry and biology, in CRISPR stuff, in vaccine stuff, over in material science, forever and ever expanding outwards. And now our algorithms are starting to do that. The AI scientist does that. ADAS does that. OmniEpic is doing that. We have[22][23]

[22] OMNI-EPIC — paper
[23] OMNI-EPIC: Open-ended Environments from Foundation Models — paper

## OMNI-EPIC: LM-Driven Environment Generation ●

**Jeff Clune** 01:33:20

**a new paper. We're doing it in self play where we have different strategies for each side of, like, a competitive coevolutionary arms race. And even a new work that we haven't yet talked about, but I'd like to mention real fast, is we're using the AI's ideas now for AI safety. So take a new language model off the shelf. You might, before you deploy it, wanna know, does it have any surprising failure modes, And does it have some surprising new capabilities that you might be scared of or worried about that you wanna nerf or eliminate or not disallow or block train out of the system or just be aware of because it's cool and you wanna share that, hey. There's a new thing we can do. So we launch an open ended system, same recipe, where the system says, hey, can you do this? Can you do this? Can you do this? And given the stuff we've already discovered about what it can do and can't do, that is now a little library of stuff we know. And And then we ask, what's the next interesting thing to try? And the next interesting thing to try? And the next and we grow this library of discoveries of things that are surprising ways that the model fails and succeeds, and that is all automated with the principles of open ended open endedness. And so now a new frontier model could come to us and say, hey. We wanna automatically red and green team our system ahead of time, basically paint us the picture of what we don't yet know. In some sense, we're capturing the magic of human scientists in the crowd. Like, you put a new model on the line, what happens in the first 3 weeks? Everybody's saying, hey. It can't do 3 r's in strawberry. It can do this crazy new coding challenge. It can do this new reasoning challenge. And everybody is looking at what people have found so far, saying, they can't do 3 hours in strawberries. Can I do this? Can I do that? Try it. If you find that it's interesting, add it to tweet about it. Right? That's adding it to the set of stepping stones. So basically, automated Twitter. Twitter is self discovery of foundation models, capabilities, and safety problems, and we've done it in this new algorithm that Song Liu pioneered called automatic capability discovery or ACD. But the point isn't any 1 of these projects. The point is how this same basic recipe, which dates all the way back to open endedness, and Ken Stanley and Joel Lehman and the work that we've been doing for almost a decade now or more, almost everywhere we apply this set of principles, good stuff starts to happen. And so that's what makes me so confident that these directions are exciting and profitable.**

**Tim Scarfe** 01:35:36

Yeah. It's it's so incredible. So there are so many things out there in the world like, you know, culture, language, and so on. We've built an artificial system which represents many of those features. I just wanna ask you a question though. So, you know, Omni Epic is is fascinating. You you were saying that you can trace this phylogeny and you get these points of divergence in in the phylogeny. Right? So we can do topological analysis. And I just wondered, if you had access to a hypercomputer that could, you know, like do an infinite amount of computation. And presumably you could run experiments, many, many experiments of what would happen if we just created all of these phylogenies and we did topological analysis to look at the the the branching structure and so on. What do you think you would see?

**Jeff Clune** 01:36:19

It's a great question. And this is actually, I'm not sure if you're aware, but this is 1 of the most fundamental questions in evolutionary biology. So it was posed by Stephen Jay Gould.

**Tim Scarfe** 01:36:28

Yeah.

**Jeff Clune** 01:36:29

And it's called it basically goes under the metaphor, if you rewound the tape of life Yeah. And replayed it, what would happen? Yeah. And so there's a debate. Would you get humans again? And would humans be very similar? Would they have descended from ape like ancestors? Or would the dolphins have become intelligent? Or would nothing have been become intelligent? How convergent, how rep repeatable is evolutionary biology? Obviously, we don't get to live in the other universe, so we don't know. But some people have been trying to answer this question. And so there's been a huge arc of work out of the lab that I grew up in actually. Rich Linsky and his lab. He Rich Linsky is effectively the father of experimental evolution. About 30 years ago, he started an experiment where he created, I took 1 cell of E. Coli or maybe 12 cells and he's simultaneously evolving them. They started out very similar in 12 separate lineages of petri dishes or flasks actually, and they just grow forever according to evolution. And the question 1 of the many questions was, would the 12 lines go in the same would they end up in the same place or they radically go in different directions? And the short answer is it's complicated. Some things tend to kind of happen seem to happen kind of across all of them. Some things are wildly

