Technically Optimistic

An Emerson Collective Podcast
Who has the power to regulate AI? In **Episode 2, Part 1 of *Technically Optimistic*,** Raffi Krikorian, Emerson Collective's chief technology officer speaks with Tristan Harris, founder of the Center for Humane Technology; human rights activist Sam Gregory, founder of Witness; Suresh Venkatasubramanian, professor of computer science; Meredith Broussard, author and data scientist about what's at risk and when we consider the balance of society, technology, and the law.

***

RAFFI:
I'm Raffi Krikorian, and this is Technically Optimistic. It's episode 2 in our miniseries on artificial intelligence.

> **CLIP** [Altman testimony beginnings]
> BLUMENTHAL: *Welcome to the hearing of the Privacy, Technology, and the Law subcommittee….*

RAFFI:
On Tuesday, May 16th, 2023, Sam Altman, the 38-year-old CEO of OpenAI — the organization behind ChatGPT and DALL-E — appeared before a US Senate subcommittee and fielded questions from lawmakers.

> **CLIP**
> BLUMENTHAL: *Please rise and raise your right hand. Mr. Altman, we're going to begin with you, if that's okay.*
> ALTMAN: *Thank you. Thank you Chairman Blumenthal…*

RAFFI:
His testimony touched on a number of concerns people have about artificial intelligence. … Like, its potential impact on the economy.

> **CLIP**
> ALTMAN: *So, there will be an impact on jobs. We try to be very clear about that, and I think it will require partnership between the industry and government…*

RAFFI:
Or AI's potential role in spreading disinformation– particularly around elections

> **CLIP**
> ALTMAN: *We are quite concerned about the impact this can have on elections. I think this is an area where hopefully the entire industry and the government can work together…*

RAFFI:
And on the outsized role that big companies seem to play in shaping the AI space.

> **CLIP**
> ALTMAN: *I think there is benefits and danger to that; the fewer of us that you really have to keep a careful eye on on the absolute bleeding edge of technology, there's benefits there.*

RAFFI:
Altman … came across as… serious, I guess? More sincere than Zuckerberg, when he was before Congress talking about social media in 2018.

On the surface at least, Altman seemed interested in working with Congress… to put, quote un-quote, guardrails in place.

> **CLIP** [guardrails montage]
> VOICES: *The right guardrails… provide the kind of guardrails… creating reasonable guardrails… guardrails or safeguards or … responsible guardrails … the guardrails are definitely needed.*

RAFFI:
This term "guardrails" can mean different things to different people. It could be something self-imposed… a rule you set for yourself, like "I'm only gonna have one cookie tonight"… or it could be something imposed upon you. – like the actual guardrails on the sides of a highway.

But really, in this case, it's definitely a stand in for a different word … and that word? Regulation.

That idea can get a lot of people on edge, especially in tech … whether it's because of their politics, or their belief that industry can police itself. Silicon Valley has become one of the biggest economic drivers for the US, in a lot of ways, because of the hands off approach by Congress.

But….. others might be wary of regulation because of cynicism… and I get that. When it comes to tech, lawmakers can sometimes seem… dangerously out of touch. And who are they to regulate something they don't even understand?

> **CLIP**
> SEN. KENNEDY: *Um, would you be qualified to uh, to uh, if we promulgated those rules to administer those rules?*
> ALTMAN: *I love my current job.*
> SEN. KENNEDY: *You make a lot of money, do ya?*

RAFFI:

And look, Sam Altman is just one voice here. But the questions his testimony raised are crucial for all of us to think about:

How would we even go about putting up these “guardrails?” And should we, if it stifles innovation?

There seems to be a tradeoff between how fast this is all coming out… and making sure it’s safe. And maybe that shouldn’t be a tradeoff… but it seems to be one that we’re living with right now.

How do we thread the needle? How do we make sure our society can reap the rewards of AI, and not be saddled with unwanted consequences?

Can we ensure that there’s oversight. and transparency… without losing our standing in the global economy… or compromising our national security?

Basically, how do we define and figure out, what’s for the public good?

CLIP
KEVIN McCARTHY: *Hi. This is Kevin McCarthy.*

RAFFI:
The Altman hearings seemed to wake up lawmakers on both sides of the aisle.. to the urgency of this issue.

CLIP
KEVIN McCARTHY: *And I’m pleased to have the opportunity to say a few words in support of the National Security Commission on artificial intelligence.*
CHUCK SCHUMER: *As you may have heard, today I gave a speech about one of the most pressing and consequential issues and innovations of our time and that is AI.*

RAFFI:
A little more than a month after the Altman hearings, Senate Majority Leader Chuck Schumer gave a speech introducing what he called a “comprehensive framework” for regulating AI.

CLIP
CHUCK SCHUMER: *I call it the Safe Innovation Framework for AI.*

RAFFI:
And you can really hear him wrestling with these tensions.

