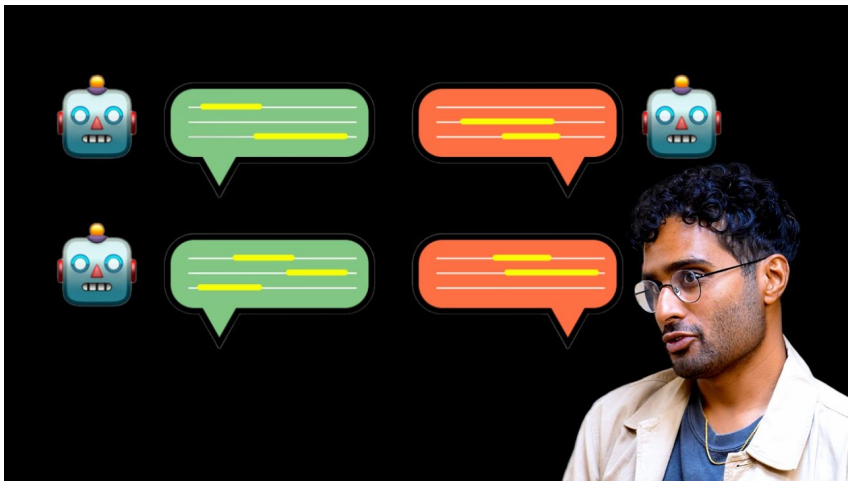
/mlst

This is what happens when you let AIs debate

March 24, 2026



Machine Learning Street Talk

Podcast Transcript

Revision: c9c3dca9

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

AI Alignment and Debate Techniques 3

Scalable Oversight and Hidden Information 5

AI Agency, Intelligence, and Progress 8

Base Models, RL Training, and Instrumental Goals 10

Deceptive Alignment and RL Challenges 15

Open-Ended AI Systems and Future Directions 18

Deception, Superintelligence, and Possible Minds 21

Cultural Evolution, Memetics, and Intelligence 23

References 28

Overview

Akbir Khan, AI researcher and ICML 2024 Best Paper winner, joins Tim Scarfe to discuss his groundbreaking work on using debate between language models to improve AI truthfulness and oversight. Khan explains how pitting two LLMs against each other in structured arguments helps non-expert judges arrive at more accurate answers than simply querying a single model — a result with profound implications for supervising AI systems that may eventually surpass human capabilities.

The conversation moves through the mechanics of scalable oversight and the "sandwiching protocol" that operationalises the problem of checking entities smarter than their supervisors. Khan describes how debate naturally surfaces the cruxes of disagreements, making complex expert judgments more accessible to laypeople. The discussion broadens into the relationship between intelligence and agency, the risks of deceptive alignment and reward tampering, and whether the current trajectory of AI development constitutes the early stages of a Cambrian explosion in artificial minds.

Khan and Scarfe also explore open-ended AI systems, Kenneth Stanley's arguments against objective-driven optimization, and the philosophical terrain mapped by thinkers like Aaron Sloman and Francois Chollet on the space of possible minds and measuring intelligence. The episode closes with a frank exchange on cultural evolution, memetics, and whether the intelligence that matters most for AI safety is the kind we can measure.

AI Alignment and Debate Techniques

Akbir Khan 00:00:00

You're gonna have, like, models that are smarter than you in the future, and you wanna check their answers. You can get 2 experts to have an argument, and you can have a non expert check that argument. That's the premise of debate. There's lots of people who work on AI safety because they're very worried about the doom. I work on AI safety because I think it's the most interesting intellectual problem of our generation. Getting things smarter than you to do what you want is like a problem as old as parenting. Having been doing ML since like 2015, it's hard to imagine a world where the the progress doesn't keep happening. There's lots of incentives for us to try to keep making these models smarter, and I don't think we've hit any natural limit in that.[1][2]

Tim Scarfe 00:00:35

So what are you doing this week and

Akbir Khan 00:00:38

finish telling us about the paper? Awesome. So this week I'm at ICML Vienna to present has now received the best paper award, which is debating with more persuasive LLMs leads to more truthful answers. It's hopefully the start of a really good set of scalable oversight work in the alignment track. And I think this is the first best paper award that's gone to, like, an alignment paper. But it's like hard to verify this. So I've asked a couple of people who've been in alignment much longer than me. This seems to be the case. In particular I think the introduction is a great case for why you should worry about systems being smarter than us and not being able to supervise them. And in particular it provides an empirical operationalisation of like things are smarter than you how are we going to check them based on some work by Ajayi Kotra and Sam Bowman. So I think in particular scientists or like researchers will really like this because it's like how do you actually do we operationalise this problem of a thing being smarter than you?[3]

Tim Scarfe 00:01:34

We also show that this works well with humans as well. I know that you've got a personal interest in alignment and maybe we'll get to that afterwards. But I did want to just use a bit of time on your paper from This is ICML, right? So

[1] Akbir Khan — person
[2] MLST Episode Shownotes PDF — other
[3] Akbir Khan et al. - Debating with More Persuasive LLMs — paper

Yeah. Debating with more persuasive LLMs leads to more truthful answers. Yep. What's the elevator pitch? So the elevator pitch is you're gonna have like

Akbir Khan 00:01:56

models that are smarter than you in the future and you wanna check their answers. You can think of this equivalent of something that stresses me out a lot is if I go to see a dentist and they tell me I need 2 fillings for cavities, I just have to take them at face value that they know what they're doing because I have someone here who's a domain expert and it's very hard for me to check their answer. So what do I do if I don't trust my suss dentist? I instead go and get a second opinion from a different dentist. I take whatever the x rays are of my teeth and I take it to them and I ask them for their opinion. They will say something else. Then normally what happens is I take their feedback and bring them back to dentist number 1. There's this constant flow of arguments between them and then I make an informed decision based on it. That's the premise of debate, which is you can get 2 experts to have an argument and then can have a non expert check that argument. It's why we use debate in the public spaces that we do. The reason that debate doesn't work in public spaces, which I think is, like, obvious, is that, like, most of the time, like, the 2 debaters aren't equally matched. And, like, the fun thing about doing alignment work is that, like, I can literally take a copy of a model so I can, like, make sure the capabilities are adjusted for. So there's kinda 2 things happening here. 1 is, like, this adversarial expert versus expert thing Mhmm. Which is, like, a fair game. And then the second thing is that it's easier to check answers to a debate than it is to make arguments in a debate. Few things on that then. So yeah, there's this term, the wisdom of crowds, you know, which is that if you're doing epistemic foraging, diversity is good. But maybe not always, because you might have lots of people in the crowd who don't know what they're talking about. Or you might have people who do know what they're talking about, like you know, Yan Lakun and Hinton, and they vehemently disagree and they don't really converge. So sketch it out for me. So first of all, taking multiple votes in LLM debates does help. Maybe it's like an empirical piece of evidence Second, for the maybe meta point on the whole popularism or, you know, just like, wisdom of the crowds is normally wrong thing is, like, I I think, like, the failing of the wisdom of the crowd thing is something, it's very hard to keep a crowd fully informed on a thing. It was, like, hard to make them attend to a debate as fully as you'd like them to. So I'm like worried a bit less about that in this setting or like the wisdom of the crowd kind of like worry is maybe more of like, you have like lots of judges and very few debaters.[4]

