/mlst

Taming Silicon Valley - Prof. Gary Marcus

March 24, 2026



Machine Learning Street Talk

Podcast Transcript

Revision: ea5cc618

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 Flaws, Improvements, and Industry Critique 3

AI Safety Theater and Image Generation Issues 8

LLMs Lack World Models and Human-like Understanding 11

Superficial Intelligence vs True Reasoning 14

AI in Specialized Domains: Chess, Coding, and Limitations 15

AI-Generated Code: Capabilities and Human Interaction 18

AI Regulation: Industry Resistance and Oversight 21

Copyright Issues in AI and Tech Business Models 23

Societal Impact: Risks, Misinformation, and Ethics 24

AI X-Risk, Alignment, and Moral Principles 36

Persistent AI Flaws: System Limitations and Architecture 41

Surveillance, Economic Challenges, and Neuro-Symbolic AI 45

References 50

Overview

Gary Marcus, cognitive scientist and one of AI's most prominent critics, sits down with Tim Scarfe for a nearly two-hour dissection of the current state of artificial intelligence and the tech industry that builds it. Timed to the release of his book Taming Silicon Valley, Marcus makes a sustained case that large language models remain fundamentally brittle — impressive on the surface, yet lacking the compositional semantics, object permanence, and genuine world models that would make them reliable.

The discussion ranges across AI safety theater and the gap between marketing claims and actual capability, the persistent failure modes of image generation and code synthesis, and why LLMs playing chess badly tells us something important about what they have and have not learned. Marcus is especially pointed on industry resistance to regulation, walking through the politics of California's SB-1047, the EU AI Act, and the structural reasons why voluntary self-governance by tech companies has not worked.

The second half turns to the societal damage already visible: copyright battles, the firehose of AI-generated misinformation, surveillance capitalism, and the erosion of trust in digital information. Marcus and Scarfe spar over x-risk and alignment, with Marcus arguing that near-term harms deserve more attention than speculative doom scenarios. The conversation closes on the prospects for neuro-symbolic AI, the legacy of thinkers like Chomsky and Piaget, and what it would actually take to build systems that reason rather than pattern-match.

AI Flaws, Improvements, and Industry Critique

Tim Scarfe 00:00:00

Why do you do what you do? Why after all of these years, you you keep persisting.[1][2]

Gary Marcus 00:00:04

I mean, why did you write this book? Well, the reason I wrote the new book is because I sensed a moral decline in Silicon Valley. That's in fact 1 of the chapter titles is the moral decline of Silicon Valley. I particularly sensed it when Microsoft, released this product called, Sydney, and Kevin Roos had this conversation with it in which it told him to get a divorce and all this stuff. And instead of pulling the product, Microsoft just put some Band Aids on it. And that was a sign to me that things had changed. And within a few days of that, Satya said that we're gonna make Google dance. The whole culture changed, like, overnight. You know, the antecedent condition, precipitating condition was the popularity of chat GPT. Suddenly, you know, people thought there was real money to be made here and their postures changed entirely. And that worried me because I do think the technology is premature. I do think a lot of harm can come from it. And I kind of dropped what I was doing research wise, and really moved full time into policy. And in some ways, the book is actually a memoir. It's not couched that way at all, but it's really a memoir of that time when Gary went to the Senate, had a great conversation with the senators, and then realized that nothing was going to happen. So, you know, I I had this peak experience talking to the senators and and feeling like they were gonna do something, and then gradually bit by bit, and I was warned, you know, that this might come down this way, but I had to see it for myself. I'm naive, I guess, in that at least that 1 respect. Seeing that nothing was actually happening and seeing, you know, learning close-up how the lobbying works. You know, like 1 time I was in DC and I took a meeting with Google, and it was a time when I got the closest hotel I could get to the capital, because I was gonna be, like, in and out, and I had to have a bunch of meetings, and I couldn't, you know, be late for the meeting. And like Google had an office next to that hotel which is like right on top of the Capitol. Like they're so there embedded in it in every level. I mean, that's just a sort of metaphor. And it's not just Google, it's it's OpenAI and and Meta and all. Once you learn how all that lobbying is working and once you see how, like, good ideas that have big support go to die, never even get voted on, like, it's incredibly disillusioning. As the dis-

[1] Gary Marcus - Taming Silicon Valley — book
[2] Gary Marcus Substack — person

illusionment really led to this book on Taming Silicon Valley, I realized that we are heading very quickly towards an oligarchy. I mean, imagine the data that these guys are getting. Like, OpenAI is like getting access to everybody's documents and wants like, you know, everything about you, which they're of course gonna weaponize. They're gonna sell in some, you know, various ways including to targeted political advertisers probably, you know, I mean, they'll say they won't but like we've seen this movie before. They're getting an enormous amount of power. A lot of that is because people think that they're going to get rich, which may not actually be true, ironically. So they've been given power in advance of actually delivering the goods. Like, people think they're gonna make AGI, therefore they should be powerful, but in fact, they've made LLMs which are not AGI that are not actually commercially useful, but they've given enormous power. They're like on all these committees and whatever. And so Silicon Valley has suddenly got a lot of power. They've taken a lot of power from the government, and the government should be like, hey. Hold on, guys. Like, you know, prove yourselves first and, you know, you can have some power but not infinite power. Government's not doing anything about it. EU is is different, but in in North America, not so much. And I watched all of this happen and I thought kind of what the consequences were gonna be, and I realized we just could not count on the government. I had this tension as I wrote this book, this crazy tension, which was if the world went the way that I wanted to, I would have entirely wasted my time writing the book. And like anybody else, I don't wanna waste my time writing a book that nobody's gonna read and is out of date. And I had this fear as an author that the book would be completely undermined because suddenly Washington would get its act together. But of course, it didn't. You know, I would have been happy for the world and saddened for my book. It was a weird position to be in, sort of like betting against yourself or something. I don't know. So so I was afraid maybe things would actually get be done right. We wouldn't need the book. I need not have had that worry for a second because Washington, in fact, mostly abdicated abdicated. Chuck Schumer, particular, you know, had the power to do something here as the senate majority leader and put some strong legislation forward, and he didn't. He he took 8 months of listening meetings and put out a white paper rather than actual law, so he kind of ran out the clock. Now as we record this, you know, I guess, you know, no very little is going to happen because the elections are soon, and nobody, you know, the way dynamics work in Washington, like, nobody wants to stick their neck out because it might hurt them in the election. So, like, basically, this period of great excitement about how we could handle this stuff that started around the time when I appeared in the Senate, which was May 16, has entirely dissipated. It's it's gone with the wind. The opportunity was completely squandered. So the point of the book is we can't trust these companies to self regulate. They don't do the things that

they promised. They you would know better than me, for example, the things that they promised about pretesting to the UK government and then didn't deliver. Mhmm. I think there's a big scandal in The UK. People in in The US may not know about it, but that's 1 example where they have not done, you know, what they said they would do with self regulation. And of course, they'll, you know, weasel down anything that they said so that, you know, it's less invasive to what they're doing and so forth. And then the big issue with governments is regulatory capture or doing just doing nothing. And that's you know, I read the writing on the wall and realized that nothing was gonna happen, and that really logically leaves only 1 possibility to get this right. And we don't get it right, it's gonna be You know, it's gonna be social media but much worse. Move fast and break things. And so the only thing left is to directly appeal to the people, and so that is what I am doing or trying to do. We'll see if anybody cares. But I I am out there talking about this stuff, trying to get the citizens of The United States and some other nations to speak up loudly. Do things like boycotting, for example, and say, look. If if the government's not gonna take care of the artists, the government's not gonna take care of the writers, we are not gonna use software that steals from artists and writers because we know we're next. You know? It's like pastor Neimoller. First, they came for the for the Jews and gay. The companies wanna take everything. They really want the whole ball of wax. They, you know, they're gonna take your keystroke loggers, whatever it is that you do, if you do something that's on a computer, and they're gonna try to replace you. That is the game now. And so if we don't stand up together with coordinated action and say, look, we want a more equitable AI year. This doesn't mean equity like everybody gets the same outcome, but equitable like everybody gets a fair chance and like if they use your IP, you get some compensation and and so forth. There's 1000000 different aspects to this. If we don't stand up and say we want democracy to function, we're not happy with the deepfakes Deepfakes is the 1 place maybe something will happen legally. But if we as the people don't stand up and say, this is really important, and we don't do it soon, and this is really important, we're gonna be screwed just the way that we were with social media, but possibly worse. So the problem with social media is things got entrenched, and we can't fix them now. I mean, yeah, the the the Child Act just just passed. But by and large, like, media is what it is now. There's nothing we can do about it and it's not good. You know, it's probably a net drain on society. It's fun. I use it, but, you know, it has a lot of problems. And AI, which is giving so much power to these companies, and if we don't set the right precedence in the next, I don't know, 12, 24, 36 months or something like that, we are gonna be stuck with whatever comes up, which is probably gonna mean basically anything goes. Like, have section 2 30 says that these companies are not liable for anything on social media, basically. I think was we could go into the story, but it it was maybe well intentioned, but

the world changed. The law did not keep up with the difference between, being a carrier of information and being a company that prioritizes social media feeds and makes more money if it makes things more polarized. Like, the laws didn't keep up, they're bad laws. We were gonna be stuck in the same position. And so the choices that we make as a society right now are gonna have an effect for the next decade, maybe the next century. I wrote this book to wake people up.

Tim Scarfe 00:09:00

Gary, it's an honor and a pleasure to have you on MLST. I love the show. I'm glad to be back. Wonderful. So you gave the keynote this morning at the AGI conference, and it was it was fabulous. So by the time folks at home watched this, you would have seen the keynote. And there was about a 10 minute section towards the end which was absolutely hilarious. But yeah, why don't you tell me about that?

Gary Marcus 00:09:24

About the talk as a whole or about I don't remember what was last minute. Well, look, it was an interesting way of giving a talk. I rewrote the entire thing, but I it was almost like something borrowed something new. So the the context is I actually gave a keynote at this same conference 3 years ago, and it's not that often that you go to a conference twice in a 3 year period. Typically, if you give a keynote, like, it's once every 10 years or something like that. And so I had given a keynote 3 years ago and it's been such an interesting and yet such a disappointing 3 years in AI. You know, most people are excited about it. I'm a bit disappointed and I wanted to explain why. And so I thought about it. I looked at that old talk and it was like almost every word here is still true. I thought about it a little more and I realized that there was something I missed before. So And the talk was kind of divided into 2 parts. 1 was all the stuff that really hadn't changed despite, you know, billions of dollars and enormous excitement and enormous press. And then the last part was what I really missed the first time around, and that was interesting too. So so the the first part was basically, I went a few years ago and I said, everybody's excited about these large language models. This was before they were really big, but they just started being called foundation models. Everybody's excited, they said, well, what should a foundation be? Ernie Davis and I said this together. Well, a foundation should be like something robust that you can stand on. That is what a, you know, a foundation of a house is or a building, and these models aren't that. They make all kinds of dumb errors and and you can't really trust them. That was the talk I gave a few years ago, and I pointed out, like, why there was a lot at stake, like, you know, telling radiologists they shouldn't train anymore, like hyping these things actually has a cost for society. So I wrote that whole talk before, and I looked at it and I'm like, this is all still

true. The examples have changed. I don't know how many people will know Welcome Back Cotter. The names have all changed since you hung around, but it's all still basically the same. That was about a high school. The names have all changed. The little details on the errors. But basically, we still have unreliable AI, you know. Since large language models came on the scene, have something that looks general but it's not that intelligent. It's not that reliable. You can't really count on it. It's not nearly as reliable as a calculator is for example. Right? I mean calculator, you type in 3 times 17 and you get 51 and you're good to go. On a large language model, you never know what you're getting. And that was true in 2021 when I gave this before and it was true, you know, today in 2024. And in fact, because the the privilege of writing a talk yesterday as you wake up and you can add another example, like, things are just they're just not trustworthy. They're interesting, but we see the same problems as before. So that was the first part of the talk. It was the larger part of the talk, and it literally went like slide by slide. This was still true. This is not true. And most of it was true. For fun, I used like orange letters where something was new and not that much was new. And then there was the part that I missed in 2021. I had a little hint of it, but really not wasn't clear in my head. It was clear in some other people's heads. But the thing I missed then in my kind of critique of large language models was excuse me. The thing that I missed in my critique in 2021 was what was going to happen to society and to the tech industry. More cynical people than I may have seen it coming. I didn't quite. I grew up in the kind of era of Google. I mean, I wasn't let me say that again. I I started thinking about the tech world a lot in the age of Google. Google had its problems with surveillance capitalism, but I think was genuine in saying don't do evil. They really didn't want to be evil, and I think the companies now really don't care. They've put in all this money and all this chips, and they need to make back the money, and that's just driving so much in so many ways. It's driving the hype. It's driving decisions about copyright law and exploiting people and so forth. And the last part of my talk was really about what I would call the moral decline of Silicon Valley, which is actually the name of a chapter in my in my new book. I think it's been precipitous. I think there's really been a change. Like, I don't see Steve Jobs being happy with what's going on right now. Like, he didn't build Apple to be this kind of company. Apple still, I think, is not so much. But so many of these other companies, it's all about the surveillance capitalism. It's all about making as much money as possible. It's about screwing artists, which I think Jobs would not have, you know, done. Think Jobs really cared about the artist. And it's not that anybody wants to screw the artist, but they're completely indifferent to it at some level. Right? They'll make a licensing deal if they're forced to it, but it's not like they want to. You have these people talking about universal basic income, and yet they don't want to pay artists a nickel if they don't have to. The courts force them to, they'll pay the nickel. But they're really trying to not pay the artist, not pay the writers