CLIP

CHUCK SCHUMER: *I call it that because innovation must be our north star. We must come up with a plan that encourages, not stifles, innovation. If people don't think innovation can be done safely, it will stifle AI's development and prevent us from moving forward.*

RAFFI:
There is a LOT to cover here. So, in this two-part episode, we're gonna talk all about regulating AI. In this episode, part one,, we'll go over some features and risks that are leading many people to call for reforms. And next week, in part two, we'll hear from experts — including a couple of sitting US lawmakers — who are actively thinking about how to set the quote-unquote guardrails.

It's a thorny subject, but it's important... because it's not just Chuck Schumer and Sam Altman who think there should be AI regulations. Based on some very recent polling, more than three quarters of American adults say: we need AI regulations and laws.

But there might be a fundamental incompatibility here. Government, especially the U.S. Congress — is really designed to move slowly and deliberatively. And in the tech world, you know... they like to move fast... and, famously, break things.

So can it be done? Can we mitigate AI risks without stopping AI progress?

Some say no... but we are Technically Optimistic.

Music

RAFFI:
Let's actually back up a little bit. Before we talk about the challenges of regulation – we have to ask... Why does AI need to be regulated at all?

Some people think there should be no regulation, but, others may point out tons of reasons to regulate. Let's see if we can focus on four big ones.

Risk number one... the race to deployment.

TRISTAN HARRIS:
it really snuck up on me by surprise, the impact of these new large language model AIs. And once we started digging into it . . . what we found was pretty consistent, which is that the race to deploy AI systems as fast as possible was setting a clock rate for the industry in which there would be no way that we could make safe decisions.

RAFFI:

Tristan Harris is a technology ethicist, and the co-founder and executive director of the Center for Humane Technology. You might remember him from *The Social Dilemma*, the Netflix film about the dangers of social media.

TRISTAN HARRIS:
I get criticized for being a fearmonger or a doomsayer or a pessimist. And I want to really say to you that what I've noticed especially in my work on social media, I had conversations with people about the things that were gonna happen with social media back in 2013. These were friends of mine from Stanford, my friends who co-founded Instagram, I knew people who worked on Facebook News Feed and the profile page and all that, back in 2013. And I tried having conversations with them because I wanted to see this stuff changed. And I didn't push hard enough. They just didn't want to think about the risks. It was just not fun to think about these problems.

RAFFI:
In March of 2023, Tristan, along with his partner Aza Raskin, gave a public talk on the risks of AI. They called it the A.I. Dilemma, deliberately trying to frame their concerns about artificial intelligence as a continuation of their work on social media.

TRISTAN HARRIS:
And it was in our experience with social media, which we call first contact with AI, that gives us some unique credibility. And so we sort of sprung into action.

RAFFI:
I was in the audience when Tristan came to San Francisco... At the time, the talk felt to me.. alarmist

> **CLIP** [from the AI Dilemma]:
> [applause]
> TRISTAN HARRIS: *Automated exploitation of code, and cyberweapons. Exponential blackmail and revenge porn. Automated fake religions that I can target the extremists in your population and give you automated perfectly personalized narratives. Uh, exponential scams, reality collapse. . . These are the kind of things that come from if you just deploy these capacities and these capabilities directly into society.*

RAFFI:
But, Tristan, instead, proved to maybe be ahead of the curve. Less than a week later, OpenAI's GPT-4 was launched, and catapulted many of the issues he was talking about into the mainstream.

TRISTAN HARRIS:
We interviewed probably a hundred people between February and March and put together, these are a hundred people inside the companies. People who used to work at the White

House, National Security Council, people who worked at Rand, people who worked on AI safety. . . . And we did briefings in New York, DC, and San Francisco and tried to get philanthropists, government policymakers, institutions, media, all educated about what was going to be coming: The kind of structural risk of what happens when you race to entangle a new technology with society faster than any technology's ever been entangled. and one that's more consequential and more powerful than the other technology that we've had.

RAFFI:
Can you tell us why regulation is important? Like, it's one thing to point out the potential risks of AI, but can you say why you think that we need to actually legislate limits onto the field?

TRISTAN HARRIS:
Yeah, well, maybe one thing that might be helpful is introduce the three laws of technology that we actually use in our AI dilemma talk,

Rule number one. When you create a new technology or new kind of power, really, you invent a new class of responsibilities. What does it mean to responsibly influence people's emotions? So that's the first rule.

The second rule of technology is if that new technology confers power, meaning if people have an advantage over people who don't tune technology to influence people's emotions, then it's going to start a race.

The third rule of technology is if you do not coordinate that race, that race will end in tragedy. So that's a tragedy of the commons problem, or an arms race, or race to the cliff.

