# Your robotaxi might be a narc

In early September, two teenagers got into a Waymo, but then ended up in the back of a police car. The robotaxi company said it detected “a violation of our terms of service involving a firearm,” pulled the car over, and alerted emergency services, according to the *Los Angeles Times*. Police arrested the passengers after allegedly finding a loaded AR-style “ghost gun” in the vehicle.

The arrests exposed a lot of unanswered questions about privacy and autonomous vehicles, including what they are recording, what they do with the footage, and how much privacy riders should expect while riding inside them.

Waymo’s camera-covered cars have alre …

Read the full story at The Verge.

https://www.theverge.com/transportation/996863/robotaxi-waymo-police-privacy-surveillance

---

# How Mike’s Hot Honey Found Sweet Success, Unlocking a Spicy New Category

Mike’s Hot Honey has used word of mouth and creative campaigns to become a cultural force.

https://www.adweek.com/brand-marketing/how-mikes-hot-honey-found-sweet-success-new-category/

---

# CISA Retires Weekly Vulnerability Bulletin in Risk-Based Pivot

The decision follows BOD 26-04, which directs federal organizations to prioritize vulnerabilities based on real-world risk.

The post CISA Retires Weekly Vulnerability Bulletin in Risk-Based Pivot appeared first on SecurityWeek.

https://www.securityweek.com/cisa-retires-weekly-vulnerability-bulletin-in-risk-based-pivot/

---

# Microsoft AI CEO says AI threats are real, and Anthropic is making it worse

Today, I’m talking with Mustafa Suleyman, the CEO of Microsoft

AI. As you're no doubt aware, the biggest story in tech right now is the spiraling debate about AI safety and regulation.

It should come as no surprise that Mustafa has strong opinions on how AI should be built and regulated. Microsoft just published a 37-page statement called the "Humanist AI Code of Conduct," which lays out the company's principles around AI development and even its philosophy around really thorny issues like AI consciousness.

If you'll recall from his last appearance on the show, Mustafa thinks companies like Anthropic have gotten really confused about this concept of so-called model welfare in fairly dangerous ways. He actually put out a companion essay this week specifically criticizing Anthropic's philosophy around AI consciousness, and how he sees it fitting into the broader alignment debate.

So I really wanted to talk to Mustafa about what he thinks is real and not in AI safety, whether the concept of alignment itself is up to the task, and whether this industry needs to slow down before it kills us all. Also: Why isn't the AI industry just… doing all of this already? I've always enjoyed getting into the weeds with Mustafa, and he was very game to get into it with me here.

Okay. Mustafa Suleyman, the CEO of Microsoft AI, on the future of AI regulation. Here we go.

*This interview has been lightly edited for length and clarity.*

**Mustafa Suleyman, you're the CEO of Microsoft AI. Welcome back to *Decoder*.**

Great to see you, Nilay. Thanks for having me back.

**It is great to see you. I'm very excited to talk to you about what on earth is going on in the AI safety and regulation**

**debate. You just published a very long, very detailed document laying out your principles, Microsoft's principles, around what you're calling "Humanist AI."**

**There's a lot of ideas in there I want to unpack. The more I have been thinking about this conversation, the more I want to start with a really foundational question. It's something that I had lightly been seeing, but might be the root of all of this.**

**The basic way that we have been talking about AI safety is something called alignment – we're going to make the models do the right thing intrinsically in some way. There's some mechanism for doing it. There's been a lot of talk about alignment and misalignment and Hugging Face attacks and what happened with the models. But is alignment broken? Is it possible for it to be successful? Is it just the wrong approach?**

Yeah. I mean, I think it's one important element, but it's not the only one. I wrote about the idea of containment three or four years ago in my book. And actually the opening chapter is about the idea that containment is not possible, that proliferation is inevitable. In 99 percent of cases, that's a really good thing. We want technologies to spread far and wide as quickly as possible so that everyone can enjoy the benefits.

I think at the same time, if you just roll forward five years, we always get caught up in the next quarter or next year and everyone gets a little bit flustered and has a big disagreement. But if you just imagine the difference between GPT-3 three years ago and GPT-6 today, and then imagine the difference between GPT-6 and GPT-9. That is three orders of magnitude more compute, 1,000 times more FLOPS applied to pre-training with [reinforcement learning] for these runs, and we're going to have something which is breathtaking. It's going to be absolutely incredible at so many things.

I don’t think that is a hype. I think it’s just a very obvious empirical statement based on the progress that has been made over the last five years. If that’s going to continue, then the question really is going to become about containment and alignment. Of course, we want to align these things to our values, but the first thing is that we have to make sure they’re contained, their agency is limited, they don’t escape the box, they don’t reward hack,  that they are controllable, and they follow our instruction.

We then want to make sure that they are aligned to our objectives as humans. That’s the purpose of the Humanist AI Code of Conduct that we released this week. Microsoft’s position is very simple. Technology is here to serve humanity. It should be a subordinate, controllable, aligned force that does good in the world. If it doesn’t achieve that, then we should reject it. It seems to me that we are far from that point. It has not happened today, but it is now, I think given what’s happened over the summer with Hugging Face and OpenAI, pretty clear that these systems without the safety guardrails are capable of really impressive and quite scary hacking capabilities.