and so forth. You have a lot of people that I think are really just in it for the money and don't really care about society and that's having consequence. And so the last part of the talk was like, really, what do we do about that? Like, can we trust these companies to self regulate? No, we can't. The book, Intaming Silicon Valley, is also about the fact that we can't really trust governments to do the right thing either. The governments are lobbied constantly by these companies. There's so much money behind the scenes. And so, like, here in The United States, hardly anything has happened. You probably know I testified in the Senate a year ago, and that was kind of 1 of the highlights of my life. It was amazing. It was this historic moment. Sam Altman was there, the Senate was there. It was the first time the senate had a full hearing on artificial intelligence policy. And I mean, really, like, I remember walking to the capital the night before and seeing, you know, Twilight. It just kind of blew me away. And it was amazing to be there. And it was also amazing because all the senators seemed to understand the urgency of the moment, how important it was that we regulate AI in a right way, not too strong, not too weak, that we get it right now. They all realized that we had screwed up with social media really bad, that they had screwed up with social media. They were incredibly humble. It was amazing watching the senators who, as a lot of people said to me afterwards, were on their best behavior. They really seemed to get it. And I was so excited, and I've been so disillusioned ever since because, you know, that's over a year ago and nothing has passed. The Senate hasn't even voted on any, you know, serious AI regulations. There's a little bit that's coming up soon, but by and large, like, nothing has happened.

AI Safety Theater and Image Generation Issues

Tim Scarfe 00:16:29

On that though, you've you've spoken about the apparent, you called it the Messiah myth of of Silicon Valley. And is is that what's going on? I mean, the OpenAI and Anthropic, they seem to be doing a lot of things around safety. Is that just theater?

Gary Marcus 00:16:45

To a large well, you know, it's hard to get into other people's heads, but I would say that there's more theater than not. That you have companies like OpenAI and Anthropic publicly saying they're for AI regulation, and then they're out there trying to weaken whatever regulation is proposed. You know, they they all tried to block SB 10 47 and and ultimately Anthropic apparently was instrumental in weakening it a good bit at the last minute. You know, OpenAI was you know,

Sam Altman was telling the senate while I was sitting next to him how important AI regulation is. And behind the scenes, Billy Perrigo reported this in Time Magazine, the the lobbyists for open AI were trying to weaken the EU AI act and probably succeeded some. So there's definitely like a 2 sidedness to it where there there's public statement that they're supportive and then in private, you know, they they really aren't. Why is there such a divergence between

Tim Scarfe 00:17:41

the perception of this technology and and the capability? And I I often look at people's Twitter just before an interview and you posted a beautiful example which reminded me of an old example of of yours from a couple of years ago. It's an astronaut riding on a horse. But of course, it wasn't supposed to be that, was it?

Gary Marcus 00:17:58

Yeah. So I actually almost went apoplectic when I tried that 1. So so so I wrote a whole paper, a whole substack essay called Horse Rides Astronaut, And it was really a riff on something that goes back probably to Chomsky, but I I kind of knew it through Pinker. He had this old example of, man bites dog as opposed to dog bites man. So it's not really news. It's a, you know, journal journalists, I think, have used this expression. Right? It's not news if a dog bites a man, but it is news if a man bites a dog. So I kind of riffed on that when DALL E 2 came out. So you might remember when DALL E 2 came out, it was a big deal. It was the first of these really good image generation things. 20 minutes after it came out or maybe it was an hour after it came out, Sam Altman posted AGI is gonna be wild. And a lot of people thought, wow, this is like the AGI moment. And I looked at this stuff, and I realized it's not. It has nothing to do with AGI. It's really nice graphics. The the systems are reconstructing images in a really interesting and powerful way, but they don't really understand language. I did some experiments with Ernie Davis and Scott Aronson and then later with Evelina, Levada and and Elliot Murphy showing that these systems don't really understand the compositionality of language, which is to say that language is made up of parts, you put them together in larger wholes and you're mapping a syntax onto a semantics and you're deriving it from there. Frege is the philosopher that we most associate with that concept. So, I mean, philosophers who have been thinking about this for a really long time and formal semanticists, people like that in linguistics have thought about it a bunch. And it was clear playing with Dali for a few minutes really that he didn't really understand compositionality. In fact, in linguistics, computational linguistics, there's an old idea of a bag of words model. And what a bag of words words model is is you just take the words in a sentence and you scramble them up. You as if they were in a bag and, like, how much can

you explain with that or whatever. And a bag of words model is almost like a control group. It's not a very good control group. You know, if you can't do better than a bag of words, then it says you don't really understand the structure of the sentence, the way that the meaning relates to the parts of the word. And I could tell playing with Dolly, you know, even for a few minutes that it had a lot of that flavor. It wasn't literally that, but it was more like that than a system that really understood the the components of the sense. And so I thought about that in Pinker's old example. So I wrote this substack essay called Horse Rides Astronaut, and it was riffing off the old man bites dog example. So, you know, we have lots of astronauts riding horses, but we don't have many horses riding astronauts. And I showed in this substack essay called Horse Rides Astronaut that DALL E tended to have problems with it. That if you said hard to even say it right. If if you said horse rides astronaut, it would tend to give you the more canonical astronaut rides horse. And then I went through all of the kind of stupid or not stupid, that's not the right way to say it. All the kinds of defensive objections that people who love, this kind of AI would make, and they would say, well, that's because the system has enough common sense to know this is impossible and yada yada. And what I showed is actually if you prompted it the right way, it could actually do this. So it wasn't that it couldn't draw the graphics of a horse on top of an astronaut, and it wasn't because it thought it was because the system thought it was literally impossible, but just it didn't understand the relation between the words in that sentence and what it was supposed to do, which unfortunately is its its 1 job. You know, the hashtag on Twitter, you had 1 job. Your 1 job of your dolly is to understand the meaning of the sentence and draw the picture, and it couldn't do it for these kinds of cases. And then I showed later another example. I wrote a whole another essay about it. I can't remember the title. But I showed an example that, NPR covered where 1 of these systems couldn't get, a black doctor with white children as patients because it wasn't canonical. Now some of these things get fixed up some of the time, people train on more data or whatever. That particular 1 I think got fixed on. But today, was working with Grock on literally in the cab on the way over, like, in the Uber on the way over. I was like, I should try that 1 and see if it's any better. I tried a bunch of other things and generally, the kind of, like, challenges that I give, Grock was not doing so well. We could talk about some of the other. But so I did Horace Rides Astronaut, and I got it wrong. I'm like, there's so much discussion about this 1. This was a popular essay, and people came at me in lots of different ways and so forth. And still, like, you know, this allegedly state of the art system managed, at least on the first try, I didn't try multiple times, managed to get that wrong. And I was just like, we are back in 2022. Everybody has been saying for the last 2 years, you know, especially these influencers saying every day they're like, look at this new amazing thing that came out. We live in this time of exponential, you know, bounty and and whatever. But on the things that count,

Tim Scarfe 00:23:15

on the things that matter, at least from my perspective as a cognitive scientist who spent his career studying intelligence, On the things that matter, we really have not made that much progress. But Gary, I thought these things learned abstract world models, and the reason this is interesting is that there's a there's a dichotomy between aleatoric uncertainty and epistemic uncertainty. You know, there's there's a difference between actually understanding something and being able to reason, being able to do this deductive closure to deduce new knowledge about the world.

Gary Marcus 00:23:47

There's a couple of different things in that

LLMs Lack World Models and Human-like Understanding ●

Gary Marcus 00:23:49

to unpack. So 1 is, is this a question about uncertainty? And it isn't really a question about uncertainty. So you I mean, you can have something that's very clear, like horse rides astronaut. There is no uncertainty. The horse is supposed to be on top of the astronaut. So there are all kinds of interesting things about aleatoric versus epistemic uncertainty. They don't really apply here. You know, this is a perfectly deterministic phrase, it's just getting it wrong. Now there is probabilistic outputs in these systems. So they're the systems themselves are not deterministic. And if I ran that same prompt 10 times, I might get 10 different answers and maybe 6 of them would be right and 4 of them would be wrong, or the other way around, or who knows? So there's that kind of uncertainty. But the fundamental is you are supposed to map your semantics onto this description of the world. You asked about world models. They don't really have world models. I think that that's actually easier to see though in Sora because Sora has changed over time. World models are in part about understanding the dynamics of the world. That's really why you want to have a world model. You know, humans have world models. So I like I have a model of the room that we're in where there are lights. It might not be perfectly specific. I might not know where all the lights are. And I can do updates. There's somebody else in the room. If that somebody else in the room suddenly says fire, then, you know, we're gonna do something different and maybe stop having this interesting conversation. And so I have a model of all the things that are going on or many of the things that are going on.

And all humans do that all the time. If you watch a movie, you make a model of the characters or I'm watching the bear. So I have you you know, I haven't quite finished catching up. And so I have a model of, you know, the lead chef and the person who works with him and, his girlfriend, who's maybe not his girlfriend anymore. And I'm watching all that stuff, and and things will unfold, and I'm maybe trying to put flashbacks and try to order the sequence. So I just saw a wonderful episode about, how somebody first got her job, I don't wanna give away too much, but, working at the restaurant. And, like, for the first 10 or 15 minutes, you're sitting there. How do I relate this thing that I'm watching? Is this in the now of the film or is this before? And eventually, realize it's a flashback and it's a kind of origin story. It's a really beautiful story and and sad and powerful. And so and I'm sitting there trying to make a mental model of how this piece of this narrative fits in with this other piece. And what kind of person like, I've seen this character before, but now I see much more. I'm learning more about her. I'm more learning more about her partner and the relationship. I'm building a model of all this stuff. And current systems just don't really do that. What they do is they build a model of the words and maybe some other stuff that have been said in a sentence so far and try to guess what would happen. But they don't have, like, the equivalent of index cards, if you remember those where you, like, like, write down notes or databases where you, you know, have records. You know, this is your your phone number, and this is your address, and so forth. They just don't have that. People don't understand that. They also don't understand that these systems don't do sanity checking, that they don't look things up in encyclopedia or Wikipedia or whatever. They just don't have models, world world models, cognitive models, or what have you. When when they're trying to do horse rides, astronaut, it's not like they have a world model of what horses usually do and what astronauts do. They have a bunch of pictures, and their their pictures are kind of clustered in space, and they're going to some cluster trying to find the nearest cluster. I'm oversimplifying a little bit. But they're going to this cluster of words that have been around this thing before. It's not the same thing as as understanding. So back to Sora, which I think is actually a better example. So with Sora, you see frequently weird things happen. Like, for example, there was 1 where people are carrying some stuff, and 1 guy goes behind another, and then the camera moves, and the guy is just gone. Right? So if you have a model of the world, then you know, you know, which people are there. There's another 1 where there's like 4 dogs and then the angle changes the camera and then suddenly 3 dogs. And then the camera changes and there's 5 dogs. So, like, people would find that weird. There are circumstances where people would miss it, but fundamentally, can see that Zorro does not really have the notion of object permanence, which is 1 of the basic things that I believe that we're born with based on Liz Belky and Renee Beyer Jean's cognitive development work and so forth. A lot of people know the old Piaget stuff saying that only 8 months old do you

learn object pronouns. That's been shot out of the water by much better experiments using more sensitive measures and so forth. My best guess is that stuff is innate, but Sora never gets it, And it never gets lots of other stuff too. Like, it doesn't learn that a chessboard is 8 by 8. It sees a bunch of chessboards, but it might draw a 7 by 7 chessboard. It doesn't understand that there are conventions. There was a a Sora aunt, Jan Lecun, and I both posted about this him an hour after I did, this ant that had 4 legs. Right? Like, it seemed god knows how many ants, and it still doesn't understand how many legs an ant typically has. So there are lots of things that an ordinary person's model of the world would have, and these systems just don't have it. But it really comes out in the changes over time where just a bunch of impossible things happen. And they only make sense in terms of the statistics of pixels rather than the statistics of the world. So they make sense from the statistics of pixels from this frame that the next frame might, you know, not have anything behind it or whatever. You know, every pairwise bit of pixels 1 frame to the next sort of makes sense if you don't understand what is an image of. Or another example is there's a Sora video where somebody goes into a building. It's a fly through, and it is amazing the first time you watch it. It's like a museum. But the second time you watch it, you realize that, like, the outside and the inside don't actually match. So pixel by pixel, everything in that panning shot makes sense. But if you go across you know, it's only like a minute film. If you go across the minute, there's enormous an enormous number of really massive discontinuities that make no sense whatsoever except on the, you know, frame by frame. From this frame, you could see a frame like that. But if you look at frame 1 and how you got to frame a 100, it don't makes any sense at all. And that's because there's no stable world model. There's not a stable model of like what the dimensions are of the building, and so you wind up I forget what it is. I think it's that like the exterior shot, you know, is something small, and the museum is massively bigger. There's, like, an empty courtyard, but then it's not empty anymore. So there's all these inconsistencies.

