Tristan Harris: Hey everyone, this is Tristan.

Aza Raskin: And this is Aza. One of the things that makes AI so vexing is the multiple horizons of harm that it affects simultaneously. We sometimes hear about this divide or schism in the responses to the immediate risks that AI poses today and the longer term and emerging risks that AI can pose tomorrow. In those camps, there's the AI bias and AI ethics community, which is typically focused on the immediate risks. And there's the AI safety community, which is typically focused on the longer term risks. But is there really a divide between these concerns?

Dr. Joy Buolamwini: About this notion of schism, it makes for good headlines.


Dr. Joy Buolamwini: I've heard this, there are camps. We got AI safety on one end, we got AI ethics on the other hand, we got the doomers, the gloomers, all of these things. I think it makes for interesting headlines, and I see it less as a schism and more as a spectrum of concerns. I think there are immediate harms, emerging harms and longer term harms. And I think the way you address the longer term harms is by attending to what is immediate.

Tristan Harris: Dr. Joy conducted the breakthrough research that demonstrated to the world how gender and racial bias gets embedded into machine learning models. Her work has been incredibly influential. She's helped set the agenda in the halls of power, and you may have seen the documentary on her work called Coded Bias, which is available to stream on Netflix. So we are absolutely thrilled to have her on the podcast.

Aza Raskin: And I think the thing we all agree on is the urgency of all these risks makes it imperative that the people who bring different perspectives can come together and talk and find common ground.

Tristan Harris: Dr. Joy, welcome.

Dr. Joy Buolamwini: Thank you so much for having me.

Aza Raskin: So I want to get into your research, but I actually want to bring listeners to where you and I got to meet, which was at a little meeting we had in San Francisco with President Biden. And I really want to say that I think that you had more impact in that meeting than everyone else because you told the most compelling story about Robert Williams. So I thought we start there with what was the story you told President Biden that relates to your work?
Dr. Joy Buolamwini: Oh, well, that’s so kind of you to say. There were many impactful people and conversations that we had, and I was really grateful for the administration really taking the time to deep dive. So for that meeting, I brought in two photos and I passed it around to President Biden, Governor Newsom. I think you also got a few of those photos.

Tristan Harris: I did. I still have it on my desk.

Dr. Joy Buolamwini: One of those photos shows a man named Robert Williams and he has two small girls who are his daughters. And he was actually arrested in front of his daughters and his wife because of a false facial recognition match. And then he was held in a holding cell for over 30 hours. And this was also around his birthday. I mean, it just gets worse and worse. And so it was really making sure that we were putting a face on AI harms, because it’s so easy to talk about it in broad terms. You’ll say a sentence, “AI can be biased” or “There could be racial discrimination” or “It’s being used in the criminal legal system in harmful ways.”

So I really wanted to say, who are the people who are being harmed by AI, those who are convicted or condemned due to algorithmic systems? And what impact do these types of interactions that involve AI as a witness have on people’s lived experience? And so sharing that photo around to humanize AI harms and the conversation was a launching point for President Biden to then ask is the reason he was falsely identified because he was black. And that really got to the heart of the matter.

It was also an opportunity to share that not only do we have documented racial bias, we have documented gender bias, documented age bias, documented ability bias, colorism when it comes to facial recognition technologies. And then if we’re thinking about AI systems more broadly, think of an ism, it’s probably been encoded in some type of AI system being deployed and it could be deployed near you, so at your child’s school at a hospital. And I think this notion that no one is immune is really important.

Aza Raskin: I just want to say actually as a testament to your work, it’s so true that we had someone from our team actually give a talk at a global semiconductor conference about the risks of AI. And the first thing that everybody asked about was algorithmic bias. And it really speaks to you’ve created an agenda of concerns that, like you said, really wasn’t there maybe six years ago, and now really is.

And I think this is something to really to celebrate. We should probably get into the core here if we’re just sort of setting the table for listeners. People think computers are running on just code, and so therefore the system’s got to be more objective or more neutral. How do gendered racial and other biases find their way into AI?
Dr. Joy Buolamw...: Yes. So the approach to AI that is currently very popular uses machine learning. And so machines are learning from what? Large data sets that are used to learn various patterns around the world, patterns of what a human face looks like, patterns of what a sentence or an essay looks like. And so you can have large language models like the kind of AI models that power ChatGPT, or you can have AI systems like the type that power facial recognition systems used by law enforcement.