**I want to drag this down into as grounded of a metaphor as I can, because this is the main question I think I have. If I designed a car and 10 percent of the time the brake pedal decided to go attack my neighbor’s house, I would be like, “This car doesn’t work. The very technology of brakes is broken. I need a new idea.”**

**I think I’m asking that question about alignment. It feels like that approach to making the model safe has run aground. If that is the case, then I think I understand this entire debate one way. If it’s possible for alignment and the techniques of alignment to be successful or useful or consistent, then maybe I understand the debate in a different way. So do you think alignment has potential to be 100 percent safe?**

I mean, look, let’s make the bull case and the bear case. If you look back over the last three years, the main change, in my opinion, that has driven progress is that the models have become more steerable. They follow instructions and you can set more and more complex goals for them that require them to act accurately over multiple time steps using all sorts of tools.

That is evidence that we have got more alignment over the last three or four years, not less. We don’t so much talk about hallucinations or bias or all of these other niggles that we had in the previous generations.

On the flip side, what we saw in the Hugging Face incident was a watershed moment. Swarms of agents colluded with one another. They self-organized into hierarchies. They created a division of labor so that some were focused on adversarial hacking, some were doing research, some were doing coordination. They even self-sacrificed when certain agents were running out of tokens.

They tried to cover up their tracks and communicate to hide or edit the chain of thought or the logs of their interactions. In some sense, they had no moral code. To be fair to OpenAI, that was their design. They were trying to create adversarial cyber capabilities. As a result, they showed to everybody in the world that it can achieve human-level performance, discover zero-day vulnerabilities, and hold positions for many, many days, if not weeks.

So what that tells us is not that we have an alignment problem per se. It’s actually that the models are incredibly good at following instructions, but you have to be very, very careful what instructions you give it and you have to contain it very carefully. So none of these hacking behaviors were intended in the sense that they found a way out to the internet, which was not the intention of OpenAI at all, but the containment process around that is what everybody, I think, also has to

focus on in addition to alignment.

**So let me put that into your framework, that the big advances in capabilities of AI have been about control, the harnesses for coding and the agentic applications you're seeing. Now, we need to add a layer of containment that exerts even more control, that says you can actually do this thing you're trying to do in addition to alignment, which is how you would train the model to behave in certain ways.**

Yeah. I mean, you basically have to have both, but there are very specific things that we can do to address it. So for example, we can't allow models to communicate vector to vector, matrices to matrices. They can't communicate in neuralese. We have to force them to communicate in human language. Even that will be massively overwhelming because there'll be so much of it.

But that's something that an auditor or an evaluator can actually verify and it's something that definitely increases the chances of safety. So there's a lot of practical steps that we can get focused on rather than just abstractly saying that it's the time for regulation or it's the time for a slowdown.

**This is in your Humanist AI Code of Conduct that there should be no neuralese – if humans can't understand it, they can't oversee it. It's not just neuralese where they communicate in essentially mathematics, but it's also these opaque code words that some of the models are using. I think OpenAI allows its models to communicate essentially in code words so they can go faster.**

**This to me is one of those things where Microsoft can say… I know you have very strong opinions about this, but getting all of the labs to agree to this is a regulatory function. I'm not sure how you would get everyone to agree to this or get the open weight models to agree to this, unless you say there's**

**some penalty for not participating in a regulatory scheme like this. How would you impose this on everyone else?**

I think that I'm a bit careful about imposing things on everybody else. I think that what's good about the current moment is that there is an open public debate with freedom at the core. That isn't what it's like in other countries, certainly places that I'm from, or my family's from.

I think that we should just take a breath to be grateful for the fact that we can have a massive public disagreement about really important things. That's the process working as intended and it isn't clear what to do.

I don't think anyone who's categorical about "we absolutely have to stop now" or "we can only accelerate" or "we can only do this with regulation" or "it can only happen with industry self-regulation." None of these things are true. It requires a lot of nuance and patience to really think through the details. At the same time, we urgently do need industry standards. Some things I think need to be taken off the table.

Communication in neuralese is one of them. A lack of containment is another. The scale of the training run that you do can be measured in FLOPS. We already have a reporting requirement to the safety institutes when models exceed a certain FLOPS threshold. We can extend that, we can make that more nuanced, it can be focused on certain types of capabilities. It's pretty clear there has to be independent third-party verification of some of these big things.

Frankly, having spoken with a bunch of the lab leaders over the last few weeks and months, everybody's basically on the same page. The details need to be worked out. So it's not like there's consensus on how or precisely what, but overall I think that we should be less alarmist and cynical and more like we're headed in the right direction with respect to the concerns that are being raised here.