**idiosyncratic to individual ones. So in 1 of the the 12 lines, basically, it invented a new species. So 1 of the species definitions for E. Coli is that in certain conditions, I think they're anaerobic, but I'm gonna get some of the details of this story wrong. In certain conditions, it does not consume this material called citrate. And it just so happened that Rich, when he made the original experimental design, because e coli could not consume citrate, it's like a common thickening agent. And so it's just sitting there in the medium in the flask, like in the in the water, you know, in the air effectively. The air for this thing, the liquid. And 11 of the 12 lines to this date is unless unless I'm out of unless I have my story is a little out of date by a few weeks. 11 of the 12 do not consume citrate. But 1 of the 12 lines violated the species definition of E. Coli and learned how to consume this meal, this resource that have been sitting there all along. Okay? So it is arguably some new species. There's debate over this. But what's really cool vis a vis this question is not just the fact that it happened in 1 and not the others, but after they it invented this amazing new capability to adjust to this resource, which is really difficult, it effectively arguably makes it a new species, they wanted to know how repeatable is that? Not across the 12 lines, but in the history of that particular line. So it turns out because they freeze the samples of the the species every so often, they have the entire fossil record of that 1 particular lineage that ended up having that amazing speciation event. And so there's a friend of mine, Zach, who has gone back and studied this question, and he basically you can go back and take some fossils, some fossil samples from the the the lineage, like a couple days before, and then grow a bunch of them. How many of them invent the citrate thing? Go back a little bit farther like a couple weeks and make many copies of that. What fraction of those things end up in the same place? And you can go further and further, you know, months back, years back, 5 years back, and say, like, was there 1 magical thing right before it happened? Some lucky lightning strike that enabled this to happen, or did some amount of mutations set the stage and make it pretty inevitable from that point forward, and how far back is that stage setting? And they've done a lot of papers that are in the top journals of the world, and basically the answer is it's complicated. There is some stage setting, and so the farther back you go, the less likely you are to get citrate, but you definitely get citrate repeatedly if you're like, don't go back too far, and it's some complicated function. So long story short, evolution has both chance driving where it ends up and some themes that are consistent. And so in our own algorithm, we could study that, and that's 1 of the beauties and the original reasons I started working on this stuff is to to address questions like that. By repeatedly doing runs of open endedness, you could study how contingent is evolution or whatever algorithm you're using, and that is beautiful. And I my guess is you've got a little bit of both. Sometimes you get random magic that that causes something that is almost impossible to repeat, but there's some**

**themes, like things get more intelligent, they get more complicated. You can also do the same thing not with the process. How how consistent is the process that led to like, the phylogeny that led to something really cool, but the final artifact. So imagine if we ran open endedness in 1000 different runs, we got 800 of them produced general intelligence. But now you could take that final intelligence and look across it. How different is it? You know? Yeah. What's the math look like over here versus over here? What does the music look like over here? Do they all have music? Do they all have humor? Do they have a notion of love? Do they have a sense of curiosity? What is the difference in their sense of exploration?**

**Tim Scarfe** 01:41:34

Yeah. It's fascinating to talking about that the space of alien intelligences. But yeah. So so themes could be complexity and intelligence. Motifs could be, you know, morphology. There's a there's a great thing, isn't there, called morphologically convergent evolution and carcinization where, you know, crab like forms appear in different parts of the of the phylogeny. And and, you know, I kind of wonder I suppose whether it could be described through emergence, through, you know, interaction dynamics of of sort of basis phylogeny functions or whether some kind of physical determinism or something like that. But I I did want to ask you another question about the the omni epic thing, which is that 1 interesting thing about evolution is isolation. Right? So, you know, we we have all of these pockets of evolution that that just kind of happen in situ and they don't share information with other parts of the phylogeny. And and that's very, very important for niche construction and so on. In something like Omni Epic, this is an operating parameter of the algorithm, which is the the extent of information sharing between the environments and the agents and so on. Would that be something that that you'd be interested in studying?

