WillnerBerkman

Wed, Mar 27, 2024 5:27PM • 1:14:00

SUMMARY KEYWORDS

moderation, ai, content, problems, people, policy, systems, human, model, moderators, question, bad, work, decisions, space, machine, good, point, values, talking

00:06
classes

00:11
started.

00:14
Thank you

00:19
hear me?

00:25
I can hear you. All right. Well, I'll speak a bit louder than thank you everyone in the room and I'm zoom for joining us for today's RSM speaker series event. We really cannot be more excited to welcome Dave Willner. For a talk entitled moderating AI and moderating with AI. I'm sure everyone is excited to hear Dave, so offer just the briefest note of introduction, Dave was a member of Facebook's original team of moderators, playing a key role writing its earliest content policies, and building the teams that enforce them. After leaving Facebook, he took on roles building the community policy team at Airbnb, and as head of trust and safety at open AI. And he is now a non resident fellow in the program on governance of emerging technologies at the Stanford cyber policy center. Dave's experience and content moderation and trust and safety spans almost the entirety of their histories as fields. So we're extremely lucky to welcome him today, if your thoughts on where the space may be heading. Dave Willner.

01:28
Hi, folks, yeah, it's great to be here just wanted to start by apologizing as as was noted, it's, I've been doing this the entire time. And it is all my fault. So sorry about that.

01:40
I wanted to sort of make a case to all of you around the use of AI and content moderation and how I expect it to change things, I have come to think that how powerful foundation models are going to fundamentally transform how we do moderation. There's been a lot of focus on the novel risks that

those models present. That's, that's fine. Those things are true. I'm not going to dwell on that today, I think it's been covered. But the models are also because of their unique capabilities and ways of working, going to be very useful in solving problems that have previously been intractable. That's those sorts of solutions also, I think, going to prove deeply relevant to align the questions in AI itself, because at least today, our current ways of controlling and steering models are themselves downstream of techniques that actually have a lot of shared DNA with how we do content moderation at present. So just to sort of briefly recover who I am why you should care about any of this, I have been working in this field for about 16 years at the forefront of sort of controlling social technology versus social media than in the sharing economy that in AI, I spent a lot of that time trying to not just grapple with the problems on emerging technology, but grapple with using emergent technology to solve the problems it creates. So in addition to working on policy itself, I've spent a lot of time at the intersection of operations policy controlling, and figuring out how we actually do the news that these platforms claim to have. And I'm going to dwell a lot on that, that question of actual performance in the talk today. Currently, I'm a fellow at the San Francisco Policy Center. I was spending a bunch of time on this subject, learning how to use LLM to do content moderation. With a guy named Simeon Chakrabarti, who was another fellow at the Center. He ran civic integrity at Mehta from 2015 to 2021. We ended up doing the same fellowship, sort of coincidentally, and we're having very similar thoughts. Beyond using off the shelf models, we're also doing some work trying to train small or large language models to be good at this task specifically. Because we think that more efficient models to be able to do this review and your contribution to the space. I bring all of that up, just to make the point from my perspective here is very much a practitioners perspective, not an academic perspective, right? I come to this as someone just desperately trying to solve these problems for the last nearly 40 years.

04:14
And very focused on what practically works, and how we use these tools in practice, not nearly.

04:23
So first, just sort of set the table about why I have this strong belief about about the importance of AI in the future of content moderation. I want to do some grounding about how I see content moderation today. Why it works, and frankly, why it doesn't work very well. I think there's a broad spread agreement. It doesn't work very well. And it doesn't seem to be controversial to say here.

04:45
Bad things keep happening to clearly innocent people watching child safety hearings in the Senate earlier this year is enough to sort of demonstrate that anyone, there are very serious social externalities that are going on and this has naturally led to

05:00
holds a lot of public theorizing as to why there's a lot of discourse in the atmosphere about why social media, moderation doesn't work well. And I've come to think we're basically have any problem of evil conversation about tech giants. The problem of evil is this idea and theology that's concerned with reconciling the existence of a benevolent and omnipotent god with suffering in the world. And we're having a problem with people conversation about Mark Zuckerberg. Right? If, if Mark Zuckerberg is good, and in full control of Facebook, why do bad things happen to people on his platform? And a lot of