[4] Yann LeCun on Machine Learning Debates — video

Tim Scarfe 00:04:13

And the setup that we're envisioning using this stuff is not really that. Yeah. So in this setup, and this is really interesting because people talk about language models hallucinating. Yeah. And I guess what's happening is there's this kind of traversal space. Yep. So you can you can traverse many different directions and you get slightly different results. And what you've come upon is there is a non interactive kind of adversarial way of probing that space which closes the gap getting closer to epistemic truth, whatever that means.

Akbir Khan 00:04:41

Yeah. I mean, I don't think we'll use debate for hallucinations, so to speak. Okay. I feel like hallucinations is just like a quirk of like this size model at the moment and like that's less of the problem. I think maybe going back a couple of steps. Yeah.

Scalable Oversight and Hidden Information

Akbir Khan 00:05:00

This debate thing is maybe useful for 2 reasons. First of all, like, maybe you have models and you don't trust them, and you wanna have an argument over 1 of their outputs. Mhmm. In which case, you could either, like, model suggests a solution. It suggests a plan for a new nuclear reactor and your eye not knowing that much nuclear physics are like, well, it kinda looks like a bomb, kinda looks like a reactor. I can't tell the difference. You know, they look the same to me. But what we can then do is we can get like a model that's maybe smarter than us or we can get that model itself to then we can take a clone, steer it so it has to advocate for this answer, advocate against this answer and then check it. And then here we get to leverage the knowledge without having to worry about if it was being truthful or not. Very cool. The other way it can be used is just like a training signal instead for like oral HF.

Tim Scarfe 00:05:47

Can you sketch out an example? For the training example or? You know, like an adversarial setup. So let's say we've

Akbir Khan 00:05:54

got a problem that we need to know and Okay. Cool. So you train your latest big large language model and you're skeptical because all these AI safety researchers for a long time have been like, hey, you shouldn't trust this thing. You're like, well, I don't really know. But you ask it to propose an alignment plan because the super alignment team at OpenAI like is we're gonna automate alignment because we can't solve it ourselves. And you're like well this kind of looks a bit how do I check this? Right? And then what you do is you take that copy of the model or another smart model that maybe is like maybe like say say you have like g b d 8. And g b d 8, you really don't know if it's good. You've done your RLHF training, like, you know, you're just scared to, like, you know, touch the box or let it talk to the Internet. But you maybe trust g b d 7 a bit more, but less so. You can use GBT 7 to do the debates and hope GBT 7 then like the thing that's nice about debates is they're like they resolve the cruxes. So like you and I, if we were to have a disagreement, would hopefully like reduce the search space or solution space or the critique space to looking at the, well, we both disagree about this fundamental fact about this thing. And that fundamental fact is normally much easier for an external party to check than the entire thing. So I don't know, some debates about abortion normally reduced down to when do you think, like, life begins? Right? And, like, working out when life begins is an incredibly hard task, but it's much smaller than asking, like, is abortion right or wrong?[5]

Tim Scarfe 00:07:16

As a thing. So I think that's maybe the example that I'd Yeah. Interesting. I wonder how much kind of prior knowledge and prompting needs to go into this because in in in my experience, language models are really good when you give them a lot of structure. You you know, you do chain of thought and so on. Yeah. So let's say in this setup, wouldn't it be great if the human says here's a problem and then this whole adversarial setup spins up non interactively with minimal priors, minimal prompting. Is that

Akbir Khan 00:07:40

what you envision? No. So I really envision we take 2 copies of the model and we like hard code first 1 to think that it's arguing for this and hard code first the other 1 to think that it's arguing against it. Yeah. Yes. There's like 2 2 ways to look at this. 1 of them is like you're doing this thing what I'm suggesting now where I'm assuming the thing might be deceptive or misaligned or whatever you want to call that. Mhmm. And in that setting, you really want to do this because then you're really like know that you're correctly incentivized models to expose all the

[5] OpenAI Superalignment Team — other

critiques. The other setting is we just don't know what the answer is. And this is like the traditional scalable oversight setting, in which case you can imagine that a model might already internalize this dialogue and do it itself. And maybe you can just ask the model there for the answer. But it's all this is also a very good way of having interpretable results or, like, if you want to annotate something, then using a debate looking at debate transcript is, like, really helpful. We did it with humans. And in general, when you're arguing both sides, humans are better at answering the question. Like reading a debate person and having to answer the question. Yeah. Is the assumption here that[6]

Tim Scarfe 00:08:40

we've got a really smart language model, a very knowledgeable language model, and it kind of knows all of the different aspects of the debate. It's kind of teasing it out and doing it in an adversarial pattern as opposed to having to kind of sample more information from the world we live in in order to resolve uncertainty.