**Jeff Clune** 01:42:38

It's a lovely point. Before I answer that question, I just wanna make 1 comment on the previous 1. There's this great quote which says, if religion did not exist, it would be necessary to invent it. Yes. Right? So you could ask that question, like, many of these 800 intelligent, you know, races that come about invent religion, for example? Does AI invent its own religion? And and any property you're interested in, could go see, to not just did they invent it? How does it look? Is it different? But how consistent is it? All of this is such beautiful science you could do if we ever cracked open endedness. Yep. It's 1 of the many reasons I find it so fascinating. Okay. To the Omni thing, Omni is really interesting because you get to make some choices as you mentioned. Okay. So imagine you have

## Evolutionary Biology Insights

**Jeff Clune** 01:43:20

a library of so so Omni right now is trying to generate just for your listener. It's trying to automatically generate an endless stream of interesting environments that an agent, if it trains in, will learn something new. Okay? So you end up, if it's doing its job and and it does work, with this massive set of environments that are all interestingly different. And at the time they were invented, the agent was getting some benefit from learning on them. Okay. Now when you go pick an environment, like, click the the cluttered restaurant scene, for example, you from an algorithmic design, when we made Omni, we had some choices. 1 thing you could do is you could say, hey, take the restaurant scene environment and oh, by the way, we didn't I didn't say this yet. But when we generate the new environment, we're asking a language model to look at the current code of the current restaurant environment and make the next interesting environment that would have high learning progress. It's interestingly different. It would have high learning progress. And then a language model spits out the code, and then we check if it's interesting. K? So when we go to ask, we could I call that an intelligent mutation operator. It's kind of like evolution. You take a thing, you change it a little bit, and you see if it's good. But you don't change it randomly now, because if you change code randomly, it doesn't compile. But instead, we take the link the the code that specifies the rest of the cluttered restaurant, give it to a language model, we say, hey. This is this is an environment the agent can currently do well. What's next? And it might say, oh, let's do it with more tables or more silverware, or let's add other agents that are actively, like, continuously coming into the restaurant, or let's make the tables kind of move up and down in the float or, you know, whatever it wants to do. It's gonna make some interesting mutation to the environment. But when you ask this system, can you make the next interesting thing? How many things do you give it? You could give it just the restaurant scene, or you could give it the restaurant cluttered restaurant scene and the kick a ball through the moving goalposts. And then maybe suddenly it's like, ah, I'm gonna combine these ideas and have a restaurant scene, but the waiter doesn't get to pick up all the dishes. It has to throw the dishes across the room into a moving tub or something like that. Right? And so you get to this and and and you don't have to restrict yourself to 2 parents. You could give it the entire archive and say, at all of the environment so far in context as an in context learning, and now what's next? Okay. So let's assume that you do the latter. You give it the whole context and it produces a new thing. Well, how do you draw a phylogenetic tree of that? In some sense, everything the parent of every new thing is everything that came before. You can't draw the tree. But