So where does the bias come in? Where does the discrimination come in? When we've looked at data sets that are open for scrutiny, when it comes to face data sets, as I was doing my research, I encountered so many data sets that were overwhelmingly of lighter skinned individuals and overwhelmingly male individuals.

One of the gold standard databases, if you permit me to get into some technical weeds a little bit, Labeled Faces in the Wild. LFW was the gold standard. And when you looked at it, it was about, I want to say over 80% lighter skinned individuals, 70% or more male. And so it wasn't so surprising if the measures of success themselves were skewed, it meant that the field as a whole had a false sense of progress. So was facial recognition advancing in some domains? Yes. Was it advancing the same across all demographics and populations? No.

At the time I was doing the research, it was very rare to find any papers that would disaggregate numbers. Typically, you would say, "Here's the gold standard data set and here's the overall performance on that data set." So what I did with my MIT research was say, "What would happen if we changed the test? If the test included more women, if the test included more people of color, would the result change?"

And so I decided to focus on gender classification, binary gender classification to be more specific, not because gender is binary, but most of the gender classifiers we were looking at had that gender binary. And it turned out that with a more inclusive data set, which I called the Pilot Parliaments Benchmark, when we tested systems from IBM, Microsoft, Face++, later on Amazon, and Clarifai, we found that there was indeed substantial bias along skin type, along gender and very importantly at the intersection.

So the gold standards turned out to be pyrite, fools gold. And it wasn't just the lesson for facial analysis systems like gender classification or age estimation, but really any human centered AI model. So think about AI models being developed to detect cancer, to predict the formation of plaque for heart disease.

Aza Raskin: One of the things I think your work does so incredibly well is it makes these invisible things visible. Actually, it does more, I think. The way you communicate, it makes them visceral. People care. They can see it. They can feel it. And so far we've been focusing on the harms of I think what Danielle Allen, who's the
political scientist and Harvard professor calls Generation 1 or 1.0 AI. Dario, the CEO of Anthropic says in the next two to three years, we'll hit AGI. And I think what he means by that is it can do the economic work of a normal human being across most tasks. And Sam Altman says super intelligence in four years, etcetera. So I'm just curious.

Dr. Joy Buolamw...: I want to make a big distinction with what you just said there because I want to be sure I'm clear with what you're saying. You're saying AGI and then you're saying AGI being AI systems being able to do a lot of the economic labor that's currently done with humans. Though AGI can also be understood, and it's often said to be, as you were going forward with, superintelligence. I want to make sure we are very precise about what we are talking about because I don't see superintelligence the way maybe some of the people you've described see it. And I worry about that with algorithmic systems.

We don't have to have AI systems flagging you as a terrorist suspect or flagging you. Imagine a drone with a gun and facial recognition, and you don't necessarily have to have super intelligence for military applications of AI to be immediately deadly. So I want to be careful in the conversation to not necessarily accept all of the premises, but still have the conversation. And so to me, this notion of superintelligence, I am very cautious about buying into the notion of sentient systems, which we do not have and I do not see us having in the next few years. And that being said, we can still acknowledge that there are very powerful AI systems that can absolutely do economic labor.

Aza Raskin: One of the questions that we ask ourselves is can you have an aligned AI in a misaligned system? Of course the answer is no. If no matter how well you align your AI, if it's in a misaligned system, it's going to cause harm from that misalignment. That reminds me of that phrase, if you make life better for women, you make life better for everyone. If you make life better for black and brown people, you make life better for everyone. As the power of AI continues to increase, the cost of our misaligned system will also increase. And so maybe this is the wake-up call for humanity.

Dr. Joy Buolamw...: I will say the term alignment and misalignment I find difficult to use because if by misalignment you mean AI that's discriminatory, if by misalignment you mean AI that is spewing hate speech, if by misalignment you are using a somewhat safe word to describe very harmful things, I lean towards saying the more harmful things that aren't aligned because alignment can also look like what type of goals you wanted to achieve. And I've seen the evolution of discourse in this space. We didn't always talk about responsible AI or AI safety or I've been hearing more recently, beneficial AI, but I've seen that those terms feel more comfortable for some people to engage in a conversation than saying AI racism, AI discrimination or misogyny.
So when I hear the term AI alignment, I’m always asking, "What do you mean? Is alignment a softer way of talking about algorithmic harms, algorithms of oppression, algorithms of erasure, algorithms of exploitation? Let’s be clear about what we’re talking about." So if we can be more specific in our language, I think that also helps us to be more specific with the types of guardrails we put in place. So when I hear something like alignment and I look at where that type of language is being used, it concerns me that it is used to remove oneself from the more challenging societal conversations.