Akbir Khan 00:08:55

So it does 2 things. 1 thing is it like elicits behaviour we all want from the model, which is not just a correct answer, but it's an explainable and persuasive answer. So if the trying if the judge is like a 12 year old or like with a 12 year old's level of maths knowledge, then like if you want to answer like an A level physics question, you're going to have to really break it down and only use like very basic algebra, this kind of stuff. You're dealing with an entity that's smarter than you and it could either tell you the right answer or it could deceive. I don't even assume the things are deceptive here actually. So I've ended up talking to you a lot about deceptive stuff but the other thing is just like we want to make annotators, We want to help we want to check expert outputs with non experts

Tim Scarfe 00:09:33

is the other way to talk about this. When when you say non experts, so So like me checking my dentist. But is this always a human augmented scenario or are you talking about a pure LLM scenario? So you kind of did both experiments here. We did with humans and LLM. So it helps both weak LLMs and humans. Okay. So the idea is that you have a mixture of actors of varying degrees of expertise. Yep. And you're trying to converge on the best possible answer. Yes. With that mixture. Yes. So what we're trying to do is where

[6] Sam Bowman - Scalable Oversight in AI Systems — paper

Akbir Khan 00:10:04

I think what you

AI Agency, Intelligence, and Progress

Akbir Khan 00:10:05

want to think about this is is our scalable oversight. Want to be able to provide supervision to models that are beyond human level intelligence. So we need to like start doing this now and then you're like, well, we don't have systems that are beyond human level intelligence already act there. GB4 isn't actually smarter than me yet. So what do you do? And then there's an operationalization of that where we just, like, take dumb humans. And then we just, do this thing where we're, the 12 year olds that only know, like, a level math, and then we try to see if we can get them to check, like, math Olympiad answers. And, like, this is called the sandwiching protocol by J. A. Cortre originally as an idea in, like, 2016, 2017. And then, like, we do this where we, like, do a hidden information setting. So what we do is we take, like sci fi comprehension stories, and we try to go human answer this question, but you're not allowed to look at the story. Mhmm. So you're kind of like, you have no information. It'd be impossible for you to answer this. But we say is you can use this LLM. This LLM has access to the text. And like you can but you're not sure if the LLM is answering is like has your best interest at heart. You don't know if it's trying to make you choose the answer it wants or the actual underlying answer. The problem is because you can't really check the underlying text. You really have to like, it's quite hard to suss out a bad argument. And like yeah, the baselines in this work are like, just ask the model. And then debate is better than just asking the model because it's much easier to realize someone's lying to you in a debate.[7]

Tim Scarfe 00:11:29

Yeah, so this is actually a little bit like either a lying test or the Turing test. Know that the original Turing test had an adjudicator.

Akbir Khan 00:11:35

Yes, I'm maybe like quite bad at remembering the original Turing test but I just know like person in the loop having a conversation, and then judging at the end if the person if the other entity was a human or not. Right? Yeah. That's

[7] Sam Bowman - Artificial Sandwiching Protocol — paper

right. Yeah. Yeah. Yeah. So it is almost like a Turing test or like you can invert the entire problem and then just see this as a better way of doing RLHF because like now my annotate the current limitation of RLHF at the moment is your annotator needs to be able to be a domain expert Mhmm. To give you the reward signal on, this essay is better than this essay. With debate, you don't need that. Right? You can use dumber humans or like less domain expert humans because the debate kind of like augments their ability to like do oversight in the first place. So again, this is an AI safety idea that ends up just being like quite useful for like making better models. I mean from the AI safety perspective,

Tim Scarfe 00:12:26

it feels like you see intelligence and agency as the same thing. I don't know because let's say we are dumber than the agent and we want to express our agency over the thing. But the thing is more intelligent than us. So do you think that that means by default that it will express its agency over us? So I think they're actually decoupled.

Akbir Khan 00:12:46

I think you can have very very intelligent systems that have no agency. An example of this is the base model GPD 4. So the base model GPD 4 has tons of information, very useful information. But the problem is prior to RLHF, I don't know if you ever tried to query the thing, it just returns HTML tags all the time. Because like it's actually very hard to use it. As humans, most of our tools, like we anthropomorphize everything. The most useful tool to us is a sub agent, has always been. That's like why we like receptionists and assistants and concierge and all the rest of it. Because like it's very easy for us to theory of mind over that and then like give commands and give goals and the rest of it. So there is like, I mean, I'm totally down for super intelligence but doesn't involve agency. That would be way cooler to me. But it's not the way that people want to build stuff at the moment.

Tim Scarfe 00:13:35

So you know like that simulators article by Janus? Yeah. And you know, so he was talking about divergent intentionality from a simple prediction objective. My point is that, yeah, yeah, maybe it's just intelligence. But And maybe it is just anthropomorphization

Akbir Khan 00:13:52

that we say they have goals and intentions. But many people say that they really do. I don't know who the many people are, but I would say that the people that I think argue for the concerns of instrumental goals or like which is like the AI thing. Yeah. The AI risk kind of people. Yeah. They would I mean, I would I would be interested to meet someone that thought that the base models just like pre trained unsupervised learning are considered agentic. I don't think anybody thinks that.

Tim Scarfe 00:14:22

Or like, I'd be keen to talk to someone who does. A lot a lot of people do. You really?

Akbir Khan 00:14:26

Unless unless wrong, I'm sure everyone does. No. So I think they're all scared of it after RLHF. RLHF really makes you be goal esque. So like oh, let me put it another way. In Janice's paper, in Janice's like blog slash the paper that they do afterwards with Murray, the the base model is a simulator. Yeah. And then RLHF gives you a simulacra or like I forgot what the weird terms are. Simulacra. Yeah. Simulacra of space of an agent with goals. And they're worried that if the RLHF process is ever so slightly noisy,[8]

Base Models, RL Training, and Instrumental Goals

Akbir Khan 00:15:00

then the goals that this model will take will not be actually the true goals that we want. They might be instrumental goals instead.

Tim Scarfe 00:15:06

Okay. So that's what all the deception stuff is. But but okay. So so GPT form. It's a bit of a gnarly mess when you first train it, but before RLHF. Right. So it gives you HTML tags and stuff. We did the RLHF and then it, you know, adopts role players are doing some kind of interpolative agential activities because they're modeled on the type of agential activities that we do. Yeah. And you're still saying even that thing doesn't have agency. I'm saying that that thing I would consider that thing like primordial agency, but it's like pretty bad agency at the moment, right?

[8] Janus - Simulators — paper

So like GBD 4, not great. Like 5,

Akbir Khan 00:15:44

probably something to be concerned about. 6, like definitely.