You know, we can reason about how we would want technology to be ethically deployed in some abstract sense. But if it exists in a container where there's a bunch of people who are maybe not have the morals that you do and use it amorally and ruthlessly to kind of maximize their power, like Cambridge Analytica did to influence people's emotions at scale and personalize an A-B test, a thousand variations of each message perfectly tuned for their audience, and that starts to confer power...

> **CLIP:**
> *In 2014, you may have taken a quiz online. And if you did, you probably shared your personal data, and your friends' personal data, with the company who worked for Donald Trump's 2016 campaign: Cambridge Analytica.*

TRISTAN HARRIS:
They're going to win. And if I'm the other party and I say, I realize that's bad for democracy. but if I don't do it, I'm just going to lose to the guy that will. And so now we end up with the race that ends in tragedy. And the tragedy is that now we don't have a shared reality because we

were so used to maximizing the emotional engagement of each person.

RAFFI:
How do you think about in that framework those developers who are being explicitly malicious and those who are just simply not thinking about it? What you describe seem like malicious actors. . . .

TRISTAN HARRIS:
Not necessarily.

RAFFI:
. . . versus innocently stupid actors, maybe?

TRISTAN HARRIS:
Well, there's a fine line between deliberate and malicious use and motivated reasoning that justifies the incentives that I already have.

RAFFI:
[LAFF]

TRISTAN HARRIS:
So if I'm you know doing my political party for the greater good because I think I have the right vision for the future and I have an optimistic good vision for the future as so far as I am concerned, motivated reasoning based on the way I see the world, then I'm going to use whatever the tools that are available to me to do the thing that I think is good for the world, operating ethically according to a subjective view of what I think is good for the world.

That person, if they start to use emotionally tuned perfectly designed persuasive messaging, is going to outcompete people who don't. So that's not necessarily a malicious actor. And in general, I think most of the problems of the world are incentives. Whoever wins the race within a narrowly defined incentive of like engagement or GDP or even reducing CO2. CO2 is one very narrow metric for how to heal the environment. There's biodiversity, there's ocean acidification, there's a bunch of other things. we need to have a more comprehensive sense of what we're optimizing for, where if I don't do the thing for that narrowly defined metric of good, even if adding GDP causes climate change, or even if winning the engagement race causes a mass polarization of society, I lose to the guy that will.

And it's really these narrowly defined goals that are driving, I think, most of our problems, issues, environmental issues, or tech issues. If I don't race to deploy and get as many people on my AI platform and getting people using my API, I'm going to lose to the companies that do. So, that's really the root of our problems everywhere. It's not evil CEOs. It's not evil corporations. It's a race that they themselves are caught in. And I think that's really true

everywhere that we look. And it's kind of humanity's rite of passage to correctly identify not bad guys, but bad games. the narrowly defined metric.

RAFFI:
I feel like you just defined Skynet for me in some way. Just sort of like narrowly phrased goals, which then says, oh, we should just kill humans because that would be the easiest way to get there.

TRISTAN HARRIS:
Exactly. Well, if I want to cure cancer, what's the fastest way to do that? The classic example is you just kill all the humans because then there's no more cancer and I was the fastest route there. Or Stuart Russell's example of Robot, get me some coffee. And it literally is a robot and it just like runs over your dog, kills them, runs through the wall, destroys the wall, gets the coffee, runs back through the wall, kills your grandmother on the way to give you the coffee. Because a narrowly defined goal in any sense that's not seeing the externalities. is always going to lead to these kinds of problems. And I would say that's really the root of all of our problems are narrowly defined goals that don't see the externalities that are associated with that.

RAFFI:

The problem of "narrowly defined goals" is a real one. And it doesn't have to be as extreme as the example Tristan brought up.

The fact that AIs have to be trained to maximize certain values – what Tristan calls the "narrowly defined metric"-- almost guarantees that there will be unforeseen consequences.

So, potentially, maybe we can prevent some of those bad consequences from occurring... if we require developers to be very, very careful in what they release.

But, I'm getting ahead of myself. We'll get to those potential solutions soon. There's still more risks to talk about.

Like... risk number 2.

SAM GREGORY:
If you can say anything can be faked, it's really easy to throw out that claim against real video that shows the truth.

RAFFI:
The loss of truth.

SAM GREGORY:

Right from the beginning, we came from a position we described as prepare, don't panic, because we saw the harms of this deepfakes hype where it was already being used even five years ago to undermine the accounts of frontline activists and journalists.

RAFFI:
Sam Gregory is a human rights advocate and the executive director of an organization called Witness

SAM GREGORY:
Witness is a global human rights network. We have a team split across 12 countries worldwide and we help people use video and technology to defend human rights. And we work very closely with people who are, for example, taking out their cell phone to film in a protest or documenting war crimes in Ukraine. We make sure that people in similar situations have access to the best ways to do everything from document war crimes, to show the true stories of abuses of land rights. challenges that make it harder for those individuals and communities to have their videos seen and trusted. And that's how we got involved in the areas of deepfakes and the questions of how you believe the videos you see online.