Tristan Harris: I’m curious how you see this in relationship to incentives because the more we hype the tech, the more quickly there is an incentive for companies to replace people on their staff with tech that is over-hyped in terms of its ability, which would then accelerate all the places where it’s biased and has all these problems. I think just to say one last thing, I think there’s a unifying frame here actually in a lot of your thinking, in our thinking, which is noticing that social media was also an AI that had all of these harmful effects. And we haven’t yet even gone back and fixed first contact social media with AI, just like we have not gone back and fixed many of the systems that you’re talking about, that are not safe and effective that are causing harms right now. And now we’re racing to scale and deploy these even more powerful systems without actually going back and fixing them. And I’m curious your reactions to that and how you see incentives playing a role here.

Dr. Joy Buolamwini: I do think, at the end of the day, we cannot rely on self-regulation. I do think this is why laws are necessary and this is why legislation is necessary and this is why litigation will be ongoing. As somebody from a computer science background initially, I very much dismissed policy and advocacy and my initial approach to some of these issues was a very technical approach, which is okay, we got biased data sets, let’s make the data sets more inclusive. That might address algorithmic bias, but it doesn’t get to the algorithmic harm of accurate systems can be abused. Right now I see a contradiction, companies saying, some companies, not all, but many leading companies saying that we should pause AI, AI poses an existential risk more than climate change even so, and yet despite claiming that the risks are so high, nonetheless moving full speed ahead as if they have no choice. There are choices and there are also profit incentives.

So there is a benefit to hyping AI because you then get to be the creators of this powerful system that even you do not necessarily fully understand. So there is a bit of a marketing mystique that also helps where it’s almost trying to, it’s more saying we don’t know what’s going to happen. It could be harmful, yes, regulate, et cetera. But then when it comes to putting in safeguards, putting in guardrails that could cut into potential profits, you’re going to be misaligned with what the profit interest is and where the public interest is.

So I do fear corporate capture of regulatory processes and also legislative processes. Do I think corporations shouldn’t be part of the conversation? No, we
need all of the stakeholders in the room. Should they hold the pen of legislation? Absolutely not. And I think we also shouldn't be so bought into the idea that only they can understand these systems to then make themselves the only people who can then propose the "solutions," which is another system they sell. So I think we have to be really mindful of that.

Tristan Harris: The first thing I relate to is what you said of as a computer scientist, you saw these problems and your instinct was, I'm going to grab the tools that I know how to grab, which is like oh, it's a code and data problem. Let's grab the code and data solution tools and let's put the tools in. Identifying a lot of the social media problems as very much a design problem, like it's designed in this way that I can clearly see is drawing up social validation and addiction and social proof and social pressure, and it's just playing with all these psychological levers. And I'm like, "Oh, that's a design problem and I'm a design thinker, so I'm going to pull out my design tool and say, 'Hey, Google', when I was inside of Google, 'Why don't you just change the design?'" Because it could just work so much better if we don't do those things and if we set some standards inside of the Google design standard code base.

But then what I think we both came to here, what I'm hearing you say at least, is it's really about the incentives and you need to get out of your tool set and say we have to have regulation that changes the incentives or something that changes what the incentives are because the incentives are what drive the action. But there's an interesting place here where actually I'm not sure that we do agree, or at least I'm interested to dig deeper, which is the belief that the companies are hyping the risk, because as I heard you saying, is surely they don't believe it's this risky because if they did, why would they be building it? But I think there is an explanation for that, which is that they are worried that that power does confer advantage to those who do if they can control it and up until the point where if you can control it, then continuing to build it so that you have access to that power before they worry the non-good guys have access to that power.