the popular explanations focus on this idea of benevolence, right there is either the notion that they aren't trying that the tech platforms don't care, they're indifferent. There's the notion that they're greedy, that they sort of do care, but they don't wanna spend money, or there's the notion that are actively malicious that they have bad values that are very antisocial and that we don't want, I those things may or may not be true, I don't think they're the root of the problem. The root of the problem is that they're very far from omnipotent. We're Put another way, we're bad at content moderation. Because we're bad at content moderation, we're not good at doing the core activity. And to understand why we're bad at it, it's important to take apart moderation into couple of components, values, and the actual classification tasks, the values piece of moderation receives a lot of attention. There's a lot of discussion about what the rules should be. Community policies are primarily understood, I think, has expressions of company's values. That's not untrue. But it is not the most significant thing that those policies expressed, I've come to believe that the focus on values is a form of bike shedding by charting is this idea of sort of realizing the slippery sort of us gaslighting, that if we were all on the border of a nuclear power reactor, all else equal, we would spend more time discussing the color of the new bike shed of the reactor than we would discuss a nuclear safety, because more of us can have opinions about colors and bike sheds that are qualified to have opinions about nuclear safety values function that way in this discussion, everybody has value, so it's easy to have opinions about.

07:13
The reality is that the sorting task underneath the values, the classification task, is the thing that we are very bad at. And that dominates any possible set of values.

07:25
To get into why sort of classification matters, and gives you some examples of how that is the case.

07:33
There are lots of situations where the value proposition that you want to achieve is not particularly in dispute, or where the ability to do it is very, very hard. To give you some social media examples. Reclaimed slurs are a great example of this. It's very intuitive to say we want to allow people who are members of community to use certain language, but actually doing that requires you to know at scale, who is a member of that community who they're speaking to what the African context of the conversation. So doing the thing is hard even if agreeing on whether LSAT is good or bad is hard.

08:06
Similarly, the controversies here often make are often made worse by public pressure, Facebook, breastfeeding photos, controversy back in 2010, was very much one of these problems in classification. The question of should breastfeeding mothers be able to upload photos of their children on Facebook is not really that interesting of a policy question, getting a moderation system to very consistently distinguish between nudity where there's babies involved, and it counts as breastfeeding. And to be fair, there's not it's hard and flawed. And so the ability to execute on policy is challenging. The napalm girl incidents that happened in 2014. Napalm girl is a reference here to a photo of a Pulitzer Prize. It was a photo of a girl who'd been attacked in Vietnam.

08:51

It's very, very intuitive to say, Oh, the Pulitzer Prize winning drone should be able to be on Facebook. But you have to know what all the Pulitzer Prize winning photos are, and get everybody doing moderation to know that to where you can't actually achieve plausible.

09:07
So why are we bad at large scale estimation? We're bad large scale classification because fundamentally, we're trying to solve an industrial scale problem with pre industrial solutions. Social media is this mass distribution machine that requires no intervention allows billions of people to talk to millions of other people nearly instantaneously without a direct human intervention in the communication itself is Heuer mass production of speech?

09:36
Well, we don't really have mass production capability for the ability to moderate speech we're still stuck in essentially a piecework system. piecework was a was a way of manufacturing, textiles and articles of clothing, sort of in the early industrial revolution when we invented London yarn, machines, but hadn't been invented machines that could do sock knitting. And so work would be parts

10:00
of how two people in their homes can be done in an artisanal way to a particular spec. Moderation today essentially works exactly like that, right? We have specific people sometimes distributed sometimes in one place, working against a document that tells them calorie content as well and artisanal level, except they're doing our maths. And systems like that have trouble achieving very high degrees of consistency. We don't have machinery to do those quarterbacks the process. And I'm going to sort of get into a little bit why humans struggle so well, so much to do this process as well. Even where we do have machinery for different parts of this classification process. Machinery itself mostly replicates the problems that humans introduce into the system as it exists today. And then I think, finally, the nature of language itself, probably caps how well we can do this, we're not, you know, making machine parts here or knitting fabric, we are ultimately dealing with classification of language and culture, which is a fuzzy activity inherently. And so the upper boundary of excellence is probably fairly low.

11:04
That's it until we make progress on at least some of those human or machine constraints. We're not going to see better moderation online.

11:13
I think AI is going to help with that because it surpasses both human capability and the capability of our current machines and a number of very specific ways that I'm going to get into after this after sort of digging into specifically how humans fail. Because the ways in which we are inadequate to these tasks are important to understanding the ways in which elements are helpful. Okay, so

11:35
why are you being battered classification, were better class when he shouldn't, for a lot of physical reasons, our working memories are really, really small. The length of the sort of content policies you

can feasibly write as a policy writer are like maybe a few pages, maybe five, maybe six. That's not because we couldn't write a much longer treatment of what hate speech might be. How to tell whether a photo contains nudity, it's because most people can't actually use 100 page document about hate speech might be to make 1000 decisions a day. Particularly not if you want them to stay up to date on what that document says. And you're changing it all the time.