since and and and we tried versions of that, and that is it does still push in the direction of moving goalposts, you know, like football worlds and obstacle course worlds and cluttered restaurant scene worlds, because it knows still the human notion of interestingness, but there's no longer this direct obvious way to draw a phylogenetic tree. In the current system, what we do, in case you're curious, is we what we found worked really well is you go and you get 1 type of you sample 1 environment randomly, like the clutter restaurant scene, and then we we don't just wanna give that only to the system and have it generate the next thing because it might have already seen the clutter restaurant scene before and already generate the next thing. So what we do is we get the clutter restaurant scene, then we use embeddings to say, now give me the nearest 5, for example, things to the restaurant scene. Give those 5 things to the LLM and say, here's a little cluster of environments you've made so far, generate the new thing. And it does, but then that complicates the drawing of the phylagena tree because now there's arguably 5 parents to this 1 environment. So in our actual graph, we take I believe when we draw the lines, we take of the 5, the 1 that ended up being closest to it, that's and the 1 we draw to make a pretty picture. But really, the picture is some complicated web of interconnectedness that you don't see. Also, 1 other fun fact. After it generates a new environment, like we give it the 5, we say what's next. It generates a new thing it thinks is interestingly new. We don't necessarily wanna trust that that thing is new because it may be a generated a thing we already have. So once we have that new thing, we go and we we with embeddings, we pull out the 5 nearest things to that thing, and we say, hey. You just generated this. Here are the 5 closest things that exist in the archive already. Do you still think this counts as interestingly new? And if it says yes, then it's in. And what I love about this is it basically is just like a graduate student or a professor. You read some papers. You're inspired. Hey. That actually makes me think this would be a good idea. But hopefully, if you're doing your job right, instead of going and spending a year on that new idea, you first go do a literature check now that you have the idea in hand to say, and is this idea actually has it been done? And if it hasn't been done, let's go.

**Tim Scarfe** 01:48:07

Beautiful. Even things like multiple inheritance, that there are there are so many topologies in the artificial world that could actually allow us to extend beyond the natural world in in quite clever ways. Jeff, I I wanted to talk a little bit about your background in in philosophy. So you majored in philosophy. Yep. Right? Which and of course, you're you're a computer science professor now. How has that influenced your trajectory?

**Jeff Clune** 01:48:31

Well, I mean, from a really practical perspective, I think that 1 of the skills you don't spend enough time on if you get a technical degree is communication. And so just being a humanities major, maybe you've spent a lot of time on clear writing, clear thinking, clear communication. And so I think it's been helpful in my career to prioritize. Like, if you do something really great, but you can't communicate to somebody else why it's interesting and what you did, then, you know, you have a lot less impact. And so I think that's helped practically. I also think philosophy just really trains really great clarity of thought. Like, it really forces you just to question your assumptions. And so 1 of the things that I think is a gift when I do science is an epistemic skepticism, which is philosophy talk for, like, not trusting anything, for being skeptical, as you know. And so there's been so many times in my career where I'm working with somebody, and they say, alright. We did this. We got these results, and therefore we can conclude this. And I say, slow down. Are we sure that we know that? How do we know that? What would that assumption what predictions does that assumption make, and can we test those predictions? And often, people will stop and say, actually, we don't actually know that. And this is the right experiment to try to figure that out. And sometimes when we do that experiment, it confirms what we thought, and often it falsifies what we thought. So I think being very, very skeptical is a great trait of philosophers and scientists.

**Tim Scarfe** 01:50:07

So how does a philosopher end up as a professor of computer science? It's a great question.

**Jeff Clune** 01:50:12

And it's actually kind of ironically a long convoluted circuitous path that led me here. It definitely wasn't a straight line. So I'm kind of, you know, art follows life or life follows art. And so the story is that I was working in dot the .com boom in Silicon Valley and I was actually in a marketing department and I didn't really, you know, love what I was doing. So at my desk, I'm reading articles about things that I find interesting and I came across this article from the New York Times and it just set my mind on fire. And then what happened in this article is people had created a virtual world where they could evolve virtual creatures using an evolutionary algorithm, real mutation, real survival of the fittest. And when the robot morphologies, which were changing over time, got good enough, fit enough, it would automatically send the design over to a 3 d printer. They would 3 d print the robots, they would crawl out into the real world. And I remember like a thunderclap in my mind reading this article. I was like, everything I read was so amazing. The fact that you could instantiate evolution inside of the computer, that means you can study it. That means you could ask questions like how repeatable is it? Why did sexual reproduction show up instead of asexual reproduction? What happens if you have 3 parents instead of 2? You know? How do you get agents that cooperate with each other that invent language? Everything you wanted to know about evolution was now testable. It was like amazing. And you can harness it to make really cool