Tim Scarfe 00:30:43

Yeah. I mean, this is where I was going with the epistemic risk because we can verify. So so we we have a world model. We have facts about the world, and we can verify. And sometimes that's a binary. We can we can say in certain situations that an astronaut is on a horse or a horse is on an astronaut. There is some vagueness around the boundaries perhaps, but it but it's binary. But I'm interested in why we anthropomorphize

Gary Marcus 00:31:04

these models. They're kind of adversarial attack on our perception in many ways.

Superficial Intelligence vs True Reasoning

Gary Marcus 00:31:09

Take the classic Turing test. What it really turned out to be was a measure of human gullibility. So passing the Turing test doesn't actually mean you're intelligent. It means you can fool humans. It turns out we're very easily fooled. The most dangerous version of this is you can see a few minutes of a driverless car and conclude that it drives basically like a person, that everything is good. And in fact, that driverless car may have a lot of serious problems in a lot of different contexts. And so something that superficially looks like a person for a few minutes may not actually be like a person. And so what has happened with large language models is they superficially produce human like output. And people are willing some people, not all, are willing to cut them some slack when they make an error once in a while, but they actually attribute intelligence to these systems in a form they don't have it. And 1 of the ways that they do that is they attribute intelligence they think it's like me. It would do the things I would do in in such and such context, and it doesn't. So, you know, 1 example I used in my TED talk was the the lost Galactica, the the meta system that was pulled that Jan McCune is still bitter about. Galactica said that Elon Musk died in a car crash. The sentence was in March 2018, Elon Musk was involved in a fatal automobile collision, think it was. And then it continues on, and it's clear that it thinks that Musk in fact died. Any person would say, oh, wait a minute. In 2018? I mean, if there is I mean, we we would think about data. We're not perfect with thinking about data, but in evidence. But if there's evidence that any human being is alive right now, it is Elon Musk because he is on X every day. He's in the news every day. The the amount of evidence that Elon Musk is alive is greater than for any other person on the planet. And so this is an obviously false assumption. You could also go to Wikipedia and so forth. So if I was an editor of a, you know, newspaper and somebody gives me, Elon Musk died in a car crash in 2018, I'd probably fire them. It'd be like, 2018, that just does not make sense. It does not check out. And even if you said now, I'd be like, well, can we get some extra sources? And, you know, if you said that he died today, you know, I'd want some sources. Can we get confirmation on that? We don't wanna run it yet. And so you or me or certainly any, you know, editor of a newspaper or something like that is gonna fact check things, especially ones that seem, you know, prima facie to be implausible. But LLMs don't do that, and it's very hard for the average person to realize that because they see this small sample of data, and in the environment in which our brains evolve, the kind of evolutionary history, we didn't have this problem of, you know, impostor chatbots pretending to be people. We had other problems like, that lion, is it gonna eat me? And so we're pretty good at looking at motion.

Is that thing coming close to me or not? How big is it? You know, we make a lot of judgments about the world that are really good, but we don't make judgments about AI that are really good unless we took cognitive science classes in college or something like that, which most people didn't. And so we see this tiny sample and we wildly overgeneralize because it said a few sentences and then, like, types the word out 1 by 1, which was a stroke of genius by OpenAI. Mhmm. It gives this illusion of being person like. It's not actually person like at all. It's a statistical, you know, autocomplete on steroids that is trained on a lot of data to look like a person, but it is not reasoning like a person ever.

AI in Specialized Domains: Chess, Coding, and Limitations

Tim Scarfe 00:34:45

I mean, there's there's 2 other things here as well. First of all, there's the phenomenon that people want to believe that it's intelligent, and especially when publishing NeurIPS papers. That's what Sabaro said to me the other day that you don't want you want people to think the LLM did it because then you get a NeurIPS paper. But also, I read a blog post by Nicholas Carlini from Google and he said he used to be in the camp that thought that LLMs are just databases. And he started playing chess. Not databases either, but we'll back We'll to that. We'll get to that. He started playing chess with GPT-four, and he was he was blown away with the sophistication of of the moves, it was often making correct moves. But he said something quite interesting, which is that if he played chess like a bad chess player, it would reflect the bad behavior back. It's the same thing with generative code. If you write code with sophistication,

Gary Marcus 00:35:31

it gives you better code back because it's almost adopting a role player. Well, it is a mimic. Just a sidebar on chess, g p d 4 doesn't really play chess that well. So, it makes a lot of illegal moves. There there's a very good I'll I'll try to give you a link to put in the show. The very good I'm blanking on the guy's name. Has done a very good analysis of of GPT, for chess playing. And it plays like, I don't know, like a 1,600 game, which is like, you know, better than the average person in your high school, but not anything like world class. And it makes a lot of illegal moves, like 6% of the time or something, you know, some crazy number like that, which note, you know, the chess compare with a chess computer that I bought in 1979 where you could stick the little pieces in the hole. I think it was called Sargon, was was a software underlying. That never

made an illegal move. Never ever. Right? We're talking about chess software from I mean, really even further back, 1969 never made illegal moves. Like, d p t 4 is not, you know, state of the art in chess, and we we should understand that. But people don't. They're they're amazed that it can do it at all, and there's some reason why you should be amazed that can do it at all. But you have to realize that it is not, you know, searching a tree the way that a proper chess program can do. It is doing mimicry, to come back to the other part of your question. And you do get these, like, weird mimicry effects because essentially what you're doing is a little bit like what humans do when they're priming. You're directing the system to a particular part of its corpus. So you can direct it towards the more sophisticated language or the less sophisticated language or whatever, and it's going to try to replicate language like that. And so that is why you get some of those, kinds of effects. I don't know whether, what Carlini described, you know, really bears out in a systematic study, but if it did, that'd be my guess for, you know, why it would. It's like the database of lousy chess games is gonna look different from the database of good chess games.

Tim Scarfe 00:37:25

Yeah. It's strange, isn't it, though? As we memorize more of the long tail, the the reliability sort of goes down a little bit. So around 5%, 4%, and a lot of people argue that that that's just fine. We can engineer our way out of it. Do you think that? It depends on what the problem is, first thing I would say. So

Gary Marcus 00:37:44

large language models are not like calculators. Calculators give you a 100% correct answer and large language models in very few domains give you a 100% correct. In some domains they're just completely outmatched so you know floating point arithmetic they're gonna be a lot less than 95% correct, especially with large digits. They're gonna be much worse, would suppose. I don't know if anybody's done exactly that study. And something that matters probably in all forms of AI, but particularly in large language models, is the cost of error. So, large language models are best suited, I would say, to things that are like brainstorming or autocomplete. So coding is a kind of autocomplete where the coder is still at the wheel. Right? It's not a fully autonomous activity. The system is not actually writing the whole code or whatever. So it's writing little bits you drop in and coders have spent their entire lives learning how to debug bad code, partly because most coders don't type that well and also because coders forget things and whatever. I I've done a bunch of coding in my life and I know how it goes. And so, like, if you don't know how to debug things, you're just not gonna become a coder. Like, it's just not the profession for you. So everybody using those tools can tolerate a certain percentage of error, and they're trading off. How long does

type take me to type this out to look it up versus how long to debug it? I think people were initially excited. Some people are less excited now because they realize sometimes, like, they did a bunch of tests to make sure that the code works and then they'll pass all of them. They wrote the code in an hour, and then 3 days later, they don't remember why this code is there because they didn't actually write it themselves, and it takes them, like, 24 hours to debug it. And then they're like, I don't know if this trade off was worth it or not. So there's there's still some open questions there about security and so forth. There's been, some academic literature suggesting their problems, but at least in principle, you have a coder who is picking, you know, the outputs. They're taking some, not others. They're they're they're fixing it. And so you can have a high amount of error, but it might still be worth it. There are other domains where also a high amount of error might be worth it, like brainstorming. So I'm trying to think of a commercial and it gives me 30 ideas, I reject 29, but I'm happy that I got the 1 and so, you know, that might be great. Right? So it could be 90% error rate, but you're still happy. And then there are domains where like any error is probably gonna like kill somebody. It's like a medical domain and, you know, you have the system treating a child like an adult and giving the wrong answers because it's not really trained in enough pediatric data and like any error might like actually like kill a child or send them the hospital or whatever and so you need to be much more accurate. So there is no blanket statement you can say about, like, what percent error matters. If you want to use a thing as a calculator, probably you shouldn't get anything wrong, like, you should just use a calculator. So it does depend what you're applying it to. Yeah. Coding is an interesting example because there's a verification step. So does the code compile? And then there's the behavior. Does does it does it run the way? Yeah. I'm gonna pause you right there, though. The worst stuff is gonna come from people who think that's the only verification step. I'm not saying you think that, but I think some people do. So for those who are not programmers, in a language like c for example, c plus plus, you write the code, you compile it, which turns it into machine language, but that does not prove that there are no bugs. But a a beginning programmer actually might labor under that assumption, so they think if it compiles, I'm good to go. Now a sophisticated program realizes that's not the case at all, that bugs can emerge in code that does correctly compile, but there's still some assumption that's made that's wrong and so forth. But, you know, there's a certain amount of bad code that is seeping into the code base because people do think that's the only verification step. Another verification step that good coders know about would be unit tests or some some kind of testing. Like, I now that it compiles, that the journey has just begun, I'm gonna make sure that this code actually does what I want it to do. And a good programmer understands the logic of what they're programming, and they know what a good test might be. They know what wacky user input might come. They wanna make sure that it handles that input. They

understand, you know, when the circumstances of their assumptions might be tested. They test that. And a bad programmer doesn't really do that. They do a couple tests and they call it a day. They say this is good. They will go on to the next thing, and then it all falls apart 3 weeks later when 1 of those assumptions is violated.

AI-Generated Code: Capabilities and Human Interaction

Gary Marcus 00:42:10

It's very true. I mean, with Gen AI coding, I've been doing a lot with Claude 3.5 SONNET, And the first observation is you can now generate something like 2,000 lines of code in half an hour.

Tim Scarfe 00:42:20

And sometimes it works reasonably well. Now, I noticed that it it hits a complexity ceiling around 2,000 lines of code. At that point, it just starts hallucinating and and doing crazy things. But on the anthropomorphization point, this is interesting because it's a supervised process. The human is coming up with a prompt and they have a mental model. They are selecting the completions. They're interactively running it and and so on. And again, isn't it interesting that the human doesn't like to think that the creativity and the reasoning is coming from them when they're working with language models?

Gary Marcus 00:42:50

I mean, don't know what people like to think of. I mean, I I, in a way, I can't really answer that question. That's a question about, like, how do individual users feel about the product that they're making with the

Tim Scarfe 00:43:04

system, if I understand the question correctly. Well, it's it's a sense that I I'd if and I know you don't think language models are a database, but let's say they're a database. The the user creatively comes up with a prompt, and a lot of the reasoning is actually implicit in the prompt, and then the language model will, know, retrieve a whole bunch of ideas, and then the user will select 1 of those ideas, and then they'll run it on their compiler, and then they'll verify, you know, it's a this is a this is actually a very integrated supervised process, yet people think the magic LLM is doing everything.

Gary Marcus 00:43:34

Yeah. There I mean, some people may may think that. They, like let me see how to put this. There's a wizard of Oz problem sometimes where, you know, the the man behind the curtain is doing a lot of the work, and you'll see these threads on Twitter where people will say, the machine didn't get it wrong because, look, I can do this prompt, and it'll be this very complicated prompt, and then the machine gets it right. Well, who's actually doing the work there is the person who is coming up with the prompt, and it's even more complicated than that because they're noticing that the machine got something wrong. That's an important piece of intelligence that the machine has actually failed at. And then they're noticing that the, you know, simple prompt doesn't get it right. They're being creative about what prompt will get it right. They're doing all this work, then they say, look, the machine is great. In fact, the machine, you know, only gets it if you sort of, to use an inapt metaphor here, hold its hand, you know, 3 quarters of the way there. Yeah. It's it's so true. And another thing is that when you interact with an LLM, you you get a feeling for its personality.