12:12
Our long term memories are also very unreliable. I sort of alluded to this into the in the napalm girl context, but sort of notions of art, which is again, the thing I think we all think is probably good. And what's happening on Facebook, or notions of what a real name is, are basically look up questions, right? There's no art pixels, art is just all this stuff we've decided as art as a society. And so in order to treat that stuff differently, you have to know what it is, which means you have to remember what it is. And people are not terribly good at remembering huge amounts of very specific facts about individual pieces of content, it can be tough, but that's what getting PhD is for. It's not something that you do as an hourly job.

12:53
We're quickly exhausted, this work is coded as introductory level work, it's entry level work. But focusing intently on content for 1000s of repetitions for eight or 12 hours a day, is intensely intensely draining, people get tired, they make mistakes, they get bored, which is another sort of understated part of this, there's the trauma and emotional part of labor, it's been much discussed, and that is very much real. But honestly, a lot of the time for work is done. Many of the things you're looking at are not interesting to classify, they're not violating they're just kind of random noise. It's a little bit like staring at my nose on a television screen and waiting to see if something meaningful shows up. And it's very hard for people to maintain focus under those conditions.

13:39
We rely on our own internal models, people don't really use the rules to make these decisions. They read the rules once, use them a couple of times, internalize some sort of approximate model of whatever the rules say, well enough to not get in trouble with their boss, and then just keep doing that until I get in trouble again. And so as these rules change, people lag in that change. We also typically can't recall our reasoning. If you asked me given moderator widing in a particular decision, when they did it yesterday, if they made 1000 decisions, they're almost certainly not going to tell you. And so while this is this human process, which sort of seems like it has meaning, the meaning is often not retrievable. And then finally, and this one is often hardest for folks to grapple with. We really do not have any shared concepts. And there's a couple of specific examples that were really important shading. So we have a warning up front, all of this is going to get a little unpleasant from a content point of view. We were trying to figure out how to classify see Sam, Child Sexual Abuse material and Facebook for the purposes of creating the photos, DNA databases that today underlie a lot of the attempts to control that material online. We had 12 folks who've been doing this for a year and the recording information from that back that entire time. These were full time employees like me, these were kids who went to Stanford and Harvard

15:00

And when we asked them to classify material and simply reports and economic reports and economic without talking to each other, they could only agree about 40% of the time. And they've been doing this for a year on what you would intuitively think is the worst thing, the easiest thing to get consensus on.

15:16
No consensus, as most other areas are actually worse than

15:20
when we first tried to outsource nudity moderation to India, we ran into a similar problem, where we had a rules that it said also take down anything that was sexually explicit beyond a bunch of particular things we listed. And immediately, a moderator started taking down photos of people kissing, holding hands, because what we meant and what they understood, those words to me are not the same, because you just can't assume share reference or shared values.

15:47
All of that is made worse by the economic way we currently organize this labor. Right, so the work is very notably poorly compensated, I think some of that is probably inevitable given both the firm scale which is done, and a bunch of Europeans are going to get into here.

16:05
Enforcing being sort of forced into consistency is pretty good moralizing. It's an alienating labor, particularly because this is legal, where people have strong moral feelings about what they're being asked to do. And so it's unnatural to being apt to be asked to put on a sort of another another morality. But putting on the other morality to get everybody on the same page is like the heart of the activity. So you kind of can't read that. And that is itself draining and not a ton of fun. So people who have better options leave, right these are high turnover jobs as a rule.

16:39
In the context of this is well known in the context of our customer support. But even for Airbnb is outsourced trust and safety teams, the average retention is about nine months. It's a very, very short period of time, because when people have the ability to do better work, they go, that undermines the approval of expertise. It also means that you have to invest the time into sort of training and updating the system, because you're constantly teaching new people to do this stuff, and having to constantly reorient them as you make changes. So the entire system was extremely cumbersome, and doesn't need to sort of update interaxial results. Cool. Okay. Hopefully, I've convinced you at this point, people are not not good at sorting things into piles.