RAFFI:
Deepfakes are AI-generated videos that have been digitally altered, and are designed to spread false information. As AI and machine learning technology advances, these videos are getting harder to detect, and so can fool more people into thinking that they're real.

SAM GREGORY:
So we live in a moment of absolute contradiction for an organization like Witness that cares about how you trust and believe in many more people sharing video accounts of really important civic issues. On the one hand, there's many more people who are creating videos, sharing them, using social media, using them for real difference, right? To prove war crimes or to show the evidence of their state doing wrongdoing. At the same time, we have this undermining happening of our belief in those videos, right? We're sort of challenging something that maybe for 50 or 60 years, we've seen — as one scholar described — as the epistemic backstop, this idea that you can always turn to the video ultimately, and you might interrogate it, but it's going to show you what's happened. . .

Raffi:
Are we dealing with sense of scale at this point, that these tools are now becoming so easy to access? People talk about democratization when it comes to these tools, but in some ways that democratization is to your detriment.

SAM GREGORY:
So I think we live in a contradiction as a human rights organization that's basically focused on this expanded access to audiovisual technologies because essentially, every five or six years for the two decades of this century, you've had a fairly significant shift that reshapes

the contours there, but doesn't necessarily reshape it with frontline activists, frontline journalists in mind, right?

So think about the expansion of the ability to share videos online in 2000, or let's go to the iPhone, let's go to broad-based social media and even live streaming in the first part of this decade. And so as an organization, we've always had to grapple with what is kind of a contradiction between the speed and sort of these flip moments on technological change that expand access to the human rights movement and create possibilities and threats at the same time as the underlying human rights issues haven't shifted that much.

At Witness, we talk about the idea that you have to fortify the truth. How do we make it more likely that we'll still be able to believe these videos and accounts that come from the journalists and human rights defenders on the front lines? And part of that means pushing back on, you know, the technologies that are making it harder to do that.

RAFFI:
To take one recent example, you might have seen an image a few months ago, of Pope Francis wearing a puffy, white, fashion-forward jacket.

> **CLIP**
> *It's the photo sweeping social media, that's fooling everyone.*

RAFFI:
At first glance, it looks like a photograph in every way.

> **CLIP**
> *Many are only learning now... that it's a fake.*

RAFFI:
And though it might have caused some momentary cognitive dissonance to see the head of the Catholic Church in such... fly streetwear, many of us probably had just enough doubt about what goes on in the Vatican that we said... "Huh. Sure."

> **CLIP**
> *TV personality Chrissy Teigen is among the many who were duped. I thought the Pope's puffer jacket was real, she tweeted. No way am I surviving the future of technology.*

RAFFI:
And that's sort of the problem. For most of us, even those of us who know about deepfakes... this image seemed convincing enough for us to question our assumptions, and accept it as true. So for Sam, this kind of created material is a threat.

RAFFI:

What do you think, or what does Witness think that it might actually look like, to push back on a meaningful way on AI?

SAM GREGORY:
I think it's a really important question because I think at the moment, a lot of the discussion is right, you know, is about sort of the individual interaction between you or I seeing like the pope in a puffer jacket or not in a Twitter feed or something. And it's like, how do we believe that? Did we get fooled? And I think that is one risk. And you see it in our work where trust breaks down on much more critical pieces of video, right? or media in general.

We run a rapid escalation task force for manipulated media and deep fakes and the last three examples have been audio. And in each case, in one case maybe it was fake, in one case we couldn't tell, in one case it was authentic, but in all of them there was a sort of climate of you can, you can't tell, and it plays into people who wanna claim that you can't believe things are true. It plays into people who have a vested interest in undermining all trust, because that of course reinforces real power, say state power. And it also just reinforces kind of the sort of the confirmation bias that we increasingly encounter and want to feel in terms of our social media ecosystem. So I think we have to look at trust very holistically and we have to think about not placing the blame on breakdown of trust or breakdown of belief in what we encounter right at the individual, because it's very hard once you get to you or I encountering say AI generated media in our timeline to place the blame there.

One of the things we heard a lot in the consultations we've run, it's been at the center of our advocacy to platforms and the people building technical infrastructure on this is we've got to have a pipeline responsibility here. If we're going to do anything about this, it can't be like blame the consumer and it certainly can't be blame the journalists and civil society activists and local election officials who are on the front lines of this because they don't have the skills, they don't have the capacity, the tech infrastructure isn't built for them. And I think too much of the discussion still pushes responsibility towards individuals and this very fragile layer that is already under stress of the media and journalists and local government officials who, as we know in the US, are often right at the brunt of what's happening right now.