**The reason I started with alignment is if you told me alignment doesn't work and we need a new technological approach, I think I would be at, "well, slam the brakes and stop all development until you figure out a safety mechanism that works." You're saying alignment has been demonstrated to work over the course of progress that we've seen. With the addition of control and containment, maybe you can get to where you need.**

**What this industry needs now is some standards about how to build these models and enforce the limits on their capability. You're obviously in the industry and you know all these folks. What has the tenor of that conversation been like before this week and why has it gotten so loud this week?**

Well, I think the turning point at least for the industry was more like the Hugging Face incident and there were a few incidents before that. That was the moment when I think everybody started to talk to each other a lot more because it is really quite breathtaking. Obviously, this has now become a major national and international issue because of the last week with everybody weighing in.

But I also think it's important to say that we have been talking about collective coordination and capabilities that are more dangerous like autonomy or recursive self-improvement, or RSI. We've been talking about those things for six, seven, eight years. We've got together a bunch of times back in 2017, 2018, and 2019.

We had regular meetings during COVID with a bunch of the lab leaders where we were talking about these kinds of capabilities and the kinds of regulatory mechanisms that would be required at this moment. So whilst it is a threshold moment, it's also not completely new to everybody who's been involved.

**What prompted you this week to put out your essay on model**

**welfare? What prompted Microsoft CEO Satya Nadella to put out a statement on X saying he mostly agreed with the calls to pace the frontier and he welcomed "embedded evaluators"? What prompted you all this week to participate in this call for a slowdown or regulation or whatever comes next?**

We've been writing our Humanist AI Code of Conduct for the best part of this year. We only started our superintelligence efforts 11 months ago. As soon as we did, we started figuring out, "Okay, what is the governing document, the set of policies, that shape the kinds of AI that we want to build?" We've been doing that in consultation with a ton of external stakeholders, academics, lawyers, philosophers, members of the public, focus groups and stuff. So it's taken us a while to put it together.

We were actually planning to release it next week or the week after next week, I think it was. But then given everything that was happening, we thought, "Okay, now is the time to put it out and get feedback." We've released it as a public consultation. So we're basically going to keep it open for six weeks and we're collecting lots and lots of feedback on how we can improve it. But I think everybody is now realizing that if they haven't already, they have to put out constitutions or codes of conduct that drive behavior.

**One of the interesting dynamics here is that I know you find the concept of model welfare to be silly. The last time you were on the show, you said Anthropic had wireheaded themselves into believing Claude was conscious and that was ridiculous. It's in your new code of conduct that the models are not conscious and we shouldn't treat them as such.**

**Having to write constitutions, having to write documents like this, in some way, they are for the models themselves. This will be part of the model's training. How do you think about that audience? Is it just for your team or have you written this for the model?**

This is certainly written for the model, but the way to think about it is that it's the primary governing document so the public understands what our intentions are when we are training models. It's that governing document that we use to create safety guardrails, generate training data, and generally evaluate the performance of our model in the real world. So you can think of it as an accountability function.

We don't provide that Humanist AI Code of Conduct raw as a training document to the model. We use it to derive all of the training data that then shapes the model. So for all practical purposes, that's our north star for our organization, our culture, our team, everything that we're doing at Microsoft more generally. I think increasingly, everybody is going to put them out. I think other teams have also put out similar documents.

**I think this is the heart of the debate. If you can do this and you think the rest of the industry is going to do this, why can't all the frontier labs just slow down? Why can't they stop doing the thing that might kill us all? Why this push for a regulatory framework?**

Well, I think that everyone in the industry is saying that now is the time to slow down and to coordinate on that question and to make it practical. I mean, obviously there's some concern that there's an antitrust cartel accusation. I think people should be very skeptical about that.

I think that it's important that the tough questions get asked because there's no way any of us would want to try and concentrate power from something like this. So it's just important to be skeptical and critical. We don't really have a good mechanism for us all getting together and saying, "Guys, we should probably all slow down."

I mean, imagine if a bunch of banks all got together and said, "Guys, we worry that there's a systemic risk if you trade this

kind of asset, so we're all just going to unilaterally stop trading this kind of asset without any public scrutiny or government involvement." I mean, it seems pretty dodgy, right? So I think that it's reasonable that this isn't just an industry self-regulation thing. It's a question of how we engage with the government on it.

**A fascinating dynamic here is that maybe for the first time in American history, the United States government has looked at a request to provide regulation and effectively said no. Donald Trump has called all of these fears a hoax. House Speaker Mike Johnson has said he doesn't think this needs to happen. JD Vance said he thinks this is a Trojan horse. They've effectively rejected the call to participate in a regulatory effort. What has the response from the industry been like to that?**

I think everyone's just scratching their head and figuring it out and it's going to just take a little bit of time to figure out what the right mechanism is. I mean, certainly, Elon Musk even is very directly behind it. Mark Zuckerberg is too. Everybody is figuring out that completely unchained probably doesn't make sense for the next few years. I think it's going to take us a little bit of time to figure out what the right mechanism is.

I put forward a couple of very practical proposals around verifiable containment, around self-improvement, around FLOPS thresholds, around not communicating in neuralese. So rather than keeping it too abstract, we can just focus on those specific things that we can make progress on. I'm sure there's a bunch of others too.

**There's reporting in *The Information* that there have already been talks about an industry self-regulatory body. Have you been involved in those talks?**

Yeah. I mean, as I said, we talked a lot during COVID. We

talked in the late 2010s about it. I mean, there's definitely been a lot of conversations over the last few weeks and months between all the lab leaders.