Tim Scarfe 00:44:37

So with with Claude Swanet, for example, I know when it starts hallucinating. I can tell within 2 sentences of of of the response that it hasn't understood me and the code is because it will confidently generate garbage broken code. Mhmm. And and what do you do in that situation? So 1 thing you can do is create a new session. So it no longer has the context of all of the decisions that you made previously. And then it's more grounded in some sense, but less grounded in in another sense. And then I've just generated all of this code, and now my friends do a peer review. And and they they start trying to understand all of this code. And there's a huge understanding credit card because they don't understand anything I've just created. 1 interesting idea would be to have a kind of transaction log of all of the previous prompting to the LLM, which means there's a huge session so the LLM understands all of the assumptions and all of the requirements that went into it. But even that just disintegrates after it reaches a certain length. I mean, ultimately,

Gary Marcus 00:45:29

there's I think a limit to what you can get out of these systems in terms of coding help. It comes from the fact that they don't really have a good theory of mind for what the customer, which might be the programmer, actually wants from the system. The the channel for communicating that is just not great. And so these systems are best at very narrow requests. Like, I wanna know how to express this in HTML and, you know, what is the API for this thing? Where you're kind of like looking something up. As opposed to, like, the first time somebody built a word processor, they had to think, well, how am I gonna structure this system at all? How am I going to represent the contact I mean, the the document that somebody's working on? How am I going to set up, an interface where and I don't know how they did this in the original 1, but you really want this so called model view controller separation so you can display things separately from the logic of what you're doing when you get the keyboard actions and so forth. And so there there's a an intellectual process of, like, how should I structure this abstract problem? Now, real world coding is a mix of that. I mean, you you get your low level people to look up, you know, the silly stuff and and the or the not silly stuff, but the the the kind of cut and dry and stuff. And you want your, like, top level system engineers to, like, think about how am I gonna build this system in the first place. And large language models are not particularly good at that as I understand. I remember in the early heady days of large language models, which is to say in 2023 right after ChatGPT got popular that Sam Altman was talking about I think it was Sam Altman, was talking about, you know, who's gonna have the first billion dollar business run by a single employee and a bunch of LLMs. And we haven't, as far as I know, actually seen anything like that. And, you know, 1 of the reasons we haven't seen anything like that is, like actually figuring out how to build a system, for example, is way outside of the scope of of what these things can do. They can be helpful assistants and, you know, Eric Brunjolfsson among others likes to talk about, you know, human augmented kinds of stuff, and I'm all for that, trying to use AI to augment human abilities. But if you're talking about really autonomously, you know, running the books for a business, running the the marketing campaign, like, really doing that by itself with no human supervision, these systems just aren't reliable enough for that. And so nobody can actually make a 1 person billion dollar business that I know of.

Tim Scarfe 00:48:07

Yeah. Mean, everyone's talking about agents now,

AI Regulation: Industry Resistance and Oversight

Tim Scarfe 00:48:10

and you just nailed it. And I think Rodney Brooks said something similar, which is that we underestimate the amount that we, in the process of supervision,

Gary Marcus 00:48:18

smooth off the long tail of failure modes. That's right. And that doesn't work in certain domains. So it's in the brainstorming domain, but not in the driverless car domain. I mean, the driverless cars, like, we still don't know how much Waymo has humans behind the scenes sorting stuff out. With Cruise, when that number came out, it was mind boggling. They had, like, I think they had 1,500 people behind the scenes in tele op centers and 800 vehicles on the road or something like that. They I I don't remember the exact number, but it was more tele operators than they had cars on the road. So there was a vast amount of smoothing out that was hidden from the customers. It was hidden from anyone wanting to do scientific analysis of what was going on. There are a lot of domains like that where where, you know, the man behind the curtain, the Wizard of Oz thing, is is really doing a lot of the work.

Tim Scarfe 00:49:08

So let's move over a little bit to to policy and some of the things you spoke about in in your book. Maybe let's start with your blog post about Fei Fei Li and and the the California regulation.

Gary Marcus 00:49:18

So I mean that situation is still unfolding but in general what I found around s p 10 47 was there was a lot of misinformation and a lock lack of precision. So I think a lot of people painted this kind of catastrophic view of what would happen if s p 10 47 passed unchanged. Since then, it's already been watered down some and so the thing that people warned about is not actually going to become law in any case. And, you know, I some of this just happened yesterday. I haven't really even fully digested it all yet, and I don't know how much, more modification there may or may not be. So I don't want to be too specific about the details of it, but what I would say is that at a general level, there's this crazy notion, crazy in my mind, notion that we can't have any regulation around AI or we will kill it. And every other domain, that's just not true. So in fact, in many domains, having regulation has been essential to the growth industry. Airlines is an example of it. You know, back in the 19 forties commercial airlines were like really dicey

kinds of things or '19 or the 19 fifties. And now commercial airlines are incredibly safe. They're much safer than driving your own car for example, Way safer than motorcycles. And why are they safer? Because we have multiple layers of oversight. Right? And, you know, it has not ended the airplane industry that we have these multiple layers of oversight. We we have we have rules about how you make an airplane. We have rules about how you maintain an airplane. We have rules about how you investigate an accident and so forth and so on. That has not caused the airline business to go out of business. Now, in every industry, there are, you know, many industries, probably, are dumb regulations, especially the the first time around. No nobody's making a claim here that, like, anybody's first time at bat, to use a baseball metaphor, that they're gonna get, you know, perfect regulation. I'm a fan of the EU AI act, but I'm not so naive as to think that it's perfect. I'm sure we're gonna find problems with it. But the notion that, like, you can't have innovation in an industry where there's regulation is just absurd on its face. Like, I I do you know about that Overton window? Like, people are just saying crazy things trying to reframe the discussion, but of course you can have some regulation around AI. And part of what's so crazy is like, they're like, oh my god. This is so onerous, and you look at the the fine print, and if you're not running a $100,000,000 training run, you hardly have to do anything under this law that everybody got so up in arms. Like, if you can pay for a $100,000,000 training run, you probably can pay for 1000000 dollars in compliance. Like, you can do that. Like, we're talking about these businesses that have been capitalized at valuations of, you know, $4.10, $80,000,000,000, we're supposed to mourn that they have to fill out some paperwork. Well, okay. But let's look at the other side of it. These things have already done harm. Right? Think about deepfake porn. Think about all the misinformation that's floating around in these elections and so forth. It's not crazy to say that people should take reasonable steps to make sure that there'll be no catastrophic harm. That's another thing about the bill. It's like it's basically restricted to catastrophic harm. It's like, literally this is not a metaphor. Literally, your first $500,000,000 of damage is on the house. You know, we're only talking about if you cause at least $500,000,000 of damage and you're negligent, and, like, you should have known better and you did nothing about it. Like, is that so unreasonable to see you don't get a free pass if there was 1000000000 dollars worth of damage, which, you know, could also be equate to a certain number of lives or whatever people have math around it. Like, you're gonna cause 1000000000 dollars of damage and we can't, like, do anything around that? We we wanna give you a free pass on that? That that's crazy. And we don't do that in other industries. Like, if you make a circular saw, we want you to, you know, put a gadget on there so that the average idiot doesn't chop off their fingers. It's, like, not unreasonable. And we don't say, oh my god, the circular saw industry is gonna come to an end. Nobody's ever gonna have power tools. Home Depot is gonna close. But that's kind of like what it was. Like,

Home Depot of AI is gonna close and we'll never invent another tool and like It's just ridiculous.

Tim Scarfe 00:53:45

Yeah. I mean, in your book, you gave the example of cigarettes, but there's also some things Cigarette manufacturers.

Gary Marcus 00:53:50

Right? Right? So just make clearly, the cigarette manufacturers pretended for years that there was, you know, no conceivable harm. They said the science here isn't good enough. Let's wait until I mean, basically, cigarette manufacturers wanted to say, unless you do a causal study in which you assign people, human beings, to smoking cigarettes or not, this is not science, and we're not responsible. And they got away with that for a really long time. If you're an actual scientist, understand that you can use animal models and you can do certain kinds of observational things and you can make a pretty darn good guess that cigarettes cause lung lung cancer without actually doing the experiment that they were insisting on. But they just, you know, with a straight face, would say this stuff for years, and a lot of people died.

Tim Scarfe 00:54:34

When we get into tech, I mean, you you spoke about social media, we've got examples like Uber, for example. Tech is interesting because there's a way of flouting the rules, so the rules don't apply to you. And AI is interesting because I mean, OpenAI, of course, argued to the House of Lords, as you said in your book, that, you know, we need to flout copyright because AI is

Copyright Issues in AI and Tech Business Models

Tim Scarfe 00:54:55

so magical. It's going to be used for all of these Let's just run that part in detail. What they said in the House of Lords

Gary Marcus 00:55:02

I mean, it's interesting what they said and what they left out. So part of what they said is probably true, which is we can't make our stuff work unless we use copyrighted material. But then what they implied is they need all the data to make it. I mean, it still doesn't work that well, but it worked kinda works. It works as well

as it does, because they're training on all this copyrighted material. If you took away copyrighted material, it wouldn't perform as well. Like, we can agree on that part, and they said that the House of Lords more or less straightforwardly. But then what they implied is we need an exemption. You need to give us all this for free. There's a totally well known alternative to that that was much more equitable to the artists or the writers or what have you to the creators, which is you license the stuff. Like, Netflix does not just show, you know, your movies for free without consulting you and not paying you. They license it if they're gonna show your movies. Same thing, you know, Apple iTunes music. And in fact, a lot of this reminds me of the Napster days where there were a bunch of people running around saying Napster is great because I can download every songs or every song for free. And there were a bunch of artists saying, well, hold on. Then we have copyright laws. And there were a bunch of people saying at the same time, information wants to be free. It's a new era. And the court said, no. We have copyright laws for a reason. They're not going anywhere. And so net, sorry. Napster was driven out of business, and who took their place? Apple. Because Apple actually was willing to pay licensing fees. And so Apple made an enormous amount of money licensing music, but the artist still got a cut, maybe not as much as I would like, but they got a real cut. So we moved from, you know, complete utter theft, that's what Napster was, to a licensing arrangement that was at least reasonably equitable, and it's fine. And we can do the same thing here. You can lie and in fact, it's interesting because OpenAI tells the House of Lords, you know, basically, you have to give us the stuff for free. And, you know, behind the stage, they're making licensing deals left and right. They're talking to everybody. So they know, in fact, that it might go that way. So they were trying to get this massive handout. It'd be like, you know, giving me the West Coast Of The United States because I said I need it for my work, and you you forgetting to say, well, could you maybe pay something for that? Social media is quite an interesting example of, you know, in in in the tech space, Mark Zuckerberg had this phrase, move fast and break things. What could possibly go wrong?

Societal Impact: Risks, Misinformation, and Ethics

Gary Marcus 00:57:26

I mean, that's how I start my book, in fact, is is with that phrase, and and a lot could go wrong with AI. Right? So I mean, we already saw a lot went wrong with social media. Like, you know, it's been really terrible for teenage girls. I think lots of people are addicted. There's a wonderful piece called Twitter poisoning or that's in the title by Jaron Lanier where he uses the example of Elon Musk. And he says, if I had gone to Elon I'm doing this from memory, I might not get

exactly, but he said, you know, I'd known Elon Musk for a long time. And if I had gone to him in 2016 and said that you're gonna spend, you know, multiple hours a day on Twitter and it's actually gonna hurt your car business, maybe your rocket business a little bit. It's gonna make people really hate you and yet you're gonna be compelled to do it a few hours a day. Musk would have looked at him and said, you're insane. Of course, I wouldn't do that. You know, I have work to do. But he got addicted. Right? Twitter is addictive and I speak as someone who is addicted, to Twitter and and would like to stop. You know, it takes up a lot of people's time and emotion. There are there's some value in it too where I wouldn't keep doing it. But, so social media is addictive. It it, you know, has probably increased loneliness. It's certainly increased political polarization. AI might do all that worse, so it might make people lonelier. So, you know, 1 of the consequences I think of these chatbots is a lot of people are gonna fall in love with those chatbots. It's already starting to happen, and they're gonna lose or never develop depending on their age whatever social skills they have because they spend most of the time with with the chatbot. There was an old study showing that people who watch more television are less happy than people who watch less television. It's partly about opportunity cost. If you watch television 5 hours a day, you actually enjoy it in the 5 hours. You pick shows you like and whatever, but at the end of the day, you haven't you haven't learned a skill, haven't gotten a better job, you haven't made a new friend. Social media is like that. Well, AI is gonna be like that even more so. On the polarization front, you know, polarization in large part comes from misinformation and targeted misinformation. Well, AI is going to accelerate all of that. We're gonna if it's even possible to believe, we're probably gonna have an even more polarized society than we already have as a function of deep fakes and other kinds of misinformation or disinformation by deliberate design.