## From Philosophy to AI Research

**Jeff Clune** 01:51:40

engineering designs and maybe even intelligence itself. So that happens. And at the time, I didn't know really, you know, I didn't know how to find out more. I will say as a philosopher, I didn't train me to these things called papers and you can keep reading. So I couldn't like learn that much more, but it just stuck with me. So I was not satisfied with my my.com job, quit my job, wouldn't travel the world for 15 months living out of a backpack, surfing, traveling, loving life. And at the end of it, said, well, I'm I'm coming back from my trip, like, do I wanna do with the rest of my life? And the answer is I said, want to do that thing that that guy was doing. So I went back, I found the article which had not left my mind. And I said, okay. It turns out it's a person named Hod Lipson. He's a professor at Cornell University. And I said, let's let's contact him. So I sent him an email, which I'd love to go find and dig up. And was like, hey, I'm a philosophy major. I've just traveled the world. I really love that that that work that you did that was on the cover of the New York Times. How do I I wanna join your lab, you know, can you

let me in? And he surprisingly wrote back and he said, well, you can't actually get up, you know, get into the PhD in computer science program at Cornell. It's 1 of the best universities in world with like a philosophy major undergrad. And he's like, so that won't work, but, you know, you seem really passionate and I'd love to try to work together somehow. So, you know, we were trying to make it work. We couldn't really find a great way to make it work. And so I said, alright, I'll apply to other schools that aren't, you know, Ivy League. So I ended up email contacting and emailing 85 different universities. Everything that Google at the time Google didn't even really work very well, but everything that Google returned for genetic algorithms, every professor I I emailed. And every 1 of them was like, no. No. No. No. You can't get a PhD in computer science with an undergrad in in philosophy. But 2 responded, and 1 of them said, hey. You can't get a p can't can't just go into the PhD program, but there's a philosopher here. His name is Rob Penuck, and he works with people who do evolutionary algorithms and a life. And I said I I looked up who these people were and I said, okay, I'm in. And so I had a secret plan. So I'm gonna go to Michigan State. I'm gonna get a master's in philosophy. When you're in grad school, they don't restrict what department you take lots of your courses in. Right? You get your core, but you can take other classes. So I said, I'm gonna take all the AI classes, all the machine learning classes, I'm gonna learn the math, I'm gonna learn to code and show them that I can do this. And so I did, I worked my tail off, I learned a program, I learned the math, I took a machine learning class, I had no idea at the beginning what I was doing, but I ended up acing all the courses and publishing papers with the artificial life lab there, meeting Charles Aphria who became my PhD advisor and Rich Linsky and Rob Penack and all these great people working on E. Coli and artificial life. And after my master's degree, I had published with them and gotten good grades and I said now can I get a PhD in computer science? And because they knew me, they made an exception and they let me in. So I'm in the PhD program, I work really hard, I publish more papers, I meet Ken Stanley, we start working together on open ended and neuroevolution and CPPNs and all these great things. Then I call up Hod Lipson after I said, okay, I now have a PhD, it's been 8 years since I first contacted you, now can I join your lab? And Ahad said, we'd love to have you be you know, I don't have funding. So I went and I found funding, secured that funding. I called them back up and I said, now I have a grant from the NSF that says I can do a postdoc wherever I want and I wanna do it with you. And he said, I'd love to have you come join the lab. So 8 years after I read that article in Silicon Valley having no idea what I was getting myself into, I was in the lab at Cornell and it was like Willy Wonka's robot factory. Just 3 d printed parts on the wall were evolving crazy soft robots every day. It was a wild and crazy adventure. And 2 years later, I became a professor at Wyoming. And you know what? I started getting emails from people being like, I like what you do. How do I get into your lab? So I like to tell this story to people even though it's a little long because I don't

think you should take no from an answer from the world, you know? Pick some, you know, some crazy thing you wanna do. Follow your notes for what seems interesting and there might it might be a long circuitous path, but eventually if you just keep going, I think you can get yourself to really interesting places and it's very fulfilling.