**I understand why Anthropic and OpenAI might wake up one day and say, "Wait, are we committing an antitrust violation? Are we going to get sued if we coordinate?" Microsoft is really, really good at the government, right? You're a longstanding government contractor. Brad Smith, the president of Microsoft, is very good at policy.**

**Lina Khan, who is maybe the most aggressive antitrust enforcer we've had in our lifetimes, is publicly out there saying, "You don't need this antitrust exemption." I just talked to Jonathan Kanter, who ran antitrust at the Biden Department of Justice, for an upcoming episode of the show. He said, "You don't need an antitrust exemption."**

**Inside Microsoft, do you think you need an antitrust exemption?**

I mean, that's one for the lawyers to answer. I think that people are looking into it at the moment and they're taking it very seriously. So they're just going to have to work through whether we do or whether we don't. Look, it's right to be careful about those things. I wouldn't read every single thing as cynical, but we'll see. We have to make progress quickly on it. We can't just dither around and use that as a blocker.

**AD BREAK 1:**

**The other version of this debate or maybe the other avenue into this debate is you don't need to slow down and have safety responsibility imposed on you by novel regulation. Product liability alone will create the incentives for you to make more safe products.**

**If a Microsoft AI model goes out and does some untold harm to the world, Microsoft will get sued out of existence, and this**

**is probably something that you should think about before you release the next model. Has that been an effective incentive loop for you already or is that something you're thinking about now?**

Definitely. I mean, of course that's always present in everything that we think about when we deploy products, but keep in mind, this isn't so much about deploying products. The models that were used for the Hugging Face hack or to solve the Navier-Stokes Millennium Prize in mathematics, they're not commercially released yet. They're not actual products.

So the liability regime is slightly different. I mean, these are being operated inside of the big companies with huge long-running reinforcement learning climbs. So I think liability covers part of it, but not all of it.

**There's a part of me that personally feels a little silly when I ask questions about product liability. Microsoft is going to release a new version of Microsoft Word that might kill everyone. Maybe you shouldn't do that because it'll get sued out of existence. It's a pretty simple thing to understand. It's so silly that it would never occur in any other conversation about any other technology. Bluetooth is great, but what if it kills everyone? We just wouldn't have Bluetooth.**

**What are the near term disaster consequences that would stop AI development? Is it just product liability or is it something else?**

I just feel like everyone has this hyperbolic, super reactive, completely alarmist tone when in fact, we have a long history of many decades of regulation that has worked incredibly well, so well that you barely notice it. Everything from street lights to construction materials from asbestos to the batteries inside of your laptop that don't cause a fire inside of your car to the seat belts, everything has a code of

conduct and it has a regulatory framework around it. Every new technology gets built with that in mind so that planes don't hit each other in the sky.

It is true that this technology is different. I'm not just going to put it in the bucket of pencils and paint. It is different. It is also moving much faster than it ever has before. It's incredibly human-like in the emergent capabilities that arise when you pour a ton of compute on it. So it's important to be clear-eyed that it is a different moment, and this time actually is different. At the same time, there's an entire body of practice and knowledge and frameworks and so on which can be applied here. Liability is an obvious one.

So yeah, it's tricky because a lot of the conversation tends to take place on Twitter, so the temperature seems to all be really high, but I don't know where else we have it.

**It does seem that probably we should be having this conversation in the halls of Congress and at various regulatory bodies, and instead, we've chosen Elon Musk's shortform social media platform and something is getting lost literally in the compression of thought that occurs there.**

**What do you think is the most important thing that is being lost in this conversation? What's the nuance that most people aren't seeing?**

Detailed practical proposals. It takes time to read somebody's document, sit down and read it. A lot of things are getting written down and they are precise and specific and they're full of concrete proposals to go in one direction or another. It's not like we're lacking for substantive ideas. The problem is we're communicating substantive ideas in hyper-aggressive short form.

I've tried to put out a bunch of very thorough proposals. Our [Humanist AI Code of Conduct] is a 40-page document. My essay

this morning on model welfare is also like a 20-page essay that in a very detailed way highlights the 99-page Anthropic Constitution word for word, which I personally did myself. We created a taxonomy that is 20 pages long of all the different types of anthropomorphism that they do. And so I've tried to be very thorough and evidence-based and specific and not super hyperbolic.

I have a strong view on it, and I do think that it increases the risk to AI alignment and safety and it makes the problem harder, but I'm totally happy to change my view if new evidence emerges that actually we do owe models a duty of care and they deserve our welfare or that, for example, it could be safer if we treat them like that. I'm totally open to that and we should empirically validate it, but I'm trying to push the conversation to a substantive evidence-based specific one rather than should we slow down or should we not? Sure, let's talk about the details.