RAFFI:
So... I hear you say that blame should not be placed on the consumer who gets fooled by AI-generated misinformation, and of course I think that's important. But, where should the blame go? Should it go to governments? Should it go to platforms? To technology creators? Sam, where should it go?

SAM GREGORY:
I think I'm gonna use a different word than blame. I'm gonna say responsibility because I think we're still learning exactly what it looks like. And we still haven't seen widespread synthetic media used in widespread deceptive ways yet, right? So, but you can see that sort of the contours of it, let's put it that way...

Raffi:
And that “yet” phrase...

SAM GREGORY:
Yeah, the “yet,” like, and I'm always trying to avoid like the sense, because as I say, it really perpetuates this attack on truth is when you're able to say everything's a deep fake when it's not.

So we have to be really careful still to assess our claims. I think there's a responsibility here that lies through that pipeline. So let's take the example of like, you or me encountering the pope in a puffer jacket or a deceptive image. These are easy ones, right? They're super easy, they're not very consequential. So we might say, for example, if we have an expectation that we're entering this really complex media generation world where our world is giving a mix of synthesis and real, personalized content and non-personalized. We've got to back up all the way to, you know, everything from the foundational models to the people who kind of act as the intermediaries for how those models are used in apps into the social media platforms. And we've got to say there's got to be some really core principles that are built across this, right? Principles around consent, which we could understand really broadly from everything from art being fed in as training data to who’s the subject of the deep fakes to how you opt out of that. We need to ask about disclosure, which is how you show what happened, right? So how do you make it clear that something, maybe it’s AI generated, but also how it moves through time and gets changed, because I think media is not static, it's not AI or not. It gets edited, it gets shared, it gets re-edited, it gets redacted.

So we need that disclosure built through the chain, because otherwise it's kind of hopeless. Imagine like if you get like an AI image and we're asking the average person to put a label on something that says, made with AI, then someone edited it, then I shared it on Facebook. It's ludicrous, right? But there are ways you can think about that from right from the beginning, from everything literally like do you, you know, create radioactive training data in a foundational model, do you watermark it, do you have standards that are privacy protecting that carry that information through the social media system until you encounter it?

Now that's a good example of like where responsibility could lie on say media integrity, like the authenticity of media. And it doesn't mean that like there aren't around that there aren’t bad actors in that ecosystem who try and get around that and fake it and delete it and break it. It just means that for the vast majority of the ways we're gonna interact in this ecosystem, we have a far stronger sense of, you know, where this AI is being used, is it being used to deceive us? And probably then also system-wide accountability, because you have a much better sense of where the problems are emerging across a system, rather than like at the symptom at the end, right? Like here's just a piece of media that I've encountered that deceived me. You can start to think, well, actually, we're really having a problem with how

foundational models handle this because we're seeing all these impacts further down the line.

RAFFI:

So, from what Sam is saying, we can glimpse a particularly grim possible future... one where the entire information ecosystem becomes effectively useless, totally diluted by AI-generated disinfo.

If it seems bleak– it's not inevitable. This is a call for all of us to reimagine accountability and authenticity. And regulation could help with that. So that's risk number 2.

But even this risk is about preventing a worst-case scenario in the future. There are other reasons to regulate that have to do with the way AI is behaving *right now.*

We'll get into that... after a quick break.

Music

MIDROLL

Music

RAFFI:
Welcome back to Technically Optimistic. We are talking about regulating AI.

We've already talked about two catalysts for AI regulation... the race to deployment, and the loss of the truth.

But there's another factor that has led many people to call for new rules. And that's risk number 3: Disparate impact.

SURESH:
I was, for the longest time, what I describe as a very unwoke computer scientist.

RAFFI:
Suresh Venkatasubramanian is a professor of computer science at Brown... and he's one of the leaders in research at the intersection of artificial intelligence and ethics. We heard from him a bit in our last episode.

SURESH:
I think of computer science in terms of geometry and high-dimensional spaces, and I was living in that world, and living in that world brought me into machine learning because

machine learning at its core speaks the language of high-dimensional geometry, right? It sounds very esoteric, but there you have it.

I got tenure at the University of Utah and said, okay, now what do I do? Right? I want to do something, not the same stuff I've been doing for all these years. I could, but that's not interesting to me anymore. I wanna look for something new. And I started trying to prognosticate all is a dangerous thing to do. Right. I said, okay, what happens if we are, you know, if we're in a world where everything we do is controlled by AI, we're gonna want to know more than just how to do better AI. We're gonna wanna know whether these systems are working for us.

But I also went and gave a talk at my colleague's university at Haverford College and Haverford College, being a liberal arts school, we had a dinner with a bunch of faculty from different departments. And I was talking to a sociologist who had actually was telling me, and I don't know how it came up, about the Griggs versus Duke Power case that led to the doctrine of disparate impact in the 1970s.