17:20
Why is our automation bad at this, right? Our existing automation is bad at this because all that you really doing is statistically copying, all those people who are not going to get right, our most advanced automation techniques on blackbox. Machine learning is just predicting what a human moderator would do. If you ask them to sort of piece of content. According to a policy, it's just a mathematical simulation of the results you would get if you had bothered to ask a person, which is very useful nuclear, because a lot of the time it does a pretty good job. And it means you can avoid asking people, which is great,

avoids trauma, much faster has some real upsides. But it also means that it inherits the fuzziness and the unreliability and the non specificity that we bring to that process. There are other forms of automation.

18:10
But they're they're actually even simpler, right? They're either asserted if then rules, exact word or pattern matching, they're all even less nuanced or less stable. We have no automation that does the activity that humans are actually doing. As part of the classification process. We only have automation that simulates the outcome of the activity that they've been doing.

18:34
And actually, the automation we have today, as a bunch of other problems, their decisions are meaningless, very literally meaningless, right? They're not making an argument about why a particular piece of content fits or doesn't fit a given policy. They're simply saying 95%, this is shaped in the same way as other things you told me violated this policy. There's no meaning to the decision, which both makes it hard to debug individual decisions, and is very disturbing for frankly, people who are subject to those decisions, because we want these things to have meaning to them, to dispute them and to be able to argue with them.

19:09
Updating the models is also extremely cumbersome. Training large machine models, under current circumstances means 1000s 10s of 1000s of examples, which means every time you change your policy, not only do you have enough data for humans and wait for that to all face it and you don't have to be follows humans 10s of 1000s of things to then be able to train your machine model machine learning model. So automation is also often very out of date. So the point where it's typically not fully synchronized with whatever given platforms policy allegedly is a time which is confusing and results in outcomes that are not so also machines that are very good at classification in these circumstances. And then on top of that, I think the best we can ever hope for here is significantly less precise than what we can ever hope for in material manufacturing. Right despite all the manufacturing analogies I've been making. We are not in fact making steel

20:00
all cylinders. With a metal fairing glaze, we're playing around with words and words are inherently vague.

20:07
The language itself is just not terribly precise. And that's particularly true in mass scale social media, where people often write frankly fairly badly or unfairly or approximately, from a Darwinian point of view.

20:21
everyday language is not meant to convey precisely specific meanings. It's meant to be efficient in communicating between people with a shared context. And social media. Moderation doesn't share its context is a very radically postmodern exercise all of the author's might as well be dead, there's always a text, you're sort of just staring at these word learns, after the fact. And that renders them very, very

difficult to understand. A version of this is the conversation about cultural context that comes up a lot. Having more specific cultural context is helpful here. But that's only a version of it. And in a lot of ways, the easiest version to solve right? Local interpersonal context is as big a part of this problem as broader cultural context. If I call you an elephant, am I calling you old? wrinkly, gray, fat, wiser, not Republican?

21:11
Right? There's no there's always a no the answer to that question outside of our specific interpersonal relationship, and there never will be. And so we're sort of tapped at the maximum here. So that's the Doom.

21:25
As an aside, all of the problems I outlined are also problems for AI alignment under current circumstances. So our primary techniques for AI alignment, reinforcement learning, rely directly on curated datasets of desirable behavior that we are trying to get the machines to copy, all we're doing in RL is curating a set of prompts and responses to our from allegedly from the model that we want the monitor to behave more like, and then through a mathematical process to get the model to eat that behavior. Which means that ultimately, what we are aligning the model to do is dependent on the same kind of content classification has the same kind of content classification problems that I've just been talking about in social media context. And you can see this in the two kinds of reinforcement learning that it talked about most frequently, reinforcement learning with human feedback, which is where we're having humans do the classification, and reinforcement learning with AI feedback, which is where we're using AI to do the classification, like even in the activity itself, this is this is in there. And then all of our other techniques for controlling the output of generative AI today are just wrapping content moderation techniques around either the input prompts that people send for the model, or the outputs from the model in response to England's it's all the same stuff. And so thinking about how we can do this core task better is relevant both to social media, but actually also deeply relevant to competitions around AI safety. And you can see this in the Google, Google Gemini,

22:58
which was released from my point of view, almost certainly, either some poorly thought through alignment instructions, or some poorly thought through moderation of the moderation and modification of the incoming policy, to try to correct your models and the model itself. It's simply failures of classification technique and a lack of nuance in doing that task.