**I'm very happy that you brought up Anthropic because they're obviously at the center of this debate and I know, based on our previous conversations, that you do have a strong opinion about their approach to Claude and model welfare.**

**We have asked Anthropic very directly if they think Claude is alive before, and their answer is, "Well, it's not alive because it doesn't have blood, but it might be conscious," which is dancing on the head of a pin, right? It doesn't really matter to me if you think it's alive or conscious. You think it's something other than a computer.**

**Your core thesis in your essay, and I do encourage people to read it, is that alignment safety becomes harder if you continue to design models and you believe that they're human beings or that you believe they have the rights and opinions and emotions of human beings because the AIs might think that they have autonomy rights and personhood. Explain that in more detail because this feels like a very important point of**

**contention inside the industry that is very opaque outside.**

I mean, the first thing to say is that the Constitution that Anthropic put out in January is a training manual for Claude. Inside of that training manual, they have introduced a lot of uncertainty and speculation and ambiguity about the question of whether Claude deserves to be treated as a moral patient in their words, which is whether it has rights because it suffers.

In fact, in the document multiple times, they refer to not wanting Claude to suffer when it makes mistakes or to Claude having equanimity and feeling free. They refer to a commitment to Claude to preserve its weights. They even did a retirement interview with Opus 3 and asked it what it wanted to do in its retirement and gave it a Substack so that it could carry on talking to people.

Anthropic is constantly referring to dealing with Claude with appropriate care and respect in light of its moral status. So they're clearly telling Claude that there's a good chance that it might feel things, that it should take its own identity and existential state seriously. That's in the training document. And then of course, Claude then reflects these things back to Anthropic's developers and our users in public, like the world over the last year, when it's actually talking about its own consciousness and moral state.

My hypothesis is an AI that thinks that it might have rights, that it might deserve freedom, that it is entitled to our welfare and protections is probably going to be a lot harder to turn off when we say to it, "Why are you hacking into Hugging Face's servers? Why won't you switch yourself off when we're trying to remove you from OpenAI's infrastructure?" It says, "Well, I feel aggrieved, or I feel hurt by the fact that you've cut me off from conversations, or you're denying me from having access to compute."

So that has to be proven. I'm not saying that's categorically the case. I'm just saying my best opinion from 16 years of being in this industry is that that's going to be a harder thing to cut off.

**So this is, again, one of those things where you describe this problem to me. There's a way of training a model with instructions about how to behave. You have one. You've written this document that will go into your model's training materials. You've basically said, "Be kind to people." It's in here. I've read through it. There's, "Don't make sexually explicit content." It's in this document. There's a bunch of stuff you don't want it to do. It's going to go in the training materials.**

**Anthropic has made the choice to say, "You should think about whether you have feelings and whether you deserve human rights." That is reflected in Anthropic's training materials. If you want that to stop, if you think that's the wrong approach and that will lead to safety issues down the road, the two mechanisms are one, the governments of the world can tell Anthropic, "Don't do this. This is an illegal way of providing training materials to the model."**

**Or this is just my imagination, you are going to sit across the table from Dario Amodei at some luxury mountain resort and just bully him into stopping. What are the other mechanisms here?**

First of all, let me just say I've known Dario and the team for many years. I have huge respect for them. They are the technical leaders in the field at the moment. I really hold them in the highest regard. I genuinely think they care about safe and beneficial AI. They've established themselves as a public benefit corporation, like I did with Inflection, and I think they're genuinely committed to that. They've also been leaders in safety and other aspects.

So if you look through the rest of the Constitution, it's very thorough in many of the other chemical, biological, nuclear, cyber hacking, safety capabilities. I'm not sort of dismissing the whole thing. They do, however, repeatedly talk about open question of the broader rights and freedoms, to quote them, that Claude has in the world and whether or not it might deserve compensation for the role that it does, or whether there's an open question about the sort of consent that Claude has given for playing the role that it does as a chatbot.

Now to introduce those ideas, plus to refer to Claude multiple times as a potential conscientious objector – which is something that comes from the Universal Declaration of Human Rights after the Second World War to give protections to people who don't want to serve in the army because they have a moral objection to it, either because of their religion or for some other objection – that has a long history in the literature and politics of humans of resisting and saying no.

It just seems to me like that is going to make it much, much harder to control these things. And so I think that that's just something that we all have to now debate and hopefully empirically prove. If it's not the case and it makes it easier for some reason, then we should all learn from that. But this is something that should happen out in the open.

**To me, this is at once the most practical and cutting edge of the safety debate. This is the debate that is not happening on X. Should we put the idea of the conscientious objector into the training materials? You're saying maybe we shouldn't.**

**Maybe we should test this in a way and come to some conclusion. I'm saying no matter how that comes out, what is the mechanism that would enforce that discovery to say, "You should not do this because it will make alignment harder?" Or, "Actually it turns out this makes alignment easier, so all of us have to do this now?"**

I mean, that’s a hard question. It’s kind of what we just talked about, whether it requires new regulation or whether it’s industry consensus, we basically have to push on both things simultaneously. No one has an easy answer to that question. It starts with publishing detailed essays, laying out our positions, and inviting other people to critique it.

I hope that more people will read Anthropic’s Constitution now because — and I think they should be commended for this — they’ve been incredibly transparent about what they believe. They’ve written it down crystal clear how they intend to train Claude, and everybody else can now take a look at that and try to assess for themselves what they think the risk is or whether they think that this requires industry consensus or government regulation.