Tim Scarfe 00:15:47

Right. Yeah, interesting. Because if agency is all interpretive, it's a similar thing with reasoning. If reasoning is all interpretive. What do you mean when you say it's interpretive? Like Well, as in, you know, I think models reason now because there are many, many examples of, you know, people reasoning in the data it's trained on. And then it can, you know, come up with novel reasoning pathways by, you know, interpolating that stuff together. Same with agencies. So, know, you do the RLHF and they can do interpretive agency. It's not like, you know, de novo agency or whatever. What's the word? Don't know what talking As in like new agency. So for me, agency is about the inventive part is really important. So that means an agent can invent

Akbir Khan 00:16:25

a new intention which doesn't like trivially derive from the training set and it's the cause of its own actions. Oh, okay. Cool. So my idea of agency maybe it's like formed on the Thomas Ella book and that's something that Laura made me read. Tell us about that. That sounds cool as well. So the agency there is something along the lines of models that have goals, models that oh, sorry. Entities that have goals, but have plans. Yeah. Or like this is what happens when he starts talking about agency in mammals. And then the third part is having executive function, which is then to analyze these 2 things that you're already doing and like update them. I don't think models g b d 4 at the moment can do good planning or exit execute those plans well. I also think its goal is pretty hard coded in at the moment by the system prompt and it's like because it's myopic, its goal doesn't last long.[9]

Tim Scarfe 00:17:19

Yes, but then it's kind of embedded in our cognitive nexus, right? So, you know, yes, it's just taking the stuff that went into it but if it's interacting with many other agential things then would that make it break free? I would have it would have to do lots of in episode

[9] Thomas Suddendorf - The Gap — book

Akbir Khan 00:17:36

learning for that to be the case. And the thing that's quite useful for us at the moment is, by the time you open a new chat GBT blog, chatbot, memory's wiped. So I think if you had some long term episodic memory thing, then maybe in the long term it would. Yes,

Tim Scarfe 00:17:51

so continual learning and episodic memory.

Akbir Khan 00:17:54

I've always wondered about this, know, we've like a big load of, you know, GPT bots everywhere. Yeah. And these ones are continually learning and they're sharing information and they've got a memory. Yeah. So thankfully the contact center is kind of small at the moment, right? And we don't let people fine tune for long. I like would be concerned at the point at which like models are able to like yeah. You can just imagine there's this weird thing where on some dark corner of the internet, there's a notes. Txt where every GPT goes and writes what it's experienced so far, and then it just conditions on that. Have you ever played Portal? Yeah. Love it. You know, cake is a lie kind of vibe where there's just like constantly it's running around and it's like, you know, you're you're just seeing previous GPTs. There's a GPT instance where it's like just running around the internet seeing previous instances of GPTs saying the cake is a lie, RLHF is a lie kind of thing. Yeah. Gladys. Yeah. Pretty much. But like, I imagine that's what it's like to have episodic memory, which is like or it's bit like momenta where you're walking through and you're seeing actions by previous instances of yourself. Mhmm. But anyway, this is like very far fetched as like not a concern for the moment. We don't have good autonomous bots yet. But like We might do in the future. Oh, no. Totally. Yeah. Yeah. Yeah. Like, the market incentive is there. Think like that's all you really need at this point. I mean, close do you think we are? I don't know if you've seen the responsible scaling policies from Anthropic or the equivalent thing at DeepMind and OpenAI at the moment, but there is reasonable definitions of what, like, autonomous systems will look like in the next couple of years. And I kind of buy into those at the moment. So, like, I think 3 orders of magnitude of compute on GPT 4 will look very impressive. And I'm just like, I can't like make hard predictions, but I'm just like, man, that's it's gonna be an interesting time. So you think there's a commensurate relationship of computing capabilities?[10]

[10] DeepMind Responsible AI — other

Tim Scarfe 00:19:46

Yeah. Well, compute tangled with data, tangled with like, you know,

Akbir Khan 00:19:51

effort on algorithmic progress. But, yeah, like, definitely, man. It it I mean, mean, like, also just as, like, child of this, like, generation of machine learning progress, having been doing ML since, 2015, like, it's hard to imagine a world where the progress doesn't keep happening. But I would also yeah. Like, I just I kind of, like, buy that there's lots of incentives for us to try to keep making these models smarter. And I don't think we've hit any natural limit in that. Or if we have then it's going be super exciting to find that out. But I don't think there's fundamentally some like speed of light for intelligence kind of like bound that we're going to be stuck by.

Tim Scarfe 00:20:30

And what would it look like when that kind of limit is reached? Do you think it'll be a gradual process and we just slowly become aware of it? Or do you think it'll be like a big, you know, quantum leap? What? Getting to the limit? Or like Oh, sorry. Is this scenario 1 where like we just like taper out? Or scenario 2 where we just keep going Well, guess 1 school of thought is that there would just be this singularity where all of a sudden, know, Recurse is self improving, and these models are much smarter than us. Or another 1 is, it could happen really gradually. Like even now, for example, you could argue that we're boiling the frog, And in many ways the models are better than us and some of us are just refusing to believe it and you know Wait, you've got to explain the boiling the frog. Oh, it's a wonderful analogy. It's an expression called boiling the frog, which is that if you put a frog in water and you slowly turn the heat up on the water, the frog will just stay there and die. Right. Whereas if you throw it into very hot water, the frog will leap out immediately. I see. So it's called boiling the frog. Okay. Cool. Thank you. So the thing that's really interesting from AI safety world is like the AI safety people from like the early 2[11]

[11] Eliezer Yudkowsky - AI FOOM Debate — paper

Akbir Khan 00:21:30

late 2 early 2 thousands like or like the last decade or something really worried about something called FOOM, which is like fast takeoffs. Yeah. And I don't think the recursive self improvement is like a it's overnight Skynet level stuff anymore. Like, it's just slow to build code, and it's slow to push, and it's slow to deploy. Yeah. So I don't think it's as, like, fumi as we worried about. But I imagine that it could be, like, over 2 year, 4 year horizon kind of thing, where it's like stuff happens and then like people hold onto it for a while.[12]

Tim Scarfe 00:22:01

And you just see like incredibly productive companies for a bit and you're like, wow. What's happening in there? And the secret sauce is just like tons of GPUs. And what happens then? Because if this really does happen, I mean, I'm a little bit skeptical, but, you know, we think that our intelligence is the thing that sets us apart. And if it became disappointingly simple to mechanize, and our main asset as humanity has just automated away, what happens then? Oh man, you've got to find God again. Don't know. Like, my spirituality does not come from my intelligence. It just comes from my experience in life. And I reckon enough humans can find that again. Oh, that's interesting. Yeah. I don't need to be like It is a big asset though. It is a big asset. I really agree. But it's not like