23:20
As an example, chat GvG when we first launched, it wouldn't tell you facts about sharks. Because we had taught the model that violence was not good. We didn't want to help you play in violence. And we also didn't want it to graphically describe violence, wildly over rotated news and got it bears are canceled, no more bears can't talk about bears, which is perfect example of this sort of classification of responsibilities as systems.

23:47
So setting all of that context, generative, generative AI, actually, I think going to be very helpful here. US used properly, it is possible to forward to exceed both humans and machines under existing real

circumstances. And by used properly I mean a very specific thing. So naming generative AI is in a lot of ways, distracting. For this purpose. It's more important to understand it as language parsing ai, 3d ai, we have machines now, that can do something that functionally is equivalent to a human reading a document and responding to what it said, which means we have a machine that can directly address the core activity that a human moderator is doing, instead of merely producing a result.

24:41
And I'm not speaking theoretically, this already works shockingly. Well. One of the first things we did internally with GPT four, when it became available to us in August of a couple of years ago, was trying to figure out how to use it to do content moderation and within a week or two,

24:59
me and a call

25:00
A lot of other engineers were able to get to 90% plus consistency with my decisions with the model, reading a document that you any of you could read, and following the instructions that are provided in order to classify. And things have only gotten better from there. Okay, I published a blog post about this middle of last year about using GPG. For content moderation, there are multiple startups pursuing this path. And it's something that I've continued to work on at Stanford with somebody who is particularly interested in fine tuning smaller models to be able to do this because the smaller you can make the model, the more broadly adoptable it will be doing content moderation with GPT. Four is a little bit like going to a grocery store in a Ferrari like, you get there, it's very expensive, and most people don't have one. And so building a smaller, more compact, more usable, more broadly accessible system seems to us to be pretty important. When I say you can in fact, use these models to read and policy text, follow it and classify content, that is not theory that is already

26:04
and used in this way. DLR is directly address a number of the problems with human moderators, right? Their short term memory is already better than the largest models have context links are hundreds of pages of text. And so you can load tons of information into the model for making a specific decision, their long term memory is or will be more reliable, using things like databases plugged into the model to be able to give very exactly we call them large amounts of information.

26:33
They don't get bored, they don't get tired, they don't lose focus, they don't seek better jobs, they don't experience trauma, right, which is a pretty important part of this, there's, there's I think we have no moral case to be made here as well. We can reasonably expect them to record what they did, and why they did it according to the text as they understood it at the time, every single time they make a decision, and store all of that information, which helps with things like the requirement for expert ability that is embedded in a lot of European law. They're also way, way, way, way faster. Even the largest models are much faster than people even doing this very cumbersome policy texts driven process. And then on the flip side, the LA Times are better than existing machinery, because again, they're directly doing the task. And so they produce responses that are suitable aren't at least feel meaningful, right?

There's a broader, like, philosophical question here about whether or not they're really reasoning. Honestly, for these purposes, I don't think it matters. Because they are producing reason shaped answers in language. And those reasons shaped answers can be used to debug the decisions the model made, by changing the instructions you gave it. So when you have a model, make one of these decisions, if you don't agree with his decision, you can simply ask him why it will make a bunch of words that you, whatever you sort of think metaphysically, those words mean, they are useful for understanding how it was functioning. And you can incorporate that feedback back into policy texts produce change behavior, and this works really, really very well. And so it's functionally equivalent to an explanation, in the sense that it is a word shaped response that helps you understand what happened and why and do something about it. And so at least to me, a lot of this sort of hand wringing around whether or not this is true reason, is relevant and is functionally for this task in shorter. That's not me dismissing those concerns in a sort of longer term, more AGI focused way. But for this purpose for something like GPT, four,

28:41
it's neither here nor there to a very great degree.

28:45
I'd also point out that when you're dealing with really any people, but certainly people in a mass bureaucratic system, I don't understand why anybody does with a two either. And it's very, very difficult to sort of get that recall or get reasons that need anything out of the systems that we have today. So it's also not super clear to me that the alternative is really well thought out clearly described reasons.

29:09
I don't want to stand here and sort of make the case that this is magically going to solve all of our problems. So please do not take me as saying that. So first, the systems themselves will have flaws. They're going to make mistakes. Some of those mistakes are downstream, or some of the sort of language limits that I talked about earlier. But some of those mistakes are simply going to be errors are problems with the performance of model, they're going to have biases. There's been a fair amount of reporting about this already, in the use of these systems for things like hiring decisions, or other kinds of adjudication decisions. That's very real. I'm not minimizing any need to work on those problems. But I wouldn't say that at least those are static engineering shaped problems. Instead of the situation we have now where all of the individual moderators also make mistakes and also have biases.