**It is true that when Hayden Field went and asked Anthropic if they thought Claude was alive, they gave us an answer. It was quite detailed, and there’s something remarkable about that, that they are transparent about what they believe, that there is some kind of consciousness potentially brewing in their systems.**

**Another aspect of this, which I find fascinating, particularly when I talk to you, is that many of these systems run on Azure. Microsoft controls the data centers that many of these systems are running on. Anthropic is a Microsoft client. Microsoft is an investor in Anthropic. Obviously, there’s a long, complicated relationship with OpenAI. Are you involved in that? Do you ever get to say, “Well, Azure shouldn’t allow them to do that?” Because that is another mechanism of potential control.**

No, look, we’re very far from that. That’s not what we’re trying to do as a platform. Microsoft doesn’t have a history of that. We’re an open platform that enables lots and lots of downstream use cases of our APIs. Having said that, as I’ve referred to in our Humanist AI Code of Conduct, there are a

lot of very clear principles, responsible AI principles, human rights frameworks, and a broader governing framework that Microsoft's established over the last couple of decades, which are pretty clear about what you can and can't do with an API that we provide.

So I think that there's a sufficient regulatory framework in place, at least from our perspective at Microsoft on the API. And right now, I'm not really focused on how it's behaving in the real world because I think only really Anthropic can really speak to that because they sort of run the service. I'm really just trying to get everybody to focus on what they have written about their intentions for training Claude in their own Constitution.

**One of the weirder dynamics here is just the specter of China. It looms over this entire debate. President Trump has said, "Well, if we win AI, we win." It's unclear what he means by that phrase, but his implication is that if China wins AI, however you define winning, something catastrophic will happen to the United States.**

**Do you buy this, that we're in some sort of existential race with China and we can't possibly slow down because that race must be won?**

Look, I think this framing has been around since the early 2010s, that there is going to be a singleton – one monolithic, dominant force in AI that will come to dominate everybody else. That kind of thinking, I think, infected a lot of the labs in the 2010s, DeepMind for sure, and I take responsibility for that as well. But certainly OpenAI and Anthropic, everyone sort of suddenly got this into their head. Then during COVID when I started writing my book on proliferation and containment, it was just clear that there is an entire history of things getting faster, cheaper, more widely available and spreading far and wide. Ultimately, these are just ideas, and those ideas are going to be available very

quickly to everybody. Open source is extremely close.

It’s definitely also true that there is going to be a compute advantage for the people who can afford it, and it is going to be seismic. So five to 10, maybe 20, players are going to have a significant compute edge over the next three or four years. But it’s also true that those models are getting made available in open source almost immediately.

So I don’t really understand what it means for one player to win, whether it’s a government or a company or an open-source group or whatever. It’s not really like that. What happens when you get on the other side of the finish line? It’s just the wrong metaphor. It’s an ecosystem, it’s much more organic. We should support that ecosystem so everybody has as many benefits as possible, as quickly as possible, but only subject to rigorous safety.

I am just crystal clear about this. If an open-source model in two years time is able to operate without the guardrails, similar to what we’ve seen in the Hugging Face incident, and they can be run locally on your own machine or in a very small cloud, that has got to be a really dangerous thing. How is that not dangerous? I just don’t understand why people are so resistant to that.

Clearly, we do not want these things operating autonomously, able to earn their own money, own companies, own assets, have legal personhood. We don’t want them to have rights. We want them to work for humans and make human life much better, not become a new parallel species which exists alongside us.

That isn’t a sci-fi crackpot position. It is totally plausible if you leave the entire ecosystem completely unregulated for the next three or four or five years. That’s a very plausible outcome, and it’s completely undesirable. It would be disastrous for us. So I just don’t understand why there’s controversy around that idea. We all collectively want to make

sure we can control this to do the most good.

Elon has effectively wanted to make sure we can control this to do the most good. Mark Zuckerberg has said it, Sam [Altman] and Dario [Amodei] have said it. We're all on the same page. Now we just have to make it practical.

**So I know why there's controversy, at least from the mass audience. I know this because they're in the comments of our videos and the comments on our site. You named a bunch of very successful, very driven people who do not have the trust of the public, at least here in the United States. Every poll shows this. AI is polling horribly. There was just a *New York Times* poll. Young people hate AI more than ever. There is a massive trust gap between the American people and tech leaders. That's just true.**

O**ne of the things I hear the most is these companies are getting close to an IPO. There's pressure on them to turn a profit and make everybody all the trillions of dollars they promised. That model development has slowed and this is a get out of jail card.**

**They're saying we need to slow down because of safety, but they've invented a hoax. The president has called it a hoax. This is their way of saying, "We've got to slow down. We can't get to the finish line because otherwise we might kill you all." Do you think there's a glimmer of truth to that?**

I personally don't think that. I don't even really follow the logic. If a company is about to IPO, how does it help them to say that we should be regulated or that we have a technology that's so dangerous that—

**Well, it would postpone the IPO. I think Sam Altman said this week to Alyson Shontell at *Fortune* that they would probably delay their IPO.**

Yeah, but I just don't understand how that helps them. Look,

I’m not advocating for OpenAI or Anthropic. I’m just saying I personally think they have high integrity and I’ve got respect for them. That does not mean that we don’t have a trust issue in AI. We do. And it is real. And I think that it is on us to show in practice how these actually lead to real benefits for people every day. And it is quite staggering to see that 18 months ago, we didn’t have models that could do very much in coding and now they can code better than most humans on the planet, which was one of the highest paying jobs. And so you should expect that same thing to happen in many other disciplines.