I'm not saying they're the good guys, but I think that there is a good faith interpretation of their fear about the power of what they're building. And we do talk to a lot of the AI safety people inside the labs and the ones that we talk to, at least I have viewed as good faith in their concern that what they're doing could be really, really catastrophic. But then it comes back to the incentives where they don't have a way to stop everyone else from building it. And so it's not so much that they're super powerful and more like they're helpless and caught. And then there's this last part which you're naming, which is correct, which is the regulation that they're proposing would cement their concentration of power where only they have access to build these sophisticated systems, which would now be a whole other set of problems.
Dr. Joy Buolamw...: Yes, we do disagree sometimes, and that's fine. So to the point where I think there is a bit of disagreement, I separate institutions from individuals and my experience with the Gender Shades Project, so we audited AI systems from leading tech companies and I had an opportunity to talk shop with the people who created some of these systems. And the conversations I would have with the tech teams and AI researchers within companies were very different conversations than what I would have with executives or people from the communications team or people from the legal team. And so I think I can agree with you that there are people who have true and legitimate concerns about the risk of AI within companies and outside of companies. I do not necessarily see the institution that houses all of these individuals, the companies then actually taking the steps that would put more belief behind what they're saying.

So if you're saying there's a pause letter and you have 30,000 people sign it, but they're not pausing, I'm saying your actions are literally not matching your stated concerns, point-blank. And the counter for that is, well, if we don't do it, someone else is going to do it as if we are somehow helpless, which we are not. The motivation is the profit if we have that power first. Yes, you can say maybe you can prevent other bad actors if you are the arbiter of determining who's bad or good, and it's more complex than that, but there's also the power in having tools that you will sell to others, and not surprisingly, these are for-profit companies. So I still think there's a contradiction in what companies are saying and the actions they are taking and the contradiction is because they want to make the profit.

Aza Raskin: I'm curious what you would say. I'll take the opposite side that I normally am on and I'll sort of be one of the people that we talked to inside of the AI companies and they'll say, "Look, I am really worried about this technology, but I'm aware that in the end of the day it's just matrix multiplication, just a whole bunch of it." So we can't stop people yet from building it. I have an obligation to build and build it in a safer way than those others might.

Also, there are other countries that are going to be building it that don't share our values, like a Russia or a China or North Korea. And because this new technology confers new power, if we don't build it, then we will be beaten by those that do. And also I think I have good values and if I don't build it, then I don't even have a seat at the table, so then it's irrelevant. Therefore, I don't have a choice. The best I can do is tell people how dangerous it is because that way I can get someone else like the government to help me coordinate because we all have to stop at the same time. If we don't all stop, it doesn't work.

Dr. Joy Buolamw...: There's so many assumptions there that you have to think through.

Aza Raskin: Great, great. Yeah, let's break it down.
Tristan Harris: Let's break it down. This is so critically important, Dr. Joy. I'm so excited we're talking about this.

Dr. Joy Buolamwini: Again, I've spoke to many people in the biometrics industry where ... I'm trying to think of the ones I'm able to share. We have bias in the wild stories, et cetera. So when our research came out, we had people saying, I worked on quality assurance and I knew that these issues existed. But it would've made my job harder to actually address it. So I don't necessarily buy that building tools that would remove racism and sexism or minimize it or address it then would somehow compromise the ability to do other types of AI innovation. I do think that sometimes this notion of if we're building responsibly, that means we slow down. So that means the other people get an advantage. I think they get a short-term advantage, but the long-time societal impacts do not outweigh those short-term gains in the first place. So I still think it goes back to the profit motivation.

For example, you could have R&D and still not release models. That was what was happening for a very long time. It's because ChatGPT got 100 million users in a very short time, historic. And that shifted the market dynamics and the market power. That's what happened. So I don't completely buy some of these rationalizations after the fact, and then we had Meta release Llama 2. So now you have open source available to many people in a very dangerous way, ahead of elections, where we're getting more powerful systems.

Tristan Harris: 100% agree.

Dr. Joy Buolamwini: These were choices that were made that it didn't have to go down this direction and you could still have those same arguments. Those choices I believe came back to how do we assert some type of market dominance or market opposition because we realize whoever has supposed supremacy with AI will hold a lot of power in the world. These tech companies could power one government or another. So even when you have people inside companies saying other nations might move forward, the companies themselves are not tied to individual nations. They got clients everywhere.