**Tim Scarfe** 01:55:47

Yeah. And from a serendipity point of view, having a multidisciplinary research is is really important because, know, otherwise you wouldn't have come in with all of these completely different ideas. Just quickly, as you mentioned, you know, compositional pattern producing networks and NEET and Kenneth Stanley and all this stuff's amazing stuff. When did you first meet Kenneth in person?

**Jeff Clune** 01:56:05

I believe it was online. So I was a person I was in an aLife lab, and a lot of people doing experimental evolution and artificial life, and I like that stuff. I did some of my early work on, like, can evolution maintain like optimize its own mutation rates, and why does altruism, which is like being nice to each other, why does that evolve? Like hardcore, open questions in evolutionary biology that we were testing in simulation. But I always had an eye towards intelligence itself. Okay. What is happening inside of my brain? How can you make an intelligent thing? And that led me to neural networks. And then I said, okay, I know evolutionary algorithms. I wanna know how you could evolve something that will kind of become more and more complex and eventually could produce the human brain. And so I ran across the neat algorithm, which is a really early algorithm, which is not from something that is now called neural architecture search. Starts out with a simple brain adds more neurons over time and it was beautiful work. And so I joined the neat users group and I started talking online, started reading about these people's papers and interacting with this guy named Ken Stanley and I keep basically posting really long sets of questions and all of the work coming out of his lab. I started reading everything that their lab does, and I'd love to know what he thought of this this upstart young PhD student early on that was asking him all of these questions. But eventually we had talked enough that I said when CPPNs came out, I was so blown away by how beautiful the work was. I said, I wanna work on this. And I started I basically went and did a little research project. It had nothing to do with my normal lab work. I said, I just have to study this new CPPN thing. It's so cool. And I started asking questions that I thought were interesting, and I put together this little lab report. And I basically emailed Ken, and I said, hey. I've done some work on that builds on your new work. You know, I'm hoping you'll take a look and just give me some feedback. And, of course, if you want, then I'd love to work together. And he wrote me back

an email that was like, I initially looked at this and I didn't think I wanted to look at it. I wasn't planning on looking at it, but I thought I'd be nice to just give like a quick read to the top. And he's like, and I couldn't put it down. I just kept reading it, kept reading it, kept reading it. And by the end of it, not only did I read it, but I wrote this long complicated response to you. So I'm sorry it's so long, but it's beautiful work. It raises all these interesting questions. And, you know, basically that was the first paper we wrote together. So that, you know, I started answering his questions and back and forth and doing experiments he suggested and I ended up winning a best paper award, if my memory is correct, at GECO, which is the conference I was going to at the time. And then we met in person, and we're off to the races.

**Tim Scarfe** 01:58:28

What was the name of the paper?

**Jeff Clune** 01:58:29

I believe the paper the the theme of the paper is effectively how do CPPNs encode regularity Yeah. In and, you know, how do they respond to how do they produce regularity in neural nets that control robots? And how do they deal with changing that regular like, do they deal with the fact that sometimes you need to have things that are irregular and sometimes you think that you need to have things that are more regular? I wish I could actually quote the paper title on that, but what we can I could look it up? We could put it in the show notes.

**Tim Scarfe** 01:59:01

Yeah. Amazing. Jeff, this has been so brilliant. Thank you so much for coming on. I I really appreciate it.

**Jeff Clune** 01:59:07

My pleasure. It's so fun to talk about all of these things. I do wanna give credit, if you don't mind, to all of the tremendous people I've worked with. So I've mentioned a lot of the people that have been there through most of the journey, like Joel and Ken. But there are so many great people that I've worked with along the way. All of the PhD students in my lab, all the people I've collaborated with, Jean Baptiste Moray in particular, very influential on me. Tim Rockechel, you know. And there's just too many to name, so I just wanna just kinda have a blanket statement that if you are interested in any of this stuff, please look at the full author list of all the papers because it is not 1 person. It is a collective community that's been building together all this work.