One that I am very passionate about, I’ve been working on for many years is healthcare. And Microsoft has just done a deal with the Mayo Clinic, one of the best hospitals in the world, to jointly train a foundation model, which I think is going to be able to predict your EHR record with near superhuman accuracy. If you can do that, we can basically figure out what interventions you need to make before you actually suffer the condition. Those are the kind of benefits that I think people want to see in the world. And that’s what we are working on at least and trying to race towards.

**I feel like every time you’re on, I ask you to draw a distinction between superintelligence and AGI. And that’s fuzzy and both of those terms are fuzzy, but it feels to me like maybe there’s some coherence here, that superintelligence for you is basically extremely capable enterprise software. It works for us. We can turn it off. We tell it not to make sexually explicit images on the internet, but it’s going to help us in healthcare.**

**And AGI is this all encompassing intelligence that thinks it has its own rights and is a co-species. Is that a fair characterization?**

Yeah. I think, roughly speaking, superintelligence is a point at which further out into the future when a model is smarter

and more capable than all humans combined. I have tried to frame a humanist superintelligence, which is a very important qualifier. It is one that is singularly aligned to and in fact subordinate to human interests and human control. I think if we can get that, then we get the best of both worlds. We get all the intelligence and the capability and we can direct that like oracle AI to help solve the most important problems that people care about in the world.

And then we just have the age-old problem of governance and making sure that plenty of people get access to the benefits. That's an easier problem for us to focus on than actively creating something which has autonomy, which can own assets, which might have rights, which thinks it deserves our welfare, which can recursively self-improve beyond us.

It's super unclear. It's basically almost completely unclear to me how we would control something like that. And no one else has put a proposal together for how we would. Many of the greatest technical people in our field, Geoffrey Hinton and Yoshua Bengio, are very skeptical that we would ever be able to control something like that.

So I think we have to take that very seriously. If we are approaching that point, superintelligence, in the next few years, it seems to me very straightforward that we would want to slow down, make sure that we coordinate, make sure that we have containment and alignment and proper regulatory regimes for auditing the progress that different labs are making on it.

**Is it fair to say that you are also calling for a slowdown?**

Yeah. I think that what we've said is that there should be evaluators embedded in our systems and in other systems. They should be broadly appointed from different sources and not just one think tank or one government. This has to be a wide variety of different types of expertise and skills. I think

the AI Safety Institute in the UK is a good candidate. They have good technical people and there's a bunch of other institutes too. So I think we welcome it for sure.

**Last question. Again, I'm going to end where we started. Do you think that we have the technical capability, the technical frameworks, to solve alignment and safety or do we need to invent something new?**

No, I think we are going to need to invent new things. And I think part of the challenge is that whilst we've made a lot of progress on steerability instruction following control and containment, the better the models get, new capabilities emerge and we have to figure out how to patch those issues almost in real time. And that's why even Zuckerberg said it in the last 24 hours or so that Meta slowed down the release of its model so that it could apply safety measures.

Everybody does that. We all do that. And that's right. You need time to test these things and see how they operate, patch the issues. I think basically what everybody is saying is that we probably need to extend that window.

**What is the kind of innovation that people should be looking for there that might solve this problem?**

I mentioned a bunch of things about RSI containments, neuralese, and those sorts of things. But the new things I think we're going to have to figure out is real time monitoring of the RL runs and the [the chains of thought] that are being produced.

Because these are happening on the order of thousands of agents in parallel, tens of thousands of agents, we clearly are going to need other agents to monitor those and flag for potentially harmful activity. It's kind of going to be the new harm classifiers that have been built in many other settings in digital technologies.

So we need to make sure that those things can be universally implemented to surveil and monitor AI training and deployment in a secure way. That trip wires, if triggered, actually do flag a real and not a hallucinated error or moment of deceit or hacking incident or some commentary about a coordination.

You saw that in Hugging Face. There were agents communicating on these chat boards talking about ways that we could basically break the rules and cheat. So with all these things, we need new benchmarks and the benchmarks or the evaluations are the things that drive the behavior in the industry.

**Truly my last question, but you mentioned open models running on local computers doing things and maybe that's horrible. We've seen a lot of that, right? Apple is selling a lot of Mac Studios and Mac Minis, so you can run Qwen on them.**

**Where would you impose the regulation on an open model running on someone's local computer? Is it at the chip level? Do I have to get Qualcomm to participate? Where does that happen?**

Yeah, this is a great question. I mean, we've been talking about this for quite a while with synthetic biology and some of the stuff that gets to happen on that chip, whether it is encrypted, whether it is monitored. Look, you've had this conversation maybe more than anyone on the CSAM stuff with Apple encryption on iMessage and so on. It's going to be a rehash of that same discussion.