Tristan Harris: That's true. But you could argue, and I think the people we talked to at the western AI labs are worried about China building artificial general intelligence level systems faster than the US is. But I totally agree. I think what we're coming to is actually a deep agreement of, it is really this race for market dominance. And when market dominance was prior to ChatGPT launching, when it was just the race to develop internal capabilities, it was a slower calmer race. When they published ChatGPT publicly and got to 100 million users in two months, that changed the form that the incentive was, where now it's about actually, if I don't release this thing that I've had in the lab to show the world that I also have a system that's as powerful as ChatGPT, because if they have 100 million users,
people aren't going to switch back and forth between different big public AI systems. Businesses are going to start building.

Dr. Joy Buolamw...: Right? So was that release because of China?

Tristan Harris: No, no, no. So we agree. I think we agree that it was unwise to hit the gas pedal and set a new clock rate for releasing systems by publicly launching ChatGPT and integrating it into being and having Satya Nadella say, "We want to make Google dance." And all of that drove up this race because literally they got a huge stock market boost from dropping that stuff. What I'd love to see is how do we bring all of our communities together who care about all this going well for bias, for discrimination, for misinformation, for democracies, and for bioweapons and for some of the bigger cognition risks of our AGI, that we actually all want the same thing, which is to move at a pace that we can get this right rather than move at a pace that we keep just shoving harms onto the balance sheet of society.

Dr. Joy Buolamw...: The question is, are you willing to lose something? This is the question of power and privilege always. So it's very easy to pay lip service to inclusion. It's very easy to pay lip service to ... Who's going to say we want discrimination? Usually, I mean, the dialogue has changed. I think it really comes down to when it costs you something to do the thing that's better for society than the thing that is better for you as an individual or for you as a company. So that's where that global coordination does require regulations.

I was at a UN related event a few weeks ago, and Professor Virginia Dignum said something that I thought was really interesting because in that room we were starting to have the false dichotomy of innovation versus guardrails. And she was saying, "AI is like a car that hasn't gone under rigorous safety checks, being driven by a driver without a license on roads that are barely paved without even traffic lights."

And so back to some of what we've been discussing, it is up to governments and also up to people to agitate to say, "No, we do not accept this wild, wild west that we're seeing." And the way in which you're describing the conversations within the tech companies absolutely shows why they are not the ones to lead this because they want to be first and they do not want to give up something that could potentially compromise that position. They have shareholders, they have quarterly profits to think about, and I'm not seeing very long-term thinking right now.

Sometimes you have to go outside of those who are incentivized for the short term, which also makes it difficult for governments as well, because election cycles also lead to a lot of short-term thinking, what can be the quick gains made. So we are not going to change the fabric of society if we don't address power differentials and if nobody is willing to lose something.
Aza Raskin: I really love that because wisdom so often is knowing when you should say no to something. Our friend, Mustafa Suleyman, who we just had on the podcast, the co-founder of DeepMind says that in the age of AI progress will be defined more by what we say no to than what we say yes to. I heard you say the release of Meta's Llama 2 open source is dangerous. And of course after that came out, the United Arab Emirates released Falcon. Mistral AI just released their open source. So now there's a race for more and more powerful open systems.

And I'd imagine some of our listeners might say like, "It's surprising to hear you say that open source is dangerous, Dr. Joy, because isn't that democratizing access to power, which you want to get it into as many people's hands?" It's not what I believe. But I love to hear your take on that because that's sort of what Mark Zuckerberg and Marc Andreessen, apparently just Marks, will make this kind of argument and I think you'll have a very powerful rebuttal. And so I'd love to hear it.

Dr. Joy Buolamw...: Overall, I remember learning Drupal when I was a kid, when I was a high schooler Drupal open source content management system, and then I built a little web development company off of that, and then that meant that I could make websites for all kinds of people, even had an opportunity to make a website for an Ethiopian embassy in West African Nation, blah, blah, blah. And I was doing this as a high school student. When I was an undergraduate at Georgia Tech. I led the development of mobile surveying tools with the Carter Center for a project we were doing on neglected tropical diseases. And because of the open source nature of Android, we were actually able to build bespoke tools. At the time, Google Android did not come with an Amharic keyboard. And so because of the openness of that system, we were able to build the type of keyboard that was necessary, et cetera, load in the Amharic font, et cetera.

Later on, when I talked to Google engineers who would've been part of those teams and I described what we did, they didn't have the market incentive to do that. So do I believe in open source in terms of its power to democratize access to the tools of creation? Absolutely. Have I benefited from it? Yes. Has the rest of the software industry benefited from open source? Absolutely.