**Tim Scarfe** 01:59:42

I'd love to get John Baptiste on it. You you just reminded me actually, Kenneth Ferm suggested we get him on.

**Jeff Clune** 01:59:46

He's fantastic. He's 1 of the smartest and most brilliant scientists that I've ever worked with really, truly. He does fantastic stuff and I've never had a boring conversation with him. You know, I I really every time I talk with him, I learn something new and I listen as hard as I can.

**Tim Scarfe** 02:00:00

Amazing. Thank you so much, Jeff.

**Jeff Clune** 02:00:02

My pleasure. Thank you for having me.

## References


[1] **Jeff Clune**
https://x.com/jeffclune
*person*

[2] **POET: Generating/solving complex challenges**
https://arxiv.org/abs/1901.01753
*paper*

[3] **POET: Generating/Solving Complex Challenges via Paired Open-Ended Trailblazer**
https://arxiv.org/abs/1901.01753
*paper*

[4] **Why Greatness Cannot Be Planned**
https://www.amazon.com/Why-Greatness-Cannot-Planned-Objective/dp/3319155237
*book*

[5] **Why Greatness Cannot Be Planned: The Myth of the Objective**
https://www.amazon.com/Why-Greatness-Cannot-Planned-Objective/dp/3319155237
*book*

[6] **Kenneth Stanley**
https://www.cs.ucf.edu/~kstanley/
*person*

[7] **Automated capability discovery in foundation models**
https://openreview.net/forum?id=nhgbvyrvTP
*paper*

[8] **Automated Capability Discovery in Foundation Models**
https://openreview.net/forum?id=nhgbvyrvTP
*paper*

[9] **NEAT: NeuroEvolution of Augmenting Topologies**
https://nn.cs.utexas.edu/downloads/papers/stanley.ec02.pdf
*paper*

[10] **NEAT: NeuroEvolution of Augmenting Topologies**
https://nn.cs.utexas.edu/downloads/papers/stanley.ec02.pdf
*paper*

[11] **MAP-Elites: Illuminating Search Spaces**
https://arxiv.org/abs/1504.04909
*paper*

[12] **Novelty search vs objective-based optimization**

https://www.cs.swarthmore.edu/~meeden/DevelopmentalRobotics/lehman_ecj11.pdf
*paper*

[13] **Novelty Search vs Objective-Based Optimization**
https://www.cs.swarthmore.edu/~meeden/DevelopmentalRobotics/lehman_ecj11.pdf
*paper*

[14] **AI-generating algorithms approach to AGI**
https://arxiv.org/abs/1905.10985
*paper*

[15] **AI-Generating Algorithms: An Alternate Approach to AGI**
https://arxiv.org/abs/1905.10985
*paper*

[16] **Open-Ended Learning Team at DeepMind**
https://deepmind.google/discover/blog/open-ended-learning-leads-to-generally-capa
*reference*

[17] **Video PreTraining (VPT)**
https://cdn.openai.com/vpt/Paper.pdf
*paper*

[18] **Video PreTraining (VPT)**
https://cdn.openai.com/vpt/Paper.pdf
*paper*

[19] **Thought Cloning: Imitating human thinking**
https://arxiv.org/pdf/2306.00323
*paper*

[20] **Thought Cloning: Imitating Human Thinking**
https://arxiv.org/pdf/2306.00323
*paper*

[21] **Go-Explore: A New Approach for Hard-Exploration Problems**
https://arxiv.org/abs/2004.12919
*paper*

[22] **Automated Design of Agentic Systems (ADAS)**
https://arxiv.org/abs/2408.08435
*paper*

[23] **Automated Design of Agentic Systems (ADAS)**
https://arxiv.org/abs/2408.08435
*paper*

[24] **OMNI-EPIC**
https://arxiv.org/abs/2405.15568
*paper*

[25] **OMNI-EPIC: Open-ended Environments from Foundation Models**
https://arxiv.org/abs/2405.15568
*paper*