30:00
But who they are changes every nine months. And so understanding correcting for controlling those biases is essentially I think, possible at present, because it is this sort of roiling, massive chaos, simply pinning down the biases to a single set of them, such that we can start to try to study understanding an engineer account for them feels better and more tractable than where we've been, where there's essentially, like, ever turning cauldron of biases that is never static and therefore cannot be stabilized.

30:34
I also do think that this is going to circling back around to my earlier points, that classification is more important than values. I weirdly, think if I'm right about this, we're gonna have more fighting about

values, because we're going to be better at doing the thing. And so what the values are going to start to matter more. And I think you're going to think you're already seeing shadows of this in some of the sort of woke AI, cultural war stuff that is starting to creep into Ayalon conversations, and some of the reaction in Germany. So oddly, as we get better at doing the activity, we're just gonna fight more about values.

31:10
That's it, I do think and I sort of like shaded this earlier, I think it's morally urgent that we figure out how to do this even even given all those flaws. Having people continue to do this work is is bad, right? And there's been a lot of focus on

31:26
ways in which the work conditions can be made better ways in which paid to be more equitable brakes can be given preventative techniques for controlling wellness, all those things are good, given no alternative. But in a lot of ways, they seem to me to be questions of like engineering a better radiation suit, when maybe we could just have a robot do it instead, and not worry so much about radiation protection, right? Like, the problem was radium girls making watch faces wasn't just that they were looking at the radium, it's that they will paint the watch faces with radioactive material, which is not a safer, good idea. And I think getting to the point where we can relieve the sort of direct coalface and labor here from humans is an actively good thing, even though it is fraught. I think that's actually doubly true for marginalized groups personally, right. Part of the sort of perverse shadow of the request for more cultural context being injected into moderation is it's essentially a call for the enlistment of people who are victimized by speech in the controlling of that speech to begin with, which is perverse. When you think about it that way.

32:31
This will mean I think job losses, particularly at BPOS.

32:36
But again, it's not clear to me the job losses are first a bad if the jobs themselves are dangerous, toxic and not conducive to human thriving. I also think it will be more jobs overseeing these systems on on the flip side.

32:50
So even with all those caveats, this is a really big deal. If you accept the case I have made to you not just because it's not just going to mean we are going to lift and drop AI in a few minutes from my moderators, right? It will change the kind of systems that are viable to have it opens up new

33:11
possibilities of moderation.

33:14
Things like super deeply personalized moderation become more feasible, ambient moderation becomes more visible. I think, in the future LLM Harvard systems are going to allow things like Siri to

prevent your grandmother from being paid, butchered over the phone, which isn't, like inconceivable thing to try to do right now. But seems very possible in this sort of future in the same way that deeply personalized moderation filters seem possible. It is utterly transformative of the policy drafting process. Right now, a lot of content policies, basically astrology about how moderators will work will react to the words that you wrote. And you're sort of guessing, because the update time is so long and retraining is so cumbersome, that you can't really do empirical testing of the outcomes of your decisions, feasibly, these systems respond instantly to your word changes, which means you can actually test different versions and approaches to things and see how that produces different outcomes, which is revolutionary in terms of the policy process directly. I think it also is going to open up new policy, VISTAs not just new processes, right. Right now we have generally global moderation standards on the social web. Because frankly, it's too cumbersome to do nation by nation monitoring for anything but the most sort of large scale blocks, that may no longer be true, right, we could potentially start to think about really localized or regionalised moderation standards.

34:43
And then similarly, like different moderation philosophies that no one has ever really, to my knowledge seriously trying to engage in at scale become possible things like deeply intersectional approaches to moderation, which no one has ever tried to do because it's just like wildly COPPA, somewhat impractical.

35:00
It might become possible with these sort of tools.

35:04
A bunch of those ideas were probably bad to be clear, like, I'm not saying all of the things I just said should happen. I'm saying they're now not impossible. And there will be other better ideas that are now not impossible, which is going to be very interesting. Similarly, changing the comprehensiveness of our moderation technology will change the kinds of platform designs that are available.