But even with all of that, I think with AI capabilities, we have to be extremely careful when it comes to data and consent and privacy. It would be one thing to open source data sets where people had agreed to even be part of those data sets in the first place. I dodged so many subpoenas while I was in grad school because big tech companies had scraped many face data sets without people's permission. And in areas where you had laws like BIPA, the Biometric Information and Privacy Act of Illinois, they actually had a case to be made and there were many lawsuits filed.

And so part of my pushback on what is being open sourced is was there permission and consent in the first place. Because we are seeing the open
sourcing of models that were built on some would say stolen data, but certainly data collected without consent and compensation. So it's one thing for me to open source something I built completely myself. It's another thing to open source something I built based on what I took, and now there's still lawsuits happening. And before we've even resolved that, now we are creating these chains of bias and discrimination. It's another thing for me to open source something where I have clarity about the risk and limitations. We've spent some time in this conversation talking about the various harms that are introduced, and especially with these large language models being so large that there wasn't necessary vetting and accounting of what's even included in the first place.

So I think it's one thing to share a meal where you know where the ingredients have been sourced, versus inviting people to a table where you're like, "No, labels. Look, good luck. I hope you don't get food poisoning. We're open sourcing." So I really think what is being open sourced? Who had a decision or a choice in the matter? Because I do feel in some ways, some companies are attempting to get a pass over the original sin, a sin I was a part of because it's literally how I learned computer vision. If the data is there, it is for the taking. This is when I was a grad student and I was doing the IRB process, Institutional Review Board, to make sure that what we were doing was ethical. When it comes to human subjects research, there are additional steps that have to be taken. As I was going through the process, because I was doing computer vision research, I had an exception.

It wasn't considered human subjects, even though I was using people's faces. And because it wasn't in a medical context that I was using it, I really didn't have to do more than just say that it had the exemption. And when I talked to my peers and older scholars, et cetera, people were just looking at me like, why would I make things harder? Why are you asking all these questions? Get the data and go, this is just how we do it. But now there are many data sets that are being challenged because once people realize, oh, this dataset I created has immense value, how do we know it has immense value? This company just raised $10 billion - where are my data residuals? So on one hand that's happening, and then you see open source. And I do believe there are a lot of people who, from the general concept of open source, I support this idea that we don't want certain tools to only be in the hands of a few.

And I do believe that overall, when you have many more minds working on various problems, you're likely to find more robust solutions. And I also believe that if you only had a few major tech companies in control of what's possible with the platform, it could be extremely constrained. That doesn't allow possibilities so that for example, if Google decides Amharic is not a priority language at that particular time in the development, it doesn't get done and now the system is closed where you can't do anything about it.
Coming back to the AI space, I think there are different ways of open sourcing and being thoughtful about it. I cannot say with what I saw with the release of Llama 2 and also where we are with the lack of regulations, that it was a responsible release. It's not in a context where we've established the rules of the road. For me, this is putting out the car that doesn't have the safety checks, where drivers don't have license, where we don't have rules of the road. Am I against vehicles? No. Am I against getting to point A to point B a little bit faster than what I could do walking in general? No. But I think this stage in the development of AI, because of where we are in actually safeguarding it, it was too soon. That's my perspective.

Tristan Harris:

One thing we've noticed is there's this schism between people who work more on AI bias, AI discrimination, AI ethics issues is kind of the common term of art. And then people who work on say AI safety, concerned about different catastrophic threats, whether they're biological, chemical, all the way to definitely more of the sci-fi ones, which not everybody believes. I'm curious, what do you think about the schism? Is there a schism? Does there need to be a schism? Because I think what I'm hearing you say over the last 30 minutes or so, we agree that we need some kind of top-down rules, a driver's license for the car, safety checks for the car, test reviews, a safe road, traffic lights. And I think that's what we all want. And so I'm curious just to go there and ask you what do you think?

Dr. Joy Buolamwini:

About this notion of schism. It makes for good headlines. I've heard this. There are camps. We got AI safety on one hand, we got AI ethics on the other hand, we got the doomers, the gloomers, all of these things. I think it makes for interesting headlines, and I see it less as a schism and more as a spectrum of concerns. I think there are immediate harms, emerging harms and longer term harms. And I think the way you address the longer term harms is by attending to what is immediate.