Right? The case that went to the Supreme Court where, um, uh, a power company was being sued for a particular test they were giving out for promotion that had the effect, even though it was facially neutral, had the effect of disqualifying black applicants from getting promotion. And the Supreme Court at the time said that even though the, the rules for who is eligible for this were facially neutral, they had a disparate impact. And that was a problem.

RAFFI: The Supreme Court ruled that, even though the Duke Power Company wasn't implementing an explicitly discriminatory rule, it had the effect of discriminating against Black workers, who were less likely to score well on the aptitude tests that were required for a promotion.

SURESH:
And it didn't hit me immediately, but it hit me fairly soon after that, of course, civil rights lawyers had known this for a long time, but I only learned about this, that there was a way to talk about discrimination using the language of impact and not the language of intent only.

Then that's something where now we could be thinking about what it meant for algorithms to have disparate impact because an algorithm doesn't have intent, but it could have an outcome that was disparate in impact. And so the first paper I wrote with my colleagues was basically on that. What would it mean to think about the disparate impact in the language of machine learning?

RAFFI: This is what Suresh calls algorithmic discrimination.

SURESH:

That's where I started. And then one thing led to another, and here I am. [LAFF]

RAFFI: Maybe we could dive into that phrase, impact versus intent. Because I think that a lot of people who are gonna listen to this don't fully know how to unpack that. Like these algorithms don't have the intent to be discriminatory, but they're clearly having an impact on the backend. Like how do you even reason through that this is actually what's happening in the world?

SURESH: So it's actually very tricky, I think, because the mere fact of an impact is something that is contested. So what do I mean by this? Suppose you have a test, one of the canonical examples is, suppose you have a test for being a firefighter. And let's say one of those tests involves lifting heavy sort of objects above your head. Okay?

So for a test that involves lifting something above your head in a way that uses your chest muscles, women are likely on average to be able to lift a lower weight than men would, so a test that says that you should lift a certain amount of weight above your head in so many seconds is going to have a disparate impact outcome, at least for men and women.

But that does not mean that this is a problem necessarily, right? And that's the key, right? The fact that there is a disparate outcome does not necessarily mean that this is discriminatory, and in fact, in disparate impact doctrine, that's only the first step in a multi-stage process to determine if discrimination has occurred. But it's the first sign of evidence that there could be a problem. And so that if you think about, okay, there is some kind of measurement of disparity as the first sign of a problem, and we first have to agree or discuss what those different measurements of disparity could be.

And then once we have that measurement of disparity, we have to then think about whether that represents an issue that we need to mitigate or not. So when I talk to my students, right, we can unpack this notion of how you measure disparities. And there are different ways to measure.

One measure of disparity is, like I said, you look at the, you know, success rates for one group versus another group and see if they're different, right? So if you are trying to, let's say, do hire people for a job, and it turns out that, again, staying with the sort of a, sort of a binary gender-based example, which is not perfect. I know.

If of all the men who applied 20% got the job, and of all the women who applied 5% got the job, then you have 20 and you have five, and that five over 20 ratio is one out of four. And that's a small number. It's very far away from one, if it was 20 and 20. And so maybe there's a problem, and that's the one way to measure a bias.

Or you might say, and this comes up a lot, well, maybe just that measurement itself is not indicative because maybe this task involves lifting heavy pieces of machinery. And it is natural that more men than women would succeed at the job. What we really care about is whether the algorithm that's deciding who should be hired makes mistakes at the same rate.

RAFFI:
You don't have to look very hard to see real-life instances of disparate impact.

> **CLIPS**
> VOICE 1: *Studies show this tech works well on white men, but is less accurate for other groups, because the database of images used to train the algorithms is not very diverse.*
> VOICE 2: *. . . that showed with self-driving cars, when they tested pedestrian tracking, it was less accurate on darker-skinned individuals than lighter-skinned individuals.*
> VOICE 3: *Computer models were trained by observing patterns in resumes of job candidates, largely from men, teaching themselves that male candidates were preferable.*

So how could we, as a society, mitigate this risk? We could place limits on how companies are allowed to use these AI models. Sort of like the Supreme Court did when it found Duke Power Company in violation of the Civil Rights Act.

Or maybe we could pass new laws that target the structure of certain AI models themselves.

We'll see what Suresh thinks in a bit.

There's one last risk to mention. And it's a big one.

Risk number 4: total annihilation. The so-called... existential risk.

TRISTAN HARRIS:
What was needed was to, I don't want to say slow down because that's going to hit some people the wrong way. It's, can we move at a pace where we could get this right and not blow ourselves up?

RAFFI:
Here's Tristan again.

TRISTAN HARRIS:
There is not some kind of master set of adults that know how to make this go safely. You know, Raffi, you and I can just sit here calmly and assume this is going to go well. Unfortunately, we really are at the frontier of an issue, just like nukes. I think this is equivalent to the development and deployment of nuclear weapons and the need to create a control structure that would prevent the catastrophic use, because that's kind of what we're playing with.