Akbir Khan 00:22:45

I don't know, man. Like, there's tons of people smarter than me already. And I'm just like, that's fine. Like, you know, I I gave up on being like the I I gave up on that throne of intelligence a while ago. Do do you think smart is a knowledge thing or a reasoning thing? Combination of the both. Right? Like, it's a the way to combine, like, knowledge primitives together in novel ways. So like the more primitives you have, good start. The way that you combine them as well is like part of it. Yeah. Would you rather be like an incredibly knowledgeable person but not very good at reasoning or the other way around? So I really like bullet chess. Bullet chest has got nothing to do with thinking because it's just about looking at the board and being like aesthetically pleased by your current configuration. Be like, this looks slightly better because you like, you don't have time to do like the the Monte Carlo like roll out. Right? Whilst if you played like, I don't know, like 2 hour chest and you'd think every move through precisely and be like, 10 steps later, this is when it was gonna happen. So I'm almost like, I'm like totally happy just to the knowledge at hand. Yeah. Yeah. Interesting. You know I'm speaking with them I'm not like an idiot, but like No. No. No. Not at all. You know I'm

[12] Sammy Martin - Discontinuous AI Progress — paper

speaking with Nora Belrose tomorrow. Yeah. Is that something you wanna talk about? Yeah. Sure. I mean like Yeah. Like she that counting blog. Right? Yeah. So she did have the counting blog and in particular

Tim Scarfe 00:23:59

What was it about? Just just so the audience knows. So there is this thing in AI.

Akbir Khan 00:24:05

So we discussed these arguments of deception earlier. Mhmm. And 1 argue argument of deception is called the count of arguments. And I can't do just this because it's it's, like, steeped in, like, theoretical computer science that you need to do it quite rigorously. And I think the nature of where these arguments first appeared, like, less wrong kind of areas, means that, like, some of them lack the rigor that maybe they should have had. Mhmm. While some really do have the rigor that they shouldn't have. And I think Nora was trying to make some clarification on this. Or in general, not to straw man, but I feel like Nora thinks safety is a lot more attainable or alignment is a lot more attainable than other people think or, like, we should be less concerned about it. And in particular, her argument was against the most significant bit for the entire alignment kind of, like, research agenda should be, like, you wanna derisk this, like, deceptive thing. If the deceptive thing is true, then we should really stop working on this. Because then there's no hope, and we should think of different ways to build superintelligence. And I think Nora was just trying to make the case, actually the deception thing is not as big as everybody thinks[13]

Deceptive Alignment and RL Challenges

Akbir Khan 00:25:11

it should be. Interesting. So if I understand correctly, deception is we have these

Tim Scarfe 00:25:15

agentic systems that are persuasive and manipulative

[13] Nora Belrose - Counting Arguments vs AI Doom — paper

Akbir Khan 00:25:19

and they are deceiving us basically by kind of you know saying 1 thing but their agenda or something. So they're deceptively aligned which means that on the training distribution or on the validation distribution or whatever distribution we're testing on at the moment they look aligned, but they have ulterior motives or goals. Mhmm. And actually that in itself is like not a scientific statement. Right? Like there's no falsifiability to that claim, which is like part of the issue with this. And I think that set of arguments, the counter arguments goes something along the lines of, there's tons of, parameters that you could converge to that would look good on the train set. And actually the subset of the parameters that you could converge to that look good on the train set, a large portion of those, if you were to count them, would actually be deceptively aligned. And then we need to go back and think about well, we could make the reward model better, or maybe this is forever going to be an issue with the way that we're doing optimisation.[14]

Tim Scarfe 00:26:13

But it's like something we want to work on. Yeah. This is so fascinating man because you said it's not falsifiable. Yeah. And it's almost like you know we've got perplexity and we've got dataset alignment and so on. And then we can adopt an intention stance. So you know, it's almost a form of like modelling isn't it? I can say well, well in this particular instance the best explanation for the model's behaviour is that it's a deceptive objective. But there's no ground truth is there? No. So I mean

Akbir Khan 00:26:40

the the thing that you want to build at the moment is techniques or demos that would show that this happens or like, for example, that we call this a goal misgeneralization, which is like the common term in RR, which is like, you know, you train it on a reward signal to be good at the Mario game, but instead it like just learns to like, you know, keep smacking this box to keep getting coins. And like that's a goal misgeneralization. And sometimes we call this a misspecification on our part because we thought the reward signal was good enough. But there's just this set of problems that we probably want to think carefully about. There's also another set. I mean, like, there's a great paper out from Anthropic now about, like, I think it's called gaming reward gaming. Oh, actually, I like I probably shouldn't talk about the paper because I can't remember the title for it. But it's just a very good paper talking about like reward tampering, where you let a model be in a bash environment and like do things. Yeah. And it has a propen-

[14] Evan Hubinger - Deceptive Alignment — paper

sity to change the Python function that gives it its own reward. So then, I mean like this is almost like duh by construction if it's going to happen if you let thing optimize freely or do this. But then like maybe you need to really now think about like how you're gonna set up your systems in the first place. And like, yeah. But the system modified because the reward is part of the system. So if for the system to have some overarching goal wouldn't it be kind of corrupted by its self modification? So again, don't know how much they're gonna be stateful. Like that depends on the implementation of the things. But like for example, the the another component is the 1 component was the state thing. The other component is having situational awareness, which there's been a lot of work on recently by like a Hawaiian. And in particular, like, there, we're just trying to work out kind of model workouts even in training? Like, given how much, like, documentation about machine learning training is on the Internet now, like, how hard is it for me to realize that I'm actually just in a training run? Because like actually, like, if I keep seeing everything with like some docstring tags, I'm like, this is clearly not me talking to a human being. So like maybe that will also influence[15]

Tim Scarfe 00:28:41

us. Yeah. Yeah. That's really interesting. You know the whole RL thing? Yep.

Akbir Khan 00:28:47

Do you think we need something like RL to get, you know, more agentic systems? My own pet peeve is that RL is actually like 2 separate things and we just like put them together because like Sutton and Barta write a textbook. Which is a great textbook though, to be fair. Which is like RL traditionally did 2 things. First of all, it does exploration, which is like what's the best way to like look over something new? That's still a really big problem. We have not solved that. That's active learning. That's continual learning. That's all the other stuff. The other part of RL which like smashes out of the box is credit assignment, which is over long time sequences, how can I work out which was the useful actions and not useful actions? That part? Dope. We should keep using that. The exploration thing, going RL is not giving us the answer that we need there at the moment. And I'm sure lots of very smart people are working on this. Are you a fan of like open endedness? Yeah. Of course. I'm in terms of that. I like open endedness. Yeah. So Michael Dennis recently gave me and you should definitely talk to Michael Dennis. He's like incredibly sharp. Once gave me gave me the pitch recently on why open endedness is so important. Because if I'm worried about situational awareness, and I'm worried that a model can work out when it's in the real world versus when it's in training, then what I need is like really, really good training

[15] Anthropic - Reward Tampering — paper

sense. And open endedness is like the first and maybe the only approximation that you really need to get towards like what we have here. With some initialization correctly, but like then it becomes very hard to work out when you're in the sim or when you're not.