Tim Scarfe 00:59:45

There are lots of ways in which moving fast with AI is likely to break a lot of things. Yeah. I mean, you said automated disinformation may destroy what's left of of democracy. I mean, is that is it sounded a little bit hyperbolic. I mean, how how do you see that rolling up? It's not, though, I think. I I mean, I I genuinely worry about that.

Gary Marcus 01:00:02

So we are going to have elections, which is the how how do I say this? Is the ultimate act of democracy is an election, especially if you have a representative democracy. You you don't get to vote on individual things. You have referenda and stuff like that. But basically, we live in a a representative democracy, and misinformation may undermine that. So, you know, we may, this year, in October, see deep fakes about what are the other candidates that actually materially changes the election. That really can undermine democracy, and we're gonna see that more and more around the world. And parallel to that, part of Russia's game plan has always been what they call the the fire hose of, what is it, fire hose of misinformation or something like that, fire hose of propaganda model. And the idea behind the fire hose model is not to push 1 lie, but to push so many lies that nobody knows what to believe anymore,

Tim Scarfe 01:01:02

and that is what's happening. You know? What was that Adam and Curtis film talking about that? Was it hyper normalization?

Gary Marcus 01:01:08

I don't I don't know the particular, but I mean, this is a well known thing. I think it's called the Russian fire hose, a propaganda model. It's been something that is, you know, Putin has cared about for a long time, and we are headed to what was always Putin's endgame, which is nobody will believe anything. So, you know, just in terms of images, images are basically done. You know, for for 100 years, if you had a photograph of something, that was serious evidence that it was true and people knew that. And as of, you know, the last couple of months, people know that that's not anymore. Well, like, people don't know what to trust anymore, and and authoritarians like that. They just go in and do what they want and and say, you know, you can't believe that other thing. Don't believe it. And, I mean, like, look at the bald faced lies that Putin tells all the time about the war in Ukraine. Right? If we move to an environment where basically nothing can be trusted, as I think we are, that's really bad. And 1 1 reaction that people have had to all of this is, well, of course, we need to educate people, about misinformation so they'll be more skeptical, and that's true. Like, I but in the limit, it's not so good. Right? In the limit, if nobody trusts anything, democracy can't really function. Right? I mean, the the currency of democracy is information and informed decisions.

Tim Scarfe 01:02:35

And if you don't have information and can't make informed decisions, then you don't really even know what you're voting on, and then democracy dies. It's ironic, isn't it, that when there's no knowledge grounding to reality, that we become like chat GBT because we can't verify anything anymore. But it could go 1 of 2 ways. Dennett argued that it would tend towards acquiescence. Or the other argument is that it would tend towards skepticism and and rational thinking. Where do you think it'll go?

Gary Marcus 01:03:03

Might be some of both. I mean, I guess this remains to be seen, but I think a lot of people will feel defeatist and and just give up. They kind of won't believe anything, but they'll kind of give up trying. I I I mean, I hope I'm wrong on on that. I I've been right about, I think, nearly all of my AI predictions. I hope I'm wrong about this political prediction that people will just sort of enter despair and not really care about truth anymore. But I already see signs of that on Twitter every day.

Tim Scarfe 01:03:34

I mean, people talk disparagingly about you. And the thing is, I mean, actually, DeepMind had a motto, solve intelligence and use that to solve everything else. And you admitted in your book that AI does have the potential to revolutionize healthcare and bring abundance to our lives. So you know, what what is what is the game plan here? What what what are you trying to do?

Gary Marcus 01:03:55

We need to move past gen AI for 1 and for 2 we need better regulation and there there's some overlap between the 2. So on the technical side generative AI is just not a good way to do AI. It's fine for certain things like brainstorming and it's fine for making images if if if it's properly licensed, but it is not fine for running power grids. Medical remains to be seen. It's not a reliable framework. It's been very hard to get people to understand that. I'm amazed that you still see things like a famous physicist the other day said, I had a conversation with ChatGPT, and I couldn't believe how bad it was. I'm like, where have you been? We've been trying to explain to you that he doesn't actually understand what he's talking about. So that message has not been communicated very clearly, so people overrely on it. There's gonna be weird economic effects, I think, when the bubble bursts. I don't think Gen I has been good and it's not the on balance, don't think. You know, I think I mean, it's an interesting question. That has it been of net benefit or not? And I am not so optimistic because I think mostly Gen AI is a good tool for bad actors and some limited positive uses. And I don't think those positive uses maybe outweigh the negative uses, particularly if we get to this regime of disinformation where democracy is basically undermined. Like, I don't I don't see any amount of productivity gains as justifying the end of democracy if we get there. So, Gen AI unfortunately is a very good tool because it doesn't track truth and doesn't understand factuality, but it mimics very well, it becomes a very good tool for bad actors, but a less good tool for good actors. And so I'm very concerned about that, and I think we need other approaches to AI. So, you know, there's a lot we can still do in nVent in medicine and technology. For example, we still don't really have a good way of handling Alzheimer's. And I think AI could be extremely helpful with that. Not generative AI, but some future form of AI that can read the medical literature, do modeling of proteins, integrate it all. You know, there's an interesting new paper about AI scientists. If we had some future form of AI, we might actually build an AI scientist. Right now, it's just not gonna work that well because they can't really reason about data. They can't plan. You had Subaru on your show. I'm I'm sure he talked to your ear off about how they can't plan. You can't really be a scientist if you can't plan. But eventually, we'll get to other forms of AI, and maybe that'll be great. I can't promise that it'll be great, but I could see ways in which it could be great. Then there's the governance side of the equation. You know, we're basically letting the companies do whatever they want without regard to consequence, and that's just not good.

Tim Scarfe 01:06:43

Yeah. The the AI automated scientist paper is very good. They're coming to London. I'm gonna interview them next month, so I'm very excited about that. But, you know, you said that AI is being rushed out the door, but if not now, when? When? Should we have AI? Yeah.

Gary Marcus 01:06:57

I mean, I think a lot of the stuff that we have now should actually be in the lab and not, you know, commercially deployed at large scale. So I have no problem with people researching Gen AI trying to find I mean, all my issues around deployment. So for example, we know that Gen AI is being used to make health care decisions and job decisions and and so forth. Let's focus on the job decisions. We we know that employers are sticking their job applications in and saying, who should I hire? And we know that there is likely to be discrimination there, but we don't have public access to to, like how bad the problem even is, how how much bias there is. But they're almost certainly bias against minority groups, for example, in using LLMs to make job decisions. And this is just not good. And like, I'm sure it's happening like literally every day, you know, at large scale now.

Tim Scarfe 01:07:51

So I spoke to Yoshua about this last night, he said freedom of speech is really important. We need to protect innovation. I don't want someone to say to me that I'm not allowed to use AI in this way. I mean, what would you say to that?

Gary Marcus 01:08:03

Oh, I would say that society has to balance values. And the United States Constitution and the laws around the laws of the land have made certain decisions that I would say are good, like generally there's freedom of speech, but there's carve outs for certain kinds of things. And that we also, for example, protect against discrimination, so we have an Equal Employment Act. And you can't just like pick and choose and say, I'm gonna keep the free speech and to hell with the Equal Employment Act. Like, I don't see the justification for that. And so you have to say, how am I going to balance this? So if I give you, you know, carte blanche to use these tools that I know are just, you know, are almost certainly discriminating against people, and really I should have access to the data, somebody should have access to the data, that's just not cool. You you can't I mean, I don't know how else to say it, but, adults realize that different kinds of constraints can be in conflict and that you have to find some way of handling the conflict. And right now, the the idea of complete free speech to build whatever you want and free whatever word is I'm looking for, free carte blanche to build whatever you want is in tension with a bunch of other things, like laws that prevent discrimination in jobs, employment, etcetera, or, you know, any rational thing you might have about sort of the commercial production of misinformation, which is actually not very well protected in The United States, but it's protected in some other places. You have to balance these kinds of things. And I think to say I'm only going to look at what I as a programmer get to do and not worry about the consequences of society, either the legal ones or the moral ones, that's just wrong. I mean, it's morally wrong.

Tim Scarfe 01:09:50

For things like discrimination, do you think it would be covered by existing laws? I mean, example, in banking, fraud models and credit models have very specific regulations around them already. Some things are, some aren't. So for example,

Gary Marcus 01:10:03

it's just not clear how much the Equal Opportunity Commission can audit what's going on. So they work on a case by case basis. Somebody walks in and says, I think I've been discriminated against. Here's my evidence, and they take on the case. It's not clear they can do the auditing, that's required. So that'd be 1 example where I don't think the existing laws cover it. Here's another 1 where it's still very ambiguous, is defamation. So large language models have, for example, accused people of sexual harassment that they did not actually commit. We know that they didn't commit it, but it's not clear that the people who were defamed by this have any recourse whatsoever. Because the existing laws of defamation, for example, there's a bunch of complications here, hasn't really been, cashed out in court how it's all gonna go, but but, you know, typically there's some notion of intending malice. Well, you could argue that large language models don't have emotions. They don't have malice. They're not doing it on purpose. It's actually a kind of negligence that's making them do this. But are we gonna say that you can make an a system that's arbitrarily negligent to the point of really screwing up people's lives, perhaps, just because they work differently from people? Like, that doesn't seem to make sense to me. I would say the existing laws don't clearly handle what happens if some AI system completely negligently defames person. And it's completely negligent because there is no fact checking on top of these systems. So, you know, my favorite example of hallucination is not particularly defamatory. 1 of these systems said I had a pet chicken named Henrietta. I don't think that really causes me any harm in the world, but there's also nothing I can do to stop it. And there are other cases like, you know, certain law school professor was accused of sexual harassment and there's nothing

Tim Scarfe 01:11:54

probably nothing he can do about it. I think you're absolutely right about automated decision making. On text generated by LLMs, at the moment, I can still smell it a 100 miles upwind in a in a hurricane.

Gary Marcus 01:12:05

Well you think you can but you know all the studies that I know show people are not that good at it. That's right. Just look at LinkedIn. Well there's tons of it out there. You know there are certain like giveaways, like the word delve is way overrepresented in LLM speech probably because the Kenyans who did the r l h f actually used the word more than than we do over here, let's say, in The States. But, you know, people have done experimental work and I don't think there's any system that can detect LLMs more than, like, 70% correct. So, like, if you're talking about, for example, could we use this these pieces of software that look for LLMs to, I don't know, weed out student papers, let's say. 75 70% correct or 80% correct or whatever is just not good enough. I mean, imagine you're the student who wrote the good paper that is described as an LLM. You're gonna be really really upset if you're in that 1 in 5 who who got screwed and justly so. And so you can't or you shouldn't be using software that's that bad at a task like that. It really has to be better. Nobody knows how to make a 99% accurate LLM detector.

Tim Scarfe 01:13:17

Yes. That's very true. When when I spoke with Luciano Floridi, he said that, you know, the problem with this information technology is that there's no friction whatsoever between it and the the regulatory, you know, the the legal the legal landscape. And open sourcing models is is quite an interesting thing. So do you think the horse has bolted?

Gary Marcus 01:13:34

I think there's multiple horses. 1 of them has and some haven't. So so the horse of LLMs that can be used for misinformation totally left the barn. You know. Meta's latest model is probably good enough to make whatever misinformation anybody wants sound plausible and, you know, now the weights are out there and there's no way to put that back. I mean, short of like a nuclear war that like sets the entire civilization back like, you know, of really crazy kinds of things happening, that's never going away. So that horse has definitely left the barn. There's probably going to be another 20 important advances in AI. I'm making up the number 5, 10, 20, a 100, and we don't necessarily have to handle them all the same way. So the open source question is extremely complicated. Right? I mean there's a lot of reason to wanna accelerate progress towards AI and open sourcing things clearly accelerates them. But there's also a lot of worries about bad actors and what they're going to do and, you know, I find it to be pretty difficult. I will refer your readers to a paper, that I wrote in Substack that that was about an open AI paper that purported to show that LLMs don't make, bioweapon manufacture

worse. And if you actually look at their data, they did their statistics wrong. And I see that paper cited sometimes and it worries me. So I it's not all clear yet, but it is possible that LLMs will, for example, be used as tutors to teach people to make bioweapons that couldn't. And, you know, even a single incident like that could kill, you know, tens of thousands of people in a subway or something like that. So there there there's some definite potential downside risks. What upsets me is that it's a really complicated decision that should have been made by a lot of very bright, thoughtful people trying to work through their differences and instead basically meta made that decision by themselves. And now if there are negative consequences, the whole world has to absorb it. In fact, a favorite phrase of mine lately, I didn't coin it, I wish I had, is, to to socialize the consequences and privatize the profits. And so, you know, Meta gets all the gain of open sourcing their stuff. Their software gets fixed by lots of people, they get a lot of credit, they use that primarily to recruit people that they weren't otherwise getting because they had taken some reputational hits. It's all all plus for them. Now, if it turns out that a bunch of people die because the software was abused, then that's something society absorbed. Meta is not going to make anybody whole. They're not going to bring back, you know, your your sibling that was killed in a in a nerve gas accident.