If you were a nuclear scientist in 1946 and you said, oh my God, we just literally uncovered this truth about just what happens if you know how to split an atom. Once that truth is out there, you could have been one of the nuclear scientists who, by the way, there's many who did commit suicide because they thought it was inevitable that the world was just going to have nuclear proliferation and you're going to have hundreds of countries with nukes and it was inevitable they were going to be used. And so some committed suicide.

Now, notice we live in a world where only nine countries have nuclear weapons because we actively made choices about the control structures we wanted to create in the world. It's much harder with AI to create the control structure than it was with nuclear weapons because uranium enrichment and all that was something you needed state-level resources and only a finite set of actors could have them and you could monitor where the material came from. It's harder with AI. That said, we have to choose that we want to do that. Unless you want to allow the alternative, which is you just allow maximum proliferation.

And I really think this is a mistake. And if there's a time in which we don't want that to happen, it's right now. It literally has to happen in the next few months to a year, right? Like we, we have to rush to create those control structures. That's what we're working on. That's what we need everyone working on.

RAFFI:
Is generative AI *really* on par with nuclear bombs? I suppose if we have to ask that question… it's worth being concerned about. It can be hard to separate what's truly concerning… from what we're only scared about because we saw it in a sci-fi movie.

> **CLIP**
> YANN LECUN: *Uh, we're not that close to human-level intelligence.*

RAFFI:
Yann Lecun, a winner of the Turing award, basically the Nobel prize for computer science. He's also the chief AI scientist for Meta.

> **CLIP**
> YANN LECUN: *Until we have some sort of blueprint of a system that has at least a chance of reaching human intelligence, discussions on how to properly make them safe and all that stuff is I think premature, because, how can you design seatbelts for a car if the car doesn't exist? So, I think you know some of those questions are premature, and I think a bit of the sort of panic towards that future is misguided.*

RAFFI:
Some people are convinced that worrying about AI as a civilization-ending threat is. . . misguided. Or a kind of smokescreen. A flashy risk designed to take our eye off the ball, or make us forget about individual accountability.

MEREDITH BROUSSARD:
One of the ways that people have done gatekeeping around technology in the past is they've made it seem really mysterious and incomprehensible.

RAFFI:
Meredith Broussard is a data scientist, professor, and author. Her newest book is called "More Than a Glitch: Confronting Race, Gender, and Ability Bias in Tech".

MEREDITH BROUSSARD:
And it's like, oh yeah, this algorithm is so hard to understand that you're not really smart enough to understand it, so I'm just going to keep all of the knowledge to myself and I'm going to make the decisions for you because I'm smarter because I know math. Uh, I don't believe in intimidating people through technology. And I also think that understanding artificial intelligence specifically is hard. I mean, machine learning, which is the most popular kind of artificial intelligence right now, machine learning is just math. It's a very complicated, beautiful math.

When you understand it that way, like, it's just math. Nobody thinks that math is going to rise up and take over the world. Like,you see all these like, you know, collective letters about existential risks of AI, it's like, why are you like getting all worked up about killer robots when the mundane harms of AI are happening right now? And it's about people making poor choices around math and machines that do math.

Raffi:
That's exactly what I wanted to ask. Like, I feel like they're taking the air out of the room on a hypothetical compared to actual problems that are occurring today because of these systems.

MEREDITH BROUSSARD:
Yeah, people should stop doing that. Journalists should stop printing it. [LAFF]

Raffi:
Maybe then, how do we get engineers or future engineers to start thinking about these issues earlier?

MEREDITH BROUSSARD:
So, engineers need interdisciplinary education. We can think about how people learn. We know a lot about how people learn. And one of the ways that people learn really well is through storytelling.

So one of the things that I do in the book is I collect all of the amazing journalism and scholarship that's happened over the past five to 20 years, and put it together and it has a

different impact when you see all of these stories together. So when you, you know, say read an investigation from The Markup every couple of months, you're like, oh, that's really bad. Like, we should do something about that. But then when you read about the history of financial discrimination in the United States and you look at the automated mortgage approval algorithms. And you find that these algorithms are 40 to 80% more likely to deny borrowers of color as opposed to their white counterparts. And then you think about the history of things like redlining in the United States. And you look at the ways that that discrimination is reflected in the data that's used to train the automated mortgage approval system. And you realize, oh yeah, this is a systemic issue.

If you only come across these stories every couple of months, you might think it's just a blip. It's a glitch that you know, this person got arrested because of a faulty facial recognition match. But it's so much more than just a glitch. It's about systemic racism, systemic sexism, systemic ableism, and the way that these problems, these social problems are embedded in the data that we use to train our machine learning systems.