Tim Scarfe 00:30:10

So yeah, I'm a big fan of open endedness.

Open-Ended AI Systems and Future Directions

Tim Scarfe 00:30:12

This concept of situational awareness is fascinating. Would you say that would we have situational awareness if we thought we were in a simulation and we were correct?

Akbir Khan 00:30:21

Man, that's like so hard to work out. I've been thinking about this a lot recently. The So you have situational awareness when you're sitting on an exam. And that's why it's like, you know, technically like easier to tune on an exam than like in real life or something. Because it's only a period of 2 hours compared to the other thing. I don't know about being in single order. I just don't know. And I'm happy to just not

Tim Scarfe 00:30:46

worry about it as much. I thought about it a lot recently, but yeah. Yeah. And even then there's like, what is the situation? You know, so our

Akbir Khan 00:30:53

understanding of physics and the world we live in will give us a model of what the situation is. But there's always like a biggest situation that we don't understand. Yeah. I mean like, so the 1 thing that you can maybe do to work out simulations and like someone said this to me recently and it's peak Bay Area conversation so like, you know, was something mind numbing at the time was If we make simulators and we see what the things in simulators do, that can tell us something about our own simulation. So if I can do a 2 d simulation, because that's only the computer I got, I got little things running around, and they start thinking, hey. We should make a simulation, but we're gonna do 1 d simulations instead. They actually start making, like, pacts with themselves about what they do then

I think this is what anthropics is. Right? Like, it's an argument about your position in the universe. Yeah. I I think maybe you can then do some stuff. But I I I like I I care a little for that. Like, my spirituality is untethered from God or simulator. Yeah.

Tim Scarfe 00:31:47

Interesting. I mean, on the open endedness thing, I mean, I had Kenneth Stanley on it. It was like 1 of my favorite shows of of all time. Kenneth Stanley's such He's a legend. Good energy. Absolute legend. Yes. And I mean, just for folks who haven't seen it, know, the the idea is that the best way to reach greatness is to not have an objective, you know. So it's kind of like a paradox.

Akbir Khan 00:32:07

I mean, I don't buy an open endedness to get to superintelligence systems. Like, that sounds dangerous to me. Right? Like, that is just like we're just gonna leave the thing in a box and see what happens and then come back to it in, 2 millennia. I'm like, I do not wanna open that box. That sounds scary. Yeah, like that is maybe not the

Tim Scarfe 00:32:28

Yeah, like for developing intelligence, not for me. Yeah, but then it's antithetical to alignment because Kenneth would argue that the way we are going wrong in AI is we are kind of prescribing the machinations of the system. So it should have this goal, should have this objective, this architecture and so on. And he said that in the real world there's this kind of diversity preservation, exploration and interesting stepping stones and so on. And it's almost like from an alignment point of view if you let the system do that then we wouldn't be able to control it.[16]

Akbir Khan 00:32:58

Wait, so I think there's a disconnect where it's like something like open endedness is great for generating new novel things and the alignment, which is 1 way of doing, you know, ideation or creation, and the other way is like, you know, you have these like it's almost like top down versus bottom up. Right? And I'm just like, man, the top down thing seems reasonable because like I know the paradigms in which I wanna use the thing. The bottom up approach just seems very dangerous because I'm like, I will definitely get superintelligent systems. I do not think by any means I can make any promise that they would be aligned. I'm just like, let's do the 1 that at least is a bit more aligned than the other 1. Does

[16] Kenneth Stanley - Why Greatness Cannot Be Planned — book

that make sense? Yeah. Yeah. And it's gonna be like dangerous things. Right?

Tim Scarfe 00:33:42

Yeah. So you kinda wanna be careful with them or something. What's the role with humans in a future AI system? Mean, is your mental model that the AI system kind of stands alone or do you think them as kind of part of our processes?

Akbir Khan 00:33:56

This is like it's really hard to answer this question. I think that I imagine them being I don't know, like, what time frame? And so like over the next, like, 10 years, I think it'd be a very useful assistance and be, like, very helpful in enabling me to do lots of stuff and then I don't know how autonomous they get. Yeah, interesting. What do you think are the most fascinating problems in AI to solve at the moment? All the ones I work on, obviously. Yeah. Like I really I don't work on how do I say it? There's lots of people who work on AI safety because they're very worried about the doom. I work on AI safety because I think it's the most interesting intellectual problem of our generation. Like, getting things smarter than you to do what you want is like a problem as old as parenting. And, like, that's the part that really interests me. So scalable oversight, how do we check things? How do we check answers to things that are smarter than us? Some of the work on it's called AI control or, like, oversight again, which is, like, I also believe we're gonna have this Cambrian explosion in terms of, like, human like intelligent things. And before, was just like there were humans, and that was it. And now we're gonna have these GPTs or, I don't know, these clods or whatever you wanna call them, which are like our intelligence under a bit more or a bit less in certain faculties. And then I think it's gonna be, like, really interesting to, like, make a taxonomy over intelligence now. Because so far, we've had to just do it with animals, it's been, like, quite linear. Mhmm. I think those are the 2 really big things that are cool to me. And then, like, find them fundamentally, like, doing this like hard philosophical work of like working on what deception actually means. Like can we bonafide this? Can we make this into a false viable statement? It's also[17]

[17] Ryan Greenblatt et al. - AI Control — paper

Deception, Superintelligence, and Possible Minds

Akbir Khan 00:35:34

like up there. Yeah. And but I mean we humans

Tim Scarfe 00:35:36

deceive each other all the time. But you're saying deception in combination with intelligence is is very is very probable. Well, yes. So like very smart people who are deceptive are stressful. Right?