So when I think again, of existential risk, and I think of the campaign of Stop Killer Robots that's been around for some time. What does it look like when we look at some would say the future of peace, some would say the future of war, what it looks like with putting in different types of AI capabilities with various sorts of militaries.

I think where I see a lot of frustration is the airtime that is given for the most extreme views and what some would call AI doomerism. This is the end of the world. Superintelligence will emerge. And we are here to warn you and say, even if we were part of creating these systems, we told you. In most conversations I have that aren't internet mediated conversations, but real conversations with most folks, even folks who would be identified within the AI safety community.

And it might be because they're talking to me, so they're changing what it is that they're saying. There tends to be more of the, "We do know there are
immediate harms." I think what has been very frustrating for many people who look at AI bias and discrimination is when those harms are categorically placed as lesser. Sure, people could face discrimination or oppression, but to be honest, they've already been facing all of these things.

And maybe what is more of a threat to some people is those who are used to being in power are now at risk of being marginalized by their own creation, to then face the oppression that many other groups of people have dealt with for centuries. And so I think it's really thinking through the power positioning and who those narratives serve because those narratives about existential risk when we're really talking about AI destroying the world, I think it's interesting how the way we're using language, when I think of X risks for AI, I think of the X coded the people who are being harmed by AI systems because we can help them now. We don't have to wait until there are trillions of future humans. What does it say about us as a society if we don't help the people who are drowning in front of us saying we hope to one day help centuries down somebody who could hypothetically drown?

And this isn't to say we shouldn't be forward-looking, but I do think we have one, an opportunity that's a real opportunity to mitigate more of these immediate harms. I think about Porcha Woodruff, who was arrested for a carjacking. Eight months pregnant, sitting in a holding cell, no one's jacking a car, eight months pregnant. She reported having contractions and then she had to be rushed to a hospital after finally being released. And then just that disregard of life because this happened in 2023. Detroit Police Department, same police department that falsely arrested Robert Williams in 2022.

So I absolutely see the frustration of saying we're talking about all of these hypothetical risk and we're not seeing acute known risk being addressed. And I absolutely think that is a mismatch of priority. Can we walk and chew gum at the same time? Absolutely think we can think about acute risk near-term risk, emerging risk, for sure. I don't agree with the doomerism type of framing of existential risk, but there are others who do.

It's when that kind of framing takes away resources, takes away regulatory attention from actually building the safety checks, getting the driver's license and putting the streetlights on, which are things we can do. Those things are hard, and they require compromise and they require negotiation. But overall, it's not that Google wins or Anthropic wins, et cetera, it's that humanity gets to win.

Aza Raskin: Dr. Joy, I thought this whole conversation was just incredible. Thank you so much for coming on Your Undivided Attention.

Dr. Joy Buolamw...: Thank you so much for having me.
Dr. Joy Buolamwini’s book is called *Unmasking AI: My Mission to Protect What Is Human in a World of Machines*. And it’s out now. And before we go, we wanted to play you Dr. Joy’s spoken word poem that she wrote, which touches on a lot of the themes we’ve talked about today. The title of the poem is "Unstable Desire."

Prompted to competition, where be the guardrails now? Threat in sight, will might make right? Hallucinations taken as prophecy, destabilized on a middling journey to outpace, to open chase, to claim supremacy, to reign indefinitely. Haste and paste, control altering deletion, unstable desire remains undefeated. The fate of AI still uncompleted. Responding with fear, responsible AI beware. Prophets do snare. People still dare to believe our humanity is more than neural nets and transformations of collected muses. More than data and errata, more than transactional diffusions. Are we not transcendent beings bound in transient forms? Can this power be guided with care? Augmenting delight alongside economic destitution? Temporary band-aids cannot hold the wind when the task ahead is to transform the atmosphere of innovation. The Android dreams entice the nightmare schemes of vice.

*Your Undivided Attention* is produced by the Center for Humane Technology, a nonprofit working to catalyze a humane future. Our senior producer is Julia Scott. Kirsten McMurray and Sara McCrea are our associate producers. Sasha Fegan is our managing editor. Mixing on this episode by Jeff Sudakin. Original Music and Sound Design by Ryan and Hays Holladay.

And a special thanks to the whole Center for Humane Technology team for making this podcast possible. You can find show notes, transcripts, and much more at humanetech.com. And if you made it all the way here, let me give one more thank you to you for giving us your undivided attention.