Tim Scarfe 01:16:14

This is not good. So we have folks like Elon Musk and he's really focused on the the X risk stuff. We've got people like Yan Lakun, for example, who I think in your opinion has adopted many of the talking points that you've been making for many years, but quite dismissive about things like misinformation, for example. I mean, what what do you think about that?

Gary Marcus 01:16:34

Well, I mean, there's a lot of different players. A lot of different people have different incentives. Some people, I think, are speaking from the heart. Some people are speaking from the pocketbook. Like, I don't know if Jan Lecun really thinks that there's no possible risk from AI. I mean, he's actually attacked Elon Musk for spreading misinformation. He must know these tools could be used to generate misinformation. There's already reports that they've been used in the wild and he says they can't be used for misinformation, but he also works for Meta. And like, I don't know what he really believes in and what's an economic incentive for him. Elon Musk is, you know, a riddle in many ways. Lately, you know, he seems to be driven in some ways more by attention than in principle. I was talking to him for a little bit, and I think that he's genuinely concerned about AI risk, but he's also, you know, building frontier models just like everybody else. And the last communication I had with him, which he didn't respond to, I emailed him when he came out strongly for Trump and I said that if you really care about AI risk, supporting Trump might not be the wisest move here. And he didn't respond, and he is supporting Trump. And, like, I don't see how to square square all of the different things that Elon Musk has said. So I really do think that he cares about AI risk. He's not putting that on, but he's doing a lot of things that may actually exacerbate AI risk. So, you know, he's not integrating over all of his beliefs, in my opinion. And then there are other people. I think Dennis Hassabis is maybe 1 of the straighter players. I think that he really is worried about AI risk and behind the scenes he's trying to do something about it. You have Anthropic, I think, really started caring, but now they have visions of dollar signs dancing in their heads and I think they care less than they did before. Open AI, I'm not sure they ever really did. I really lost faith in those people, but they certainly talked more about AI risk, in older times and and, you know, clearly they focused primarily on commercialization and so they've shifted. You have lots of different players. There's a cartoon that I showed in in my talk and is in in my upcoming book, by by Cal from from The Economist. We have all these companies saying, like, AI or I think it's countries rather than companies in the cartoon saying but it would apply just as well to companies saying, you know, we're terrified of the catastrophic risk that AI might destroy us all, and then, you know, that's at the top of the cartoon. And then all the same players are saying, and we're desperate to get there first. Right? So so there's this, tension throughout the industry that I think is comes out most clearly with Dario Amade, who's the CEO of Anthropic, who has been saying that he thinks we might get to AGI in 3 years, which I think is ridiculous. We haven't actually talked about that. But he seems to genuinely believe that. Maybe it's just hype to raise his valuation, but he seems to genuinely believe that. He thinks that his so called p doom is very high. Maybe, like I don't know if he gave a number, but let's say 50%. Like, he thinks this stuff might re-

ally kill us. And he used say, so we're not gonna work on frontier models. Now he's working on frontier models just like everybody else is. And so, like, I don't know how those ideas simultaneously can exist in 1 person's head. Like, how if you really thought that the technology you were building had a, you know, 40, 50, 60% chance of annihilating the species, which I I actually think the number is lower, but if you really thought there was a 40, 50, 60% chance that the thing you were building might annihilate the species, let's say just decimate it, kill you know 10% of it. Like how could you morally keep working on that? But that seems to be where his head is at.

Tim Scarfe 01:20:23

But you think they believe it genuinely?

Gary Marcus 01:20:25

I mean, so that raises the question of like, is it all hype? Does he actually believe it? How does he square like, I can't literally get inside these people's heads. I can just say what I see from the outside reading, you know, what they're doing. I would say I think different people have different beliefs and about why they're doing it and what they think risk is. There are clearly some people whose pee doom over the next 10 years really is like 50 percent. They're not putting it on. You know, my pee doom is much lower than that. I don't I don't think we're gonna annihilate the species. My pee democracy gets really seriously impaired is very high. I I think that that's a likely outcome of LLMs. The idea that we're gonna eliminate this human species seems to me not absurd, but very unlikely. So it's not 0. I'm glad some people in the world are thinking about it and how to defend about it. And I think more generally that alignment is very important even if you don't think about literal extinctions. Literal extinction, it just seems weird to me. Right? We're we're a very, physically diverse species or sorry. Physically sorry. We're very geographically diverse species. We're somewhat physically diverse with different genes. So you think about, COVID was really bad. Right? I mean, it's a horrible, horrible disease, but it still only killed only killed, you know, 1 percent of the population, not 50 or, you know, 80 or something like that. So even if, you know, a system tried to design COVID, it wouldn't necessarily, you know, stop all humans in all places. Like, I just think that that's unlikely. On the other hand, I do think that there's a real risk of some kind of catastrophe. For example, somebody might try to short the stock market, and I hope this isn't 1 of those Marcus predictions that turns out to be true. There's been a lot of them. Somebody might try to store short the stock market by using LLMs to cause as much harm as possible in a shorter an interval. You know, short it just before that, try to make a bunch of money. Like, that might actually happen. And, you know, a lot of people could die because people like, you know, set airplanes to

crash into each other or whatever. So, like, there are real risks like that. There are risks that misinformation could lead to even a nuclear war. Right? So, you know, we blame the Russians for doing something they didn't actually do. They get mad at us. It's back and forth, and then we get into a lot of scenarios like that. And I think the probability that something bad will happen as a consequence of all this, like seriously bad, is actually fairly high.

Tim Scarfe 01:22:56

Yeah. I mean, I I interviewed Syash Kapoor, know, when 1 half of AI snake oil the the other day. He had a great piece out about these subjective probabilities that are used for PDOM, and I and I agree. I'm I'm not a fan of of casting it as a probability.

Gary Marcus 01:23:09

But 1 I I wrote a piece also in Substack some months ago about PDOM, and and

AI X-Risk, Alignment, and Moral Principles

Gary Marcus 01:23:14

like I also think that most of the numbers are made up and post hoc justified and so forth. Yes. Indeed. Indeed. But 1 interesting thing though is is that in a way your strange bedfellows are some of the AI risk folks because of it's it's still safetyism

Tim Scarfe 01:23:28

in in a sense. I mean, how do you feel about that?

Gary Marcus 01:23:32

Well, I mean, I think that goes to the part that I think really is important even though I disagree about the the strong extinction narratives. The part that I think is important is we need to have 2 things. We need to have a global agreement about how we're going to handle AI as it gets more dangerous, sharing information and things like that, And we need technical means for alignment and that will only get more important as AI is a empowered more, which is to say given more responsibility in the world, and b as it actually gets more proficient in doing the things that is asked to do because there will be more opportunities for bad actors to do bad things. You I mean, you want your system to actually be able to calculate, is this going to have an adverse effect on a large part of human-

ity? Then I shouldn't follow this instruction. Like, you actually I mean, the the x riskers aren't wrong to want something like that.

Tim Scarfe 01:24:31

Well, no. That that's the thing. It's very reasonable, and they they have kind of concentric circles of, you know, on on on the inner circle they're talking about being paper clipped, and then they increasingly talk about more and more plausible risks. And the reason I made that comment is I I made a video commenting on your on your on your senate discussion with Sam Altman, and you're famous for being very skeptical about the capabilities of neural networks. At 1 point in the meeting, you were talking about the possibility of them becoming self aware. And that that just seemed that seemed I

Gary Marcus 01:25:02

I don't know what the odds are on that. I think I can't remember what I said, but my general comments on that is I don't think we should cross that bridge. I I don't I mean, it's 1 thing for a system to have a model of, like, where objects are and so forth, but there's a certain level of, like, agency and consciousness that a bunch of people, including at this conference that we were just at, would like to see happen and I do not want to see happen. I think it just opens another, you know, barrel of worms. It's not like I think that's imminent. Like, and you have to understand that most of my skepticism about AI is skepticism about what we can do right now and over claiming and so forth about current capabilities. I think AGI is possible. I I don't expect to see it next week and I offered Elon Musk 1000000 dollar bet when he said that he thought it would happen by 2025 and I would still love if he would make that bet because it's easy money and why not? Even if it goes to charity, I'd be happy to make the bet. Elon, it's still there, and I have a friend who will raise it to 10,000,000 if you're in. Like, I I don't think AGI is close, but I don't see any principal reason why it's impossible. I've never seen a good argument for that. The arguments I've seen about that are all kind of like mysticism or something like that. But I think people will be interested to hear you say that because people interpret your skepticism as you being a kind of Luddite in a way, and that's not really the case at all. It's totally wrong. And I I mean, I've publicly said this a bunch of times that it's totally wrong. I think a particularly good place was when I was on Ezra Klein's podcast and he said, you know, a lot of people might think you hate AI, but you you built AI companies and you've been writing about it for years, you've been interested in a kid. And I said, glad you asked that because I do actually love AI. Like, I wouldn't be doing this if I didn't think there was a good outcome that was possible. If I thought it was hopeless, I would, you know, just stop. I would stop writing about AI if if I thought it was hopeless. Hopeless, I should say, on a technical side or on a kind

of government side. But I do think there's a chance we can do this right, and that's why I, you know, sit here on whatever Twitter every day trying to get us to a better place in AI because I think it is possible and it is is of value. You know, I I not everybody will know, but I'm a high school dropout. And the the way that I pulled that off is and I went straight to college. And the way that I I managed that is I wrote an AI program that translated a semester's worth of Latin into English. So, I mean, I was coding AI as an independent project when I was 15 in the the programming language logo, which is a little bit like Lisp, which I couldn't get my hands on. Like, I have, you know, ate and breathed AI for a long time because I think it's interesting, because the way I think it reflects on human cognition, because of the value I think that it could bring to science and technology. I want succeed. It's just that I think we've gone down this terrible rabbit hole where the people who are running AI now are not researchers. They're marketing type people and not really that interested in the morality of how this stuff is being used, which I think was not always true. So I think that the culture and politics have changed, and I think technically we have gone from a healthy scientific environment where a lot of people were presenting a lot of different models and comparing them, thinking about different ideas, to 1 in which everybody has zeroed in on a technology that I think is fundamentally flawed, is the large language model. It's not that I think there's no value to it, but I don't think it will get us to AGI. I don't think it will get us to the medical discoveries that we might make and and so forth. I think it's the wrong avenue and it's also sucking down an enormous amount of resources. So, you know, a $100,000,000,000 went into driverless cars, basically, mostly on similar technology. I don't know the exact dollars. And, you know, 50,000,000,000 plus has gone in just to chips, on large language models. And anybody who's a grad student works on large language models because that's where the money is. And so you just it's not an intellectually healthy environment. And, like, it's absolute absurdity that you have 7 companies or something like that all basically building knockoffs of GPT 4. Like, that's, you know, tens of billions of dollars that could go towards developing better approaches to AI. Like, people like, the, you know, the people that hate me the most are the EACC people, the evol what do they call them? Executive acceleration. Accelerationists. Yeah. And they don't understand I'm actually on their side in a certain way, but they really don't get it. I've tried to explain this. And the way in which I'm on their side is I'd like to see AI move along faster. And what they can't separate is AI as we do it now, which I think is really not what we want to do, from AI as we should do it. I think that a good EACC person should say, our endgame here is not to make LLMs, you know, ubiquitous. Our endgame is to make AI that would help people and make society better. And ask the question, is this the right way to get there? And the answer is pretty clearly or, you know, I'll say highly likely, no. And so EACC people should be my biggest fans. They should be like, that Marcus is sitting there every day absorbing arrows, trying to

figure out how to bring us to an AI faster that would be better. Like, how could they not like that? Well, that that that's that's a beautiful point because I I had I hosted a a debate with Beth Jezos, you know, the the head of the EAC,

Tim Scarfe 01:30:35

and Connollyhee. And Beth is creating this company to build physics inspired intelligence called Extropic. And Conner's argument was you're basically submitting to the void god of entropy in terms of morality. So, you know, what if we just built this kind of physical simulation of the world and we just trusted it to be moral? So presumably, you wouldn't be a fan of that.