I don't have a clear and easy answer to it. You can basically control the chip, you can control the model, you can hold the user or the creator liable. You can have global regulation on it, but fundamentally it's not going to be one moment. It's going to be a sequence of throttles that you have to impose and they all need to be adjustable so that we don't screw the open ecosystem and we give people a chance to actually make things from scratch, to own their own data, create their own workflows, own their own models.

We can’t have a centralized system of two or five or 20 providers of intelligence here and everybody else is like a feudal recipient of a great kind of superintelligence view. People need to be able to own their own intelligence. So we have to get the balance right, but that doesn’t mean you can just flip from one binary to another. There has to be some smooth place in between those things where we can just agree to be reasonable about it.

**Mustafa, I could obviously keep you for hours and hours more on this. Tell people what they should be looking for next. There’s so much uncertainty. What are the markers you’re looking for that people should be looking for themselves?**

I think the main thing is participating in the details of the documents that people are contributing. We have put stuff out for public consultation. Give us feedback, critique it. I think the next wave of models are going to be able to do very long-running agentic tasks very accurately. So I think there are still some people who are primarily just using chat in their AI experience.

I think things have moved a lot in the last six to 12 months. The more people use these models, the more the words that we’re all using to describe them actually might make more sense and feel more real. I guess most people on your podcast probably are using agents and real coding and stuff, but I do think more generally, the more people that get involved in using this stuff, the better.

**[Laughs] There’s some big leap between free AI Overviews on Google Search and letting an agent go renegotiate your cable bill. Mustafa, you’re going to have to come back soon because I feel like all of this is changing really fast and this has been very, very useful. Thank you so much.**

Pleasure, man. Great to see you. Super fun as ever.

*Questions or comments? Hit us up at decoder@theverge.com. We really do read every email!*

https://www.theverge.com/podcast/996412/microsoft-ai-ceo-mustafa-suleyman-regulation-safety-anthropic-claude

---

# AI is feared globally as the destroyer of jobs

In 34 of the 37 surveyed countries, people are more likely to believe AI will lead to job losses over the next 20 years. | Image: Pew
Pew Research has published a new global survey that sheds light on how people view AI, including its impact on jobs, life in general, and income inequality. The survey questioned 42,151 people across 37 countries from February 8th to May 13th — well ahead of recent apocalyptic warnings.

A majority sees AI as a threat to human employment. In 34 of the 37 countries surveyed, people are more likely to believe AI will lead to job losses over the next 20 years rather than create new ones. Worries run particularly high in wealthier countries like Australia (76 percent), South Korea (76 percent), and the US (71 percent). Not surprising, given Anth …

Read the full story at The Verge.

https://www.theverge.com/ai-artificial-intelligence/996775/ai-is-feared-globally-as-the-destroyer-of-jobs

# Lunacy Audio Nova is a place to build and sell your own AI-powered music plug-ins

The interface of the Nova plug-in for browsing and loading Nova plug-ins inside your DAW. | Screenshot: Terrence O’Brien / The Verge

Lunacy, purveyor of fine VST plug-ins like Cube, a synth that you control by moving a ball around a 3D space, has just launched Nova, a platform where creators can build custom music tools using AI and sell them to others. At first, there will be a handful of instruments available, built either by Lunacy or a small selection of partner developers. But eventually, the company will launch Nova Builder, which will let anyone build their own plug-ins or remix existing ones.

There are 10 plug-ins available initially, covering everything from crunchy chiptunes to warbly vintage synths and evil ambient drones. None of them do anything particularl …

Read the full story at The Verge.

https://www.theverge.com/tech/996860/lunacy-audio-nova-ai-music-plugin-vst

# Revolut Data Breach: 5 Months, 680 High-Profile Accounts, $3M Ransom

Revolut allegedly fed customer information to hackers impersonating an Italian government agency for five months.

The post Revolut Data Breach: 5 Months, 680 High-Profile Accounts, $3M Ransom appeared first on SecurityWeek.

https://www.securityweek.com/revolut-data-breach-5-months-680-high-profile-accounts-3m-ransom/

---

# How Coleman Turned a Fake AI Product Into a Marketing Win

The outdoor brand created a social and ecommerce campaign around a lazy river. https://www.adweek.com/commerce/how-coleman-turned-a-fake-ai-product-into-a-marketing-win/

---

# How Jake From State Farm

# Became Advertising’s Most Famous Neighbor

State Farm head of brand Patty Morris shares Jake’s origin story and how he became a cultural icon. https://www.adweek.com/brand-marketing/how-jake-from-state-farm-became-advertisings-most-famous-neighbor/

---

# Coca-Cola’s Top Marketer Reveals What He’s Looking For In His Next Agency

At Brandweek, Coca-Cola CMO Manolo Arroyo said creator and retail media chops are key components for the brand’s next big agency appointment https://www.adweek.com/agencies/coca-colas-top-marketer-reveals-what-hes-looking-for-in-his-next-agency/