35:26
There's been a lot of discussion of network effects and social media. Yes, I think that under discussed aspect of the reason you see a lot of centralization in social media is how annoying, moderation is to do at scale, I have been very skeptical of federated solutions simply because I did not understand how that was going to work at the level of Mastodon except scalable to Facebook size, these sorts of systems might actually provide the ability to make that IO system work. Just as an example of the sort of consolidation I think you naturally see due to moderation. Think about Reddit and the power mod situation. Reddit is technically the sort of flat federated system. But the reality is like a few 1000 people, moderate half of Reddit, because it is in fact a full time job and has caused a bunch of concentration in the sort of bureaucratic processes to do that moderation even in a system that is designed.

36:20
So I think that's really, really interesting. Push for the extreme. Why are we even talking about post moderation at all? If I'm right about this, why aren't you ending up in dialogue with the textbox, you're

trying to write in about whether or not what you're saying is constructive? Again, maybe creepy, maybe a bad idea. But a possible idea. Now, I think there will be more versions of that.

36:43
These systems are also going to create new kinds of abuse, right? Ultimately, this technology is technology for sorting things into piles, regardless of what your plans are, and why you want to do that. So it is going to be useful for things like censorship, and surveillance, it's going to be useful for things like Jahmai, the virtuousness of these tools is simply a product of how they are used. It's not a product and the tools themselves, I think we can easily see the law start to try to specify exactly what your content moderation prompts and standards will be.

37:15
That is probably a bad idea. But I suspect we will see some folks attempted at some point, as it becomes more and more possible.

37:24
All of that said, though, I think this is coming, no matter what I'm really quite sure that some version of this is going to come to fruition. And so I think we all have an obligation to sort of embrace it and try to figure out how to use it well, now, so that we're not left on the backfoot when it becomes increasingly prevalent. So with all that said, a hope that was convincing, and being just want to leave you with an even more provocative question, which is, what would the internet look like if we weren't terrible at content moderation? The internet has been assumed to be a sort of semi anarchic space, there's been a lot of discourse about how that has become less true over time. And some I think about wistful moving for a freer version of the internet among certain quarters. But really, we're still pretty bad at content moderation. If I'm right about this, I think it might start to really seriously change the dynamics of how the web could work in ways that I think are really hard to sort of get our heads around. Maybe bad, maybe good, but we're sort of noodling on insofar as.

38:33
All right, thank you.

38:40
I'm just gonna start out with a couple of related questions from folks online. But if people have questions in the room, just fly them my colleague will be running around the mic. The first question

38:50
deals with a you mentioned sort of the ability to retrieve the reasoning of outlines when they're making these decisions. Yeah. How does that intersect and interact with the possibility of hallucination? And then also, policy statements are often intentionally kept high level and broad? And how well can MLMs capture the socio cultural context from broadly that are that are relevant to implementing broadly described policies? And a related question to that second point.

39:19

Words change for the meaning of words change pretty rapidly and increasingly so. Especially with regional dialects and how does that how does that address or how does that interact with the subtle problems that you're describing? Yeah, great question. So on their sort of recall reasoning,

39:38
really specifically, what I'm proposing there is asking the model to actually print a reason that you then store every time it makes a decision. The hallucination question is a little bit separate from that and mechanically, in the context of how we've used these so far, you're actually feeding the model the policy text every time as a prompt, and that is that sort of grounding in

40:00
In this specific document you're using it to make your asking you to use to make a decision is very helpful, not perfect, but very, very helpful in controlling hallucination hallucination is often worse when you're asking the model to sort of remember what it knows. Whereas if you're asking me, What does this document say, that is helpful in reducing hallucination? That is, it is absolutely an issue. Again, though, people also make all kinds of really weird choices. And so the question is not, is that an issue or not? It's Is it better than the status quo? And will it continue to improve? We have spent,

40:34
you know, nearly two decades and billions of dollars trying to squeeze more juice out of the human moderation lemons, and I just don't think there's any more juice in those lemons. But we have some new lemons. And maybe there's maybe there's just

40:47
you had two questions, the the changing meanings of words over time. Yes, that's happening rapidly these days. Yeah. So that definitely is a problem. Part of that is something that you can address with the sort of database and long context, stuff that I was talking about, where you simply tell the model, here's how you will understand words to be understood for this task at any given time. And so there is actually I think, in some ways, more ability to direct models there, relative to people who again, also have that sort of freshness problem. So there are very serious advantages, not keeping everything on this page.

41:24
And we have a question on the room. If

41:27
so, if you've registered developer clients this year, and I used to work on comms integration issues before I got the law school.