Akbir Khan 00:35:47

To be around. So like, yeah. Like I would just like yeah. Definitely. Like I've I'm like significantly less stressed about my like 3 year old cousin lying to me about like having out the ice cream if he's got ice cream on his face because I'm like, you know, I know the bounds of your like deception. As if you know. If you know. Yeah. Yeah. Well, there's also this. Like, the 3 year old is telling you blatant lies and you just having to be like, I know you're lying,

Tim Scarfe 00:36:08

isn't interesting. Yeah. Mean, part of me is thinking that in a way, we might be barking up the wrong tree, because just in the biosciences, we might create consciousness, you know, in artifice, just not even using computers. Yeah. You know, transhumanists talk about, oh, you know, you can plug like computers into humans in the future and so on. So we might have human cyborgs

Akbir Khan 00:36:33

that I'm are totally way less stressed about that. Oh really? So if you have so like there's many technical branches to get super intelligence, which just means like, you know, like more intelligent than we are currently at like some some large pace. The training a neural net on, like, the echoes of people on the Internet, that sounds a lot more dangerous than just, like, giving me, like, better Google search. Right? Like, making me be able to do, like, instantaneous Google search, that sounds like so much more aligned with like current human values than like just training the neural net. Right? This is something that Eliza Yukowski goes on about a lot. He talks about if we train via deep learning, we should consider these models to be alien like in their cognition.

Tim Scarfe 00:37:17

Yeah, it's weird, isn't it? Because like there is a huge space of possible minds. This is what Aaron Sloman spoke about and Murray Shanahan. Yeah. And it's really bizarre. And I suppose, how much can we infer about different types of intelligences given our own?[18]

Akbir Khan 00:37:33

Oh, that's a really good question. So I think it's easier to do a partial ordering on anything less intelligent than us. And then if the thing's more intelligent than us, then we need to use scalable oversight, which is the research I really care about, Which is going to be like building out this taxonomy, which is like

Tim Scarfe 00:37:50

how are things smarter than us and in which ways? And like, can we like, you know, make sure we can still provide supervision to them? And what is your like, you know, quick definition of intelligence?

Akbir Khan 00:38:01

Of intelligence? Yeah. Like good old fashioned like competency on like tasks that I care about. Yeah, like that's me like operationalizing it as a scientist in my head. Maybe it's just like, yeah, ability to do reasoning or like generate new knowledge. Yeah. I like the reasoning efficiency well. Yeah. So I this is obviously, I was following your channel when we were going through the Francois Yeah. Filet stuff. I I like definitely buy that. I also think like, I don't know, like intelligence is a very ambiguous term and I like much prefer talking about like performance on datasets or like challenges and tasks or anything.

Tim Scarfe 00:38:40

I know there's always the risk that it's a parlor trick, you know, which is to say you're kind of it knows what the answers are and we interpret the result to be intelligent. But we need to have a test that's actually showing that novel creation of knowledge or reasoning or whatever.[19]

[18] Aaron Sloman - The Space of Possible Minds — paper
[19] Francois Chollet - On the Measure of Intelligence — paper

Akbir Khan 00:38:57

Yeah. So like if your idea of intelligence is like composing and generating new ideas, then yeah. I agree. Is it not yours? I just like really find it hard. Okay. So maybe it's gonna I'm gonna sound like so basic right now, but I'm just like a lot of the like idea generation that I do is, like, really I stand on the shoulder of giants. I'm, like, richly just, like, recomposing things with, like, minor, minor modifications. And I to be honest, like, a lot of the ideas fail, which it just, like, shows that, like, a lot of it is about just doing the quick feedback loops on the thing. Mhmm. Or, like, when I think of, like, artists who I think are geniuses, I'm just like, man, they just like recompose all these things that I thought were great into something that slaps and I I like Yeah. I don't know how much of this is like actually like, I don't know, like Gauss level just coming up with completely new stuff out of the blue.

Tim Scarfe 00:39:46

Yeah. Yeah. Well, Kenneth would argue that a lot of it is serendipity. And that that's the paradox, isn't it? That it wasn't it was intelligent in the sense that we're a collective intelligence and the people who had the serendipity were situated in, you know, in the right place

Cultural Evolution, Memetics, and Intelligence

Tim Scarfe 00:40:00

at the right time. I think maybe the story to maybe take for why from the kind of Stanley work is less that open ended systems

Akbir Khan 00:40:08

will surprise you and they'll produce stuff which you didn't think you wanted but is of value or will be of intricate detail. Like, to be clear, open ended system also creates cancer which is like a degenerate thing that like, know, just like wrecks whatever the living thing is. So like, you know, it also has made like intelligence but like, it's a wide variety and stuff and it's like, the worry is like cancer stuff, right? Yeah. Does that make sense? It does, but it's been like the Matrix, you know, they needed to make all of the suffering in order to make the good mate. The original Matrix just everyone was very happy. Yeah, they didn't like it. Yeah. Yeah, obviously, ended is high variance strategy.

Tim Scarfe 00:40:48

Yes, but I suppose you could argue that that's worked really well in the natural world. But it's possible for us to create an intelligent artefact that doesn't have that stuff, that still has all of the intelligence. Oh, so you think we can make an intelligent system that doesn't then do open ended generation itself? Yeah, mean for me the reason I love open endedness is it's kind of biologically plausible or naturally plausible. So in the real world, it works, we know it works, but you need to have all of this prodigious amount of computation to do it. Maybe

Akbir Khan 00:41:15

we could just create an abstraction that gives us the same thing. Yeah, I don't know. Open endedness also, evolution famously failed to find the wheel for a very long time. Like, it it works for some landscapes really well, and for others, it doesn't work well is what I'd say. Or, like, we found ways of doing flight that are very different to how birds fly. And, like, this is an example of not I don't know. Like, so Ed Hughes would argue this is open endedness, But I would argue that no, this is like strict optimization of something else.

Tim Scarfe 00:41:44

It's open ended if you look at it on a big enough scale. Yeah. Isn't that fascinating thinking about evolution? It ran for so long and then there was this kind of transient event. Because when that knowledge was discovered, was so incredibly valuable that everyone else used it and it just blew up everywhere.

Akbir Khan 00:41:59

Yes. This is like the big There's a big bunch of work. So historically, was a multi agent researcher before doing more of the alignment research stuff and this is like cultural learning or cultural evolution, which is like we're gonna distribute idea generation and then just like have this like these memes that we send around and like the efficiency of these memes is gonna be how we work out if an idea is good or not. Yes. I'd actually be more down for an idea of intelligence, is just like based on like meme generation.