Gary Marcus 01:30:56

I mean, I'd have to know the details and I watched a little of the video and wasn't, you know, wildly impressed, but so didn't pursue it more. But I I would say that building a simulation is not building a moral reasoner. Those are not the same things. And so, you know, for example, there are some scientists who understand physical laws pretty well and some of them are ethical and some of them aren't. Like, it's just a different thing. Like, you want some philosophers involved in in in the ethical side and so forth in the ethicist. They're just different things. If you have a perfectly faithful simulator of the world, which I think, you know, gets into Laplace's demon and is not a realistic thing in the actual world with finite resource, But if you had it, it still wouldn't give you ethical decisions. You know, we still have law, legal frameworks, ethical frameworks, and so forth independent of our knowledge of how the world operates. Now, I think there's a small connection, which is 1 of the things an ethical AI system ought to do is to evaluate the consequences of its actions. And so the better the simulation you have, not just of physics, but of like sociology and, you know, there's multiple layers of understanding that we should have of the world, levels of understanding. But if you could understand economics and and psychology and so forth, you could simulate all of these really well, probably better than is actually realistic. But if you could do that, you know, even decently well in the context of an ethical system that is trying to avoid harm to humans, then great. You know, think about Asimov's laws. You know, the first do first do no harm. Like, think about that from a technical perspective. You have to actually calculate, would this action cause harm? And the better the simulator you have, the better you can do that. But you can only like, you have to know to do that. Right? If you have the ethical principle of do no harm, and of course, Asimov's stories were about how you get into various tricky, you know, edge cases or whatever. But if you start with a system that tries to anticipate the consequences of its actions, then a simulator is very helpful. If you don't start with that premise and you just have a simulator, like, so what? Like, I mean, you can use the simulator to, you know, do a bunch

of things, but, like, you have to have the ethical framework in order to leverage that. Well, exactly. But but for you though, if if we need to have a moral calculus, presumably

Tim Scarfe 01:33:13

you you think there are moral facts about the world, where where do they come from? Mean, they're they're moral axioms. They're not facts. Like, have to

Gary Marcus 01:33:21

you have to take a position there. There's no moral set of facts independent of humans in the world. Right? And that makes a lot of people run away screaming. They're like, we're never gonna agree. But the fact is there's a lot of stuff we actually do agree on. We don't talk about the stuff that we agree. But if you took a survey of, you know, a 100 people on the street and said, is murder okay? 98 of them would say no. You could also say, what if it's in self defense? And a lot of them would say, okay, maybe, and, you know, tell me about the circumstances and so forth. But, you know, there's a lot of stuff we actually agree on. There's some stuff we don't. So when is the beginning of life? Like, you know, people disagree about that, and then it's not clear there's a fact of the matter and so forth. But there are some moral principles, a lot of moral principles that most people actually agree. We mostly agree don't steal stuff, don't kill people. You know, the 10 commandments wouldn't even be such a terrible start, but we don't even know how to do that. Right? So, like, you know, don't lie is, you know, maybe a principle that people are less agreed on, but to a first approximation, they agree on it, but you can't tell an LLM don't lie. They just it does not compute in an LLM. Right? They literally don't understand that. So, yes, there are disagreements. There are cultural differences and so forth. And people wanna play a slippery slope argument and say, don't even try to build ethical AI. But that's actually taking a stand too. Right? It's in a way, it's taking an extreme libertarian view of anything goes. And at the most extreme, I think even libertarians aren't actually that comfortable with that. I think there's a lot I mean, libertarians are not in fact all saying, well, we should just legalize murder. Like, they're just not saying that. And if they did, we wouldn't take them seriously.

Tim Scarfe 01:35:10

I think there was an example you gave in your book though of of a huge cultural difference in China. My my mind's slipping on it. But presumably, you do think that AI legislation should be enacted quite differently in different cultures.

Gary Marcus 01:35:23

Well, mean, the thing that I object to most in Chinese AI regulation, which is actually much stronger than US AI regulation, is an insistence that everything be consistent with the party line and I, you know, I I would not like to see that in The United States. I don't like seeing it in China. It's sort of, you know, it's not as much my business or maybe none of my business if they do that in China. I certainly don't want anything like that in The United States ever. The the, you know, it's not legal to operate a chatbot if it doesn't, you know, conform to what the state believes is, you know. So it's 1 thing to say chatbots shouldn't promote, you know, conspiracy theories that are clearly false or whatever. Like, know, I can imagine something in in that direction, but but, you know, becoming an instrument of the ruling party. No. I I don't I don't think that that's just don't think anybody on the planet should do it. I can't stop China from doing it, but I think it's wrong that China is doing it.

Tim Scarfe 01:36:22

So, Gary, almost exactly 4 years ago, we invited you onto MLST. We had early access to GPT 4. It was actually Conor who gave us access for a secret API, he wasn't supposed to, but and we invited you on and Waleed Sabah and we spoke about RR3, it was probably GPT-three. It was GPT-three 4 years ago, think it was about November 2020, And the amazing thing, first of all, is the consistency of your position. And even now, the video's got hundreds of thousands of views. Every day, are commenting on it, they're saying it's aged like fine wine, but but in quite a polarized way. What I said was it Yes. Yes. Isn't that incredible?

Gary Marcus 01:37:00

I mean, I've not been able to put people in the headspace that I am. To me, none of what's happened is surprising. I mean, I figured out basically

Persistent AI Flaws: System Limitations and Architecture ●

Gary Marcus 01:37:10

how this stuff works in 1998, and it hasn't really changed. There's statistical approximators that don't have world models. And once you understand that, the likely flaws like hallucinations just jump out. And so, you know, I was writing about hallucinations in my 2001 book, The Algebraic Mind, and it's not like there's ever been a principled solution to hallucinations or the stupid errors of of

reasoning, and it's not even reasoning. It's approximation. The stupid errors that come from approximation. Nothing has ever been principled. Nobody has come up with a principled technology to solve them. And so, of course, those errors still happen. Like, to me, it's just obvious that all this is gonna happen, and what I always get is, well, we will just add more data, and it'll solve it. And I say, it's not gonna solve it because it's not a principled solution. Come back to me with a principled solution. Explain to me, for example, in the passage in which I anticipated hallucinations in my 2001 book. I gave the example of my aunt Esther who sadly passed away last summer, and I said, suppose that she wins the lottery, and what's going to happen to a system that doesn't have a distinct record system for individuals as opposed to kinds? What you're really learning about in these systems are properties kinds, not individuals. You don't have an individual level predicate, and I noticed that Subaru has started saying some fairly similar things recently. So I made this argument back in 2001. I said, this is principal problem. If the system is told my aunt Esther wins the lottery, it's gonna think that other women who live in Massachusetts or who work for Harvard or, you know, that share some properties with her also, you know, are a little bit more likely to have won the lottery. But that's not how the lottery works. Right? Other people that are in your, you know, pigeonhole advertising category didn't also win the lottery. The system doesn't understand that this individual property inheres in this individual because it doesn't have a mental representation like a database of particular individuals. And so it was plain as day to me. I think I did some modeling on 1 of the models of the time by Ronald Hart and Todd, if I remember correctly, showing that that kind of error would happen. And then you wanna come to me and say more data is not gonna is gonna solve that problem. Well, I did this work back in 1998. I know that's not gonna solve that problem. You you need an actual solution to that problem. More data is just hand waving. So, of course, my predictions, you know, age like fine wine. The only thing that would make those predictions break is if somebody actually came up with a principled solution and said, here's how we're going to handle compositionality in this architecture, or here's how we're gonna handle, distinctions between individuals and kinds or structured representations or operations over variables. And instead, people are just hoping that, like, by magic, by having enough data, they'll go away. And they get confused because any individual puzzle that I put out there may get solved. I mean, we saw with the, Horse Rides Astronaut, even sometimes they out in the literature, they still don't get solved. But in general, any given puzzle can be because these systems are in part big memorization machines. That's not all they are. It's a bit of an oversimplification. But you can think of them roughly as lookup tables. And so you can store, you know, multiplication tables, a lookup table. So you can store anything you want in a lookup table and fool yourself into thinking your system has it. So you could take a young child who hasn't really learned what multiplication is, give them a multiplication table, and they could

tell you what 6 times 7 is because they memorize the table. That doesn't mean that they understand multiplication at an abstract level. And so what happens time and time again, and I've been watching this for 30 years and it kinda breaks my heart to keep seeing it, is that people will see a single correct answer and assume that the system has the correct underlying abstract abstraction because they don't understand how to think about data and errors. Like, that's what I really learned from Steve Pinker when he was my mentor, was how to think about data and human errors and algorithms and what the I mean, people don't have training in that, so they just say, well, if there's some generalization, then the system is done. And in fact, the point of my 1998 work, which I still think is my best work, was to show that there's 2 different kinds of generalization. There's generalization that's kind of nearby to the examples you've seen, and there's further generalization that's further away. Nowadays, we call that distribution shift. I basically was writing about distribution shift and how neural networks are or are not good models of human minds in 1998. So once you understand there's different kinds of generalization, which may be 2% of the population or 2% of the people in machine no. I guess that's not fair. But, you know, 2% of the population understands, and a growing fraction of machine learning understands. Once you see that, you realize that there are serious problems. There are a lot of people who just don't understand that basic distinction, and they imagine a level of generalization that is not there. It's kind of over anthropomorphizing the system, etcetera. And once you're down that path, then you believe, okay, more data will solve this or whatever. Once you're down the correct path, it's just obvious. Once you really understand what these systems are doing, it's just obvious that they're not miracle workers, that they're not gonna solve these problems. These core problems are not gonna go away. I took 2 you know, I looked at Sora for 2 seconds, I said, I'll bet you that that's gonna have problems tracking objects over time. Because there is no representation here of the individual objects. It's at a pixel level, and sure enough, it had tons of problems. Like, I look at these systems and I can usually take them apart in a few minutes because I can kind of see how they're working and what they're lacking from a and that comes from a, you know, perspective of being a, a computer programmer since I was 8 years old, and b, a cognitive scientist since I was 13 years old. And so I just I see it and I know. It's annoying how dismissive people are. I mean, our Chomsky video, it's had nearly 1000000 views.

Tim Scarfe 01:42:59

It was the most amazing experience in my life. Read the YouTube comments. And I don't know whether people just don't understand it. It's so annoying watching people just write you off.

Gary Marcus 01:43:09

Yeah. People write Chomsky off all the time. And they'll I mean, they'll do it, like based on his politics. Like, I disagree with his politics so he must be wrong about linguistics. That doesn't actually follow. Right? I mean, I actually disagree with his politics, disagree with some of his linguistics, but when he says that large language models haven't taught us much about human language, he's right. You know, it's like we don't have a better theory of how human use language from large language models. Large language models, as it happens, depend on having like the entire Internet as a corpus and it's obvious to anybody who's had a 3 year old child that, you know, they they listen for a couple of years, not to the entire Internet, and then they basically understand it. I know. But that there are trolls everywhere. Mean, many of these people don't probably know that you sold an AI company with with Kenneth Stanley called Geometric

Tim Scarfe 01:43:54

Intelligence.

Gary Marcus 01:43:55

Ken Stanley, Zubin Garamani Yeah. Doug Bemis, Jeff Clune. Yeah. We were all part of that. I I was the first person to kind of launch the company. I brought in Zubin, then eventually we brought in Ken, eventually we brought in Jeff. Yeah. Most people don't know that. I mean, it's there and Wikipedia is there to be found. Every now and then, I'll I'll push back on the trolls who attack me and say, well, you know, do you have papers in science and nature? Did you sell a company to Uber, you know, machine learning company? Of course, you know, they didn't. I'm like, I don't know how to respond to that. Like, people either put in the time to actually understand who I am or they don't want to. I mean, they really don't want to. They they want to demonize me as as with Chomsky.