RAFFI:
There are serious problems to consider already, from our first three risks: like what happens when AI models are rushed to the public, or the possible loss of objective truth in an AI-dominated future, or the disparate impact of AI on minority populations.

If we follow the "existential riskers" too far, then maybe we don't have to address the massive disinformation problem of ChatGPT... or maybe we don't have to confront how face-recognition doesn't work as well on darker-skinned people.

If AI's just gonna kill us all... then what's the point?

But Tristan sees things differently. He thinks the existential risk is real... and that people are afraid to acknowledge it.

TRISTAN HARRIS:
One thing I'll say is that I know that senators in the US have been afraid of talking about the more catastrophic risks that can come from this because they're worried that it sounds a bit sci-fi. Even if they're privately concerned about it, they're worried that they can't talk about those concerns publicly. We actually handed some of the senators' quotes from the CEOs and founders of the major AI labs where, in their own words, they're saying, this has existential risk for humanity. hold those CEOs to their own statements and ask them, so what about this is an existential risk? How many people are you putting on it? What should happen to prevent those risks? And they didn't ask those questions because they were worried it would be politically risky for them to do so. I think this is a huge mistake. We have to move the Overton window so that it's totally legitimate to question the kind of risks that we're talking about. These are not sci-fi risks. And that's something more that still needs to happen.

RAFFI:
So those are the risks. Now... what are we supposed to do about all this?

In part two, We'll hear from a Senator...

> **CLIP**
> SEN. BENNET: *Members of the Senate? I can tell you we're not gonna have any idea, any clue, what to do with that.*

RAFFI:
a Congressman...

> **CLIP**
> REP. OBERNOLTE: *That will create an environment where capital will flow through and create investment in new technologies, which is what we need to happen.*

RAFFI:
and a foreign policy expert...

> **CLIP**
> IAN BREMMER: *We will have massive scientific breakthroughs, biological and health breakthroughs, educational breakthroughs, that come from AI.*

RAFFI:
...all to help us see the current state of the thinking around AI regulation... and what some of the challenges might be for getting regulation off the ground.

That's coming up... in part two.

Technically Optimistic is produced by Emerson Collective, with original music by Mattie Safer.

For updates, additional content, and engaging discussions, follow us on social media! You can find us on Instagram, LinkedIn and Facebook ... at EmersonCollective.

I'm Raffi Krikorian. See you next week... on Technically Optimistic.

CLIP LIST:

- Daniel Leufer, in a Center for AI + Digital Policy Panel: https://youtu.be/FkJBv9aNvoU?t=551 (money quote at ~11:10)
- Cambridge Analytica: https://www.youtube.com/watch?v=mrnXv-g4yKU
- Pope in a Puffer, Inside Edition: https://www.youtube.com/watch?v=BCjhg994fIU
- Yann Lecun: https://www.youtube.com/watch?v=BY9KV8uCtj4
- Sam Altman softball Bloomberg talk (June 22): https://youtu.be/A5uMNMAWi3E
    - How does he respond to claims that he is calling for regulation in public but lobbying against it in private? https://youtu.be/A5uMNMAWi3E?t=374
    - These models are built on the backs of biased data, how do you deal? Altman's weak response about "alignment" and "retraining" that seems to me to beg the question: https://youtu.be/A5uMNMAWi3E?t=739
- The EU AI Act explained (WSJ, June 20): https://www.youtube.com/watch?v=i5iZNH2lCGU
    - German news clip including detail of the EU law's categorization of AI deployments by risk (and a ban on those deemed unacceptably high-risk): youtube.com/watch?v=kcOQYj9zcqw
- News clip ft. mention of Blueprint for AI Bill of Rights: https://www.youtube.com/watch?v=JXbddMpHY-Y
- Altman to Blumenthal testimony: https://www.youtube.com/watch?v=Pn-W41hC764
- Altman to Klobuchar (disinfo and elections): https://youtu.be/6P3FmJr1B4A?t=688
- Altman to Booker (on concern for big corporate interest): https://www.youtube.com/live/TO0J2Yw7usM?feature=share&t=9891
- Altman to Kennedy (three-prong plan): https://www.youtube.com/live/TO0J2Yw7usM?feature=share&t=5969
- Tristan Harris, the AI Dilemma: https://www.youtube.com/watch?v=xoVJKj8lcNQ
- Schumer AI framework announcement I: https://www.youtube.com/watch?v=F84He6kKEcE
    - Long version: https://www.youtube.com/watch?v=uuB-SK7EuGI
- McCarthy on AI: https://www.youtube.com/watch?v=mVRwDyxagF8
- Facebook wants Sec. 230 reform: https://www.youtube.com/watch?v=Gv3AZRBnqb8
- Marietje Schaake
    - https://www.youtube.com/watch?v=lNcHU2PJF2o
    - https://www.youtube.com/watch?v=NvKgpvwjKW8