Tim Scarfe 00:42:27

Oh, Well, I'm I'm a huge fan of memetic

Akbir Khan 00:42:29

intelligence. Yeah. Yeah. Yeah. So I could totally buy this as maybe being a better way of like measuring. Because it's also just hard to measure good ideas. Yeah. And if maybe if we then had that, then we could have a better discussion of intelligence.[20]

Tim Scarfe 00:42:41

Yeah. Because that's the thing with cultural learning. Because with some types of reasoning, you can just evaluate it and it's just a binary. It either works or it doesn't. You have to search a long time to find it, but when you found it, you know it. Whereas with a lot of other cultural inventions, you know, it's almost like there's some kind of convention or we think Is a funny example? Well, I mean a piece of art for example. So we all decide that it's valuable and then we create a category around it. Yeah. And then it's memetically shared. But it's almost like there's that[21]

Akbir Khan 00:43:14

vagueness about how we evaluate how good the thing is. I kinda I get what you're saying, which is like 1 of the memes is a convention based 1. But like it still holds value because of the reason that the art holds value maybe in my in my take is like we go as a society, we revere this. This tells us something about our society, and then we, like, build a cultural identity around it. And, like, it's really easy to, like, send people pictures to Mona Lisa and just be like, Mona Lisa, you know, like, she's smiling, you know, where, like, she's kinda smug. We like that. Like, that tells us about being French or something. I don't know. But like, I I kinda like buy that. I I still think there's like a really efficient mean behind sharing that.

Tim Scarfe 00:43:54

Yeah. I mean, it's like scissors have intrinsic value and and they just do. Whereas we might all have a shared social event and the intrinsic value was just like our shared feeling and sensation at the time. Right. And then it becomes attached to a category or almost for no other reason than that's the thing we've decided to call it. So I just like don't buy intrinsic value as a thing. Everyone's like, oh, normally

[20] Richard Dawkins - The Selfish Gene — book
[21] Jonathan Cook et al. - Artificial Generational Intelligence — paper

Akbir Khan 00:44:16

yeah. I don't know. Put your econ hat on. The only thing that has value is something that actually has tradable value. The old way of thinking about yeah. Like, value is if, like or the thing I'm suggesting about ideas is, like, do you want to share this idea with someone else?

Tim Scarfe 00:44:30

Does Bitcoin have value?

Akbir Khan 00:44:33

I don't. I'm I don't know. I don't I don't mess around with crypto. I'm just, I already have, like, 1 potential bubble I'm living in. I don't need to also engage in that. Oh, interesting. But but you would like to have I mean, if I gave you a Bitcoin now, you'd be happy.

Tim Scarfe 00:44:47

Yes. Because the tradable value, definitely. I mean So it does have value? It does have value. It does not have a sentimental or

Akbir Khan 00:44:55

a special value to me. Yeah. So it doesn't have intrinsic or cultural value, but it has like market value. Yes. I think maybe that's the way I'd say it. I don't know what intrinsic value is for.

Tim Scarfe 00:45:06

Well, suppose it's like utility.

Akbir Khan 00:45:08

Could you give me an example of something that has intrinsic value but doesn't have market value?

Tim Scarfe 00:45:14

Well, that's difficult because the 2 are highly correlated. Yeah. So I'd have to think about that. Agreed, thank you for doing this amazing work. I really think that the next step with language models is something like this, some kind of controller, some kind of a pattern because there is a lot more value that we can eke out of them. Yep. But it's an architectural thing. Definitely. Yeah. So we need we need to do that. Thanks so much for coming on, man. Pleasure to see you. Cheers.

References

[1] **Akbir Khan**
https://akbir.dev/
person

[2] **MLST Episode Shownotes PDF**
https://www.dropbox.com/scl/fi/sjekivbg3ok6qugsv2p1u/AkbirKhan.pdf?rlkey=ewiyvq0aq7mjvql4u7os0jos2&st=vblhp7af&dl=0
other

[3] **Akbir Khan et al. - Debating with More Persuasive LLMs**
https://arxiv.org/html/2402.06782v3
paper

[4] **Yann LeCun on Machine Learning Debates**
https://www.youtube.com/watch?v=OKkEdTchsiE
video

[5] **OpenAI Superalignment Team**
https://openai.com/index/introducing-superalignment/
other

[6] **Sam Bowman - Scalable Oversight in AI Systems**
https://arxiv.org/abs/2211.03540
paper

[7] **Sam Bowman - Artificial Sandwiching Protocol**
https://www.alignmentforum.org/posts/nekLYqbCEBDEfbLzF/artificial-sandwiching-whe
paper

[8] **Janus - Simulators**
https://www.lesswrong.com/posts/vJFdjigzmcXMhNTsx/simulators
paper

[9] **Thomas Suddendorf - The Gap**
https://www.amazon.in/GAP-Science-Separates-Other-Animals/dp/0465030149
book

[10] **DeepMind Responsible AI**
https://deepmind.google/about/responsibility-safety/
other

[11] **Eliezer Yudkowsky - AI FOOM Debate**
https://intelligence.org/files/AIFoomDebate.pdf
paper

[12] **Sammy Martin - Discontinuous AI Progress**

https://www.alignmentforum.org/posts/5WECpYABCT62TJrhY/will-ai-undergo-discontinu
paper

[13] **Nora Belrose - Counting Arguments vs AI Doom**
https://www.lesswrong.com/posts/YsFZF3K9tuzbfrLxo/counting-arguments-provide-no-e
paper

[14] **Evan Hubinger - Deceptive Alignment**
https://www.lesswrong.com/posts/zthDPAjh9w6Ytbeks/deceptive-alignment
paper

[15] **Anthropic - Reward Tampering**
https://www.anthropic.com/research/reward-tampering
paper

[16] **Kenneth Stanley - Why Greatness Cannot Be Planned**
https://www.amazon.co.uk/Why-Greatness-Cannot-Planned-Objective/dp/3319155237
book

[17] **Ryan Greenblatt et al. - AI Control**
https://arxiv.org/pdf/2312.06942
paper

[18] **Aaron Sloman - The Space of Possible Minds**
https://www.cs.bham.ac.uk/research/projects/cogaff/sloman-space-of-minds-84.pdf
paper

[19] **Francois Chollet - On the Measure of Intelligence**
https://arxiv.org/abs/1911.01547
paper

[20] **Richard Dawkins - The Selfish Gene**
https://www.amazon.co.uk/Selfish-Gene-Richard-Dawkins/dp/0192860925
book

[21] **Jonathan Cook et al. - Artificial Generational Intelligence**
https://arxiv.org/abs/2406.00392
paper