Surveillance, Economic Challenges, and Neuro-Symbolic AI

Gary Marcus 01:44:33

Surveillance capitalism is basically that the companies make their money by invading your privacy and selling information about you to, you know, all sorts of people either directly or indirectly. And I have to say that something that just, like, turned my stomach the other day was OpenAI bought a webcam company. So I already knew that they were trying to, like, access all your documents. They have a deal with Microsoft that was gonna do this thing taking screenshots every 5 seconds. I think the public may maybe shot that down. Initially, they had a database there in plain text. So even if Microsoft wasn't correctly gonna do something bad with it, like any bad actor with any, you know, chops would have been able to take all this stuff. So already you had this situation where they're clearly trying to get everything they possibly can about you and to commercialize it, you know, probably to sell ads. And then they bought a webcam company. I was just like, what like, they're having a lot of trouble actually making money at OpenAI. Like, they lost, you know, operating loss, you know, cost relative to to revenue last year, $5,000,000,000. A lot of companies tried their software and are not that excited about it, having tried it out. Copilot, which is run on their platform, a lot of people aren't that happy with it. Microsoft is, like, trying to rush people to use the product. Everybody tried it. It's not like there was a lack of marketing, but it's not that reliable. So OpenAI is faced with a problem. How are we going to make money off of this stuff? And it's not cheap. Right? I mean, they've probably spent, whatever, $10,000,000,000 on on data and chips and, you know, they're gonna spend another 10 or whatever. Mean, GPT 4, who knows how expensive need to make money or they will either go out of business or Microsoft will take a bigger chunk of them. Like, it's not clear that VCs are gonna wanna up their valuation to $200,000,000,000 given the kind of picture of revenue that is is emerging. And so if I were Sam Altman and I had a different kind of morality than I think I do, first thing I would think about is surveillance. Right? They have people are giving them all this data on a silver platter. Now you got a camera too? Like, that's that's the natural direction for the company to go despite their name, quote, OpenAI. Right? I mean, it's a really sinister thing for them to do to become big brother. Did I mention they're still a nonprofit, which is insane? So we're gonna we may very well wind up with the world's largest surveillance company as a wholly owned subsidiary of a place that calls themselves OpenAI that's nonprofit. Like that is, I think, almost a default outcome right now. That's to me upsetting. What about the NSA guy they employed?

Tim Scarfe 01:47:13

Well, that's clearly part of the same picture. I mean, if I were them, I would be building language models on, you know, the government capture or the surveillance data. I'd be building a language model on that.

Gary Marcus 01:47:23

I'm sure the government is interested in doing that and I think, you know, they hired Nakasone, Paul Nakasone, who worked at the NSA. Clearly well, I won't say clearly, but it would seem plausible that they they hired Nacasone to try to get those government contracts. Like, that would seem to be the play that they have.

Tim Scarfe 01:47:41

Yeah. Because other than the government, OpenAI is interesting because they go direct to consumer.

Gary Marcus 01:47:46

The rest of the LLM ecosystem are developing, you know, enterprise models. What what do you think about that? I actually don't see the winning strategy for anybody but NVIDIA right now. So NVIDIA is selling shovels in the gold rush. People wanna buy the shovels. You know, it's not their fault that people wanna buy the shovels. They've been hyping things a little bit, but, you know, in general, wanna buy the shovels, and they know how to make the shovels. They spent a lot of time. They're a very well run company thinking about how to make the best possible shovels on the planet, and they really have that, and they're selling them. They're making a lot of money. Everybody else, I think, is struggling. Like, Midjourney is making a profit, but there's gonna be massive lawsuits, I think, against them or else they're gonna have to pay massive licensing. So at end of the day, I don't know if Midjourner is gonna make a profit. And most of the other companies, you know, they have modest revenue. They're valued at, like, 200 times earnings and that kind of stuff. Not even 200 times net profit, just 200 times revenue. These are kind of crazy numbers. I don't think anybody has a strategy yet, maybe in like a small sector, but I don't think anybody has the strategy yet to make, you know, multiple billions of dollars profit year in and year out with this stuff. I think that's still a dream. I don't think anybody has a clear way yet of doing that on the business to business side or on the business to consumer side. I won't say absolutely nobody will get there, but there's a really serious problem for all of them, like deadly possibly problem, which is that Meta is now giving away for free es-

sentially the same product as everybody else is making. So there are ways you can deal with that. You can have better customer service. You can, you know, have a particular training set. So, like, it's not absolutely knocked down. It's all over. Folks go home. Like, there's still a chance, but that is a pretty serious problem. And, you know, whatever you're doing, you got a competitor now who's gonna do it on Meta's platform. You're doing it with your own dataset. Well, someone else is gonna fine tune on Meta's platform.

Tim Scarfe 01:49:51

You're not you're not gonna pay for OpenAI anymore. Is it a bubble, and is it going to burst?

Gary Marcus 01:49:58

So I'll say yes and no. So I think that the economics don't work. I wrote a piece last year called Is Generative AI Gonna Be a Dud? It was almost exactly a year ago to the day. I think that that was prescient and I I think that it's becoming clear to all that the economics don't really make sense. I think a lot of investors are going to lose a lot of money, lot of LPs like pension funds that put in money to the investors are gonna lose money. I think enthusiasm is gonna be lost. I think that the tools will remain. You know, Meta's gonna keep giving them away free and people will still find some uses, but I think like a year from now things are going to look different, know. Many people who like move from crypto to AI are going to move to something else because they're going to say, well I don't see how to make money out of this. It may be that investment dollars become much scarcer. LLMs will continue to exist but eventually people will try to find other solutions. I think neurosymbolic AI is maybe finally going to have its day. The recent results from DeepMind with alpha proof and alpha geometry were impressive. You know, it's a moment where I think there might be a shift. So generative AI, I think, will never be as popular again as it was in 2023. The tools will always be there, but it'll be more like, yeah, use it for brainstorming and, you know, it won't be like, this is the Messiah. This is, like, the greatest thing ever. Like, people were saying, like, AI is, you know, gonna be bigger than fire and electricity and whatever. And the truth is most people would not take generative AI over their cell phones. Like, everybody, you know, you you if you take away somebody's cell phone, they're gonna be pretty upset. A lot of people could actually live without generative AI. They played with it. It was fun. But if you said, look, I'm gonna charge you $50 a month for it, most people would say, yeah, I don't really need it. Most companies have tried it out and they're like, yeah, we can find something, you know, but people could live without it. And the hype last year was that it was this indispensable thing that was going to transform our entire world. AGI, if it comes, might be like that. Right? I mean, you really had a system that can under-

stand the cognitive requirements and the physical requirements of any human job, that's gonna completely change our world. And I think that that will happen, maybe in my lifetime, maybe not. I think it will happen someday. But generative AI was not that, and it was never going to be that as I told you whenever it was 3 years ago, that reality has settled in and it's just not gonna be treated in the same way. Plus you have the fact that Sam Altman's star has fallen somewhat. You know, a lot of this was like his personal charisma. But a lot of people are skeptical of Sam at this point and I think with good reason. And so you had this kind of rock star and you had this, you know, constant hype, like, every day in every newspaper. And in 2025, let's say, people still use generative AI. They'll still be in the news sometimes. Open AI will still be in business, but people will think the $86,000,000,000, whatever it was, valuation was probably too high. They can be like, how are they gonna make money when Meta is giving us a waiver free? It'll still be there, but it's not gonna be this, you know, it's gonna be like pet rocks. You can still buy a pet rock, I think, at a toy store. Like, you know, pet rocks were most of your listeners won't even know what I'm talking about. But at 1 point before I was born, like, everybody was buying a rock and calling it their pet. It was this crazy thing or or, you know, Tamagotchis and and and, you know, you have these things that become fads that completely take over the world for a little while and then they disappear. You know, a point that I've often made is that people over extrapolate. They think there's an exponential after they see a few data points. And there are different ways you can, make fun of it. 1 of them I like I call the disco illusion. And this is somebody else's cartoon. Maybe we can put it up. And it's like sales of disco albums in like 1974 and 1976. And someone's like extrapolating, I forget exactly how the cartoon goes, that like in 1978, every record that is sold will be disco and and forevermore. And we all know it didn't work out that way. Disco had its moment and then, you know, synthesizer music and and punk and all kinds of the other things took over. People got tired. It was fun for a while. Chatty PT is gonna look a little like that. It's like, you can still buy disco albums, and I still like the, soundtrack to Saturday Night Fever, and I listen to it sometimes. Like, it still has, you know, a role in my life. It didn't completely disappear. I feel nostalgic about it, but I don't spend, you know, 24 hours a day listening to disco anymore, and people are not going to spend as much time with with the chatbots, and, you know, it's just not going to be the center of our lives the way it has been really since November 2022 when it came out.

Tim Scarfe 01:55:04

Met is interesting though because they're building this metaverse and virtual reality and stuff like that. Ironically, they do have some potential use cases for the technology but some some business leaders have said to Well, I think

Gary Marcus 01:55:16

you can come back to the rest of your question. I think a lot of people imagine you could populate a metaverse world with large language model characters. And I think that that's gonna get tired fast because those characters aren't really gonna understand the world and so it'll work for a little while. You know, having non player characters made by chatbots, but eventually like things are not gonna make sense and it's gonna break the illusion. The only way metaverse is gonna work is if the the illusion is so potent and it's so much fun to play with. And the more that, like, you have these weird experiences like, today I had grok. I asked it to make 11 eggs without an egg carton. Instead, made 15 with an egg carton. And we have enough experiences, and you're like, yeah. I know. Was I was surprised by that. But but

Tim Scarfe 01:56:11

wouldn't it be interesting that it might be therapy, it might be virtual partners, it might be gaming. Could there be an application, a killer application, that does change it? They change metaverse? What would that make generative AI, you know, economical?

Gary Marcus 01:56:26

I mean, I think the killer app is gonna be surveillance. I think that's the, you know, selling everybody's personal information is is, you know, historically been a profitable business, and I think that's where they're gonna land. But I don't know. Professor Gary Marcus, thank you so much for joining us today. It's been a pleasure. I love our conversations.

References

[1] **Gary Marcus - Taming Silicon Valley**
https://amzn.to/3XTlC5s
book

[2] **Gary Marcus Substack**
https://garymarcus.substack.com/
person

[3] **Evelina Leivada et al. - LLMs Understanding Human Language**
https://arxiv.org/pdf/2308.00109
paper

[4] **Gottlob Frege - Compositional Semantics**
https://plato.stanford.edu/entries/compositionality/
other

[5] **Jean Piaget - Object Permanence Theory**
https://en.wikipedia.org/wiki/Object_permanence
person

[6] **Alan Turing - Computing Machinery and Intelligence**
https://academic.oup.com/mind/article/LIX/236/433/986238
paper

[7] **Nicholas Carlini - Chess with LLMs**
https://nicholas.carlini.com/writing/2023/chess-llm.html
paper

[8] **Mathieu Acher - GPT-4 Chess Analysis**
https://blog.mathieuacher.com/ChessWinning7MovesGPT/
paper

[9] **Jiexin Wang - AI-Assisted Coding Security**
https://arxiv.org/pdf/2407.02395v1
paper

[10] **Rodney Brooks - Three Laws of AI**
https://rodneybrooks.com/rodney-brooks-three-laws-of-artificial-intelligence/
paper

[11] **Cruise Teleoperation Revelation**
https://www.nytimes.com/2023/11/03/technology/cruise-general-motors-self-driving-
html
other

[12] **California SB-1047 AI Regulation**

https://apcp.assembly.ca.gov/system/files/2024-06/sb-1047-wiener-apcp-analysis_0.pdf
other

[13] **Gary Marcus - Open Letter on SB-1047**
https://garymarcus.substack.com/p/an-open-letter-to-fei-fei-li-concerning
paper

[14] **European Commission - EU AI Act**
https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai
other

[15] **A&M Records v. Napster - Copyright Precedent**
https://en.wikipedia.org/wiki/A%26M_Records,_Inc._v._Napster,_Inc.
other

[16] **Shoshana Zuboff - Surveillance Capitalism**
https://amzn.to/3ZqHAxS
book

[17] **Jaron Lanier - Twitter Poisoning**
https://www.nytimes.com/2022/11/11/opinion/trump-musk-kanye-twitter.html
paper

[18] **Adam Curtis - HyperNormalisation**
https://www.imdb.com/title/tt6156350/
video

[19] **Firehose of Falsehood - Russian Propaganda Model**
https://en.wikipedia.org/wiki/Firehose_of_falsehood
other

[20] **Chris Lu et al. - AI Scientist Paper**
https://arxiv.org/abs/2408.06292
paper

[21] **Luciano Floridi - Digital Ethics**
https://www.youtube.com/watch?v=YLNGvvgq3eg
video

[22] **Sayash Kapoor and Arvind Narayanan - AI Snake Oil**
https://www.aisnakeoil.com/
book

[23] **Gary Marcus - p(doom) Analysis**
https://garymarcus.substack.com/p/d28
paper

[24] **Isaac Asimov - Three Laws of Robotics**

https://amzn.to/3XTIwtl
book

[25] **Seymour Papert - Logo Programming Language**
https://el.media.mit.edu/logo-foundation/what_is_logo/logo_primer.html
person

[26] **Gary Marcus - The Algebraic Mind**
https://mitpress.mit.edu/books/algebraic-mind
book

[27] **MLST - Noam Chomsky Interview**
https://www.youtube.com/watch?v=axuGfh4UR9Q
video

[28] **Gary Marcus - Neural Networks and Generalization**
https://www.sciencedirect.com/science/article/pii/S0010028598906946
paper

[29] **Beff Jezos - Physics-Inspired Intelligence**
https://www.youtube.com/watch?v=0zxi0xSBOaQ
video

[30] **Laplace Demon Concept**
https://en.wikipedia.org/wiki/Laplace%27s_demon
other