41:34
You know, I think the time horizon or what people were concerned about with like, social media and content moderation, I think like when you were first starting, like Facebook, way back in the day, probably a lot of folks on the Hill in particular weren't as concerned with content, moderation, they were thinking about as much I think you fast forward to today, like it probably would be. I don't know, probably

most legislators, folks on the Hill know a little bit about Secretary at least they heard about right, it seems the time horizon with like, artificial intelligence, and specifically generative AI is much shorter. But the fear and my concern with this is also my short show that the the rising concern happened much faster than I think the content moderation issues. And I'm wondering, to the extent you're speaking to like policymakers and regulators, what do you what do you or if you have thoughts for them? What do you say to them about like things they shouldn't actually be concerned about? Instead of like, maybe some of these like, culture war like boogeyman type issues? Yeah, there's been a lot of focus in the sort of focus on AI around longer term questions, there's been a lot of discussion around bio risk and similar issues. I actually think I'm not saying those aren't important. But I actually think we sort of missed the boat a little bit on some of the shorter term content related issues, less in the text model space, but in the image and video model space, open source image models are able to generate

42:55
large amounts of cesium and it's really, really difficult to control those. There's been some interesting work from SEO around this topic right now. I think ca regulators are over focused on long term versus real, some relatively specific short term problems that are happening that honestly shouldn't be super controversial, from a cultural point of view, because they revolve around some pretty core abuses that you still do have social consensus on.

43:21
Another question on there?

43:25
Go ahead.

43:27
Actually, next, okay. Seems like one difference between the internet that work fairly well, 20 years ago, and the one that doesn't work that well now is the scale that you've talked about. But maybe the answer here isn't to try to, you know, throw more AI into the column, but to

43:45
back away from the large scale. And back to a time when you had many people on many different platforms, each of which had its own idiosyncratic, moderation policy, and people were just sort themselves to communities that they felt they belonged in.

44:03
Yeah, this is essentially the Federation proposal as a solution. I am skeptical of Federation as a solution because it doesn't address the worst kinds of harms. Because some, like space owners are going to be bad actors. Though, insofar as we're dealing with harms where the issue is, I don't want to be exposed to a particular kind of content. Sorting is great. And we should encourage it. This isn't me saying it's like a bad thing. It doesn't deal with the kinds of harms where the problem is content that people see about me, right? A problem in the context of NCI, non consensual, intimate, intimate, intimate imagery isn't that I saw my news is that you saw my news, right? And so there's no really great answer sort of absent

a central authority for those kinds of problems. And those are in fact the worst harms that happen in the content space nearly universally.

44:57
From

45:00
Hi, my name is Matilda from Big Vision Business School. I work with the UN on online harms regulation, particularly Sisa. And with MIT on AI voices. So my question to you is, you have mentioned neck neck, there are many developments currently that are leading towards anti encryption. And we know first what what happened with C seven regulation after WhatsApp.

45:28
And an encrypted, how, and now we're seeing the same with

45:33
messenger, right? So how do you see these these developments that are leading towards a minimization or end to end corruption and all of these other similar developments affecting the motivation.

45:49
So this is where the sort of like wild possibility stuff gets really interesting. And I wish in a lot of ways, we had a healthier conversation about the trade offs between my mind

46:01
I wish and a lot of ways we have healthier conversation about the trade offs between privacy and safety, because I think they are real and they're uncomfortable. And that leads us to sometimes try to dodge them. When I say that, I feel like when people say they say that they often sort of steer into an overvaluing. privacy versus safety. That's not necessarily what I'm saying. I'm more saying we need to be honest with ourselves about difficult trade offs and the downsides that they carry in both directions.

46:30
I also think, though, that some of the stuff I'm pointing to might have really interesting implications about ways to compromise, you could conceivably, if on white, get an AI moderator to be really, really excellent at interdicting certain kinds of super problematic material, and lock it inside of an encrypted space such that it's never doing recording, but it is doing interception. Again, maybe a bad and creepy idea, for now, not impossible idea, in a way that previously that was sort of outside of the design space to even consider. So I think there's some interesting implications for even that tension of some of this, if I'm correct about where it leads. Great. A couple of questions from online. One is asking you about how the moderation tuning and red teaming of LMS themselves has parallels to and differences from monitoring social media, of the gender of AI companies learn the lessons they should have. And then a couple related questions about using using generative AI in linguistic and cultural context other than English and outside of the middle of the world. So in terms of the similarities and differences with social media moderation, so content is content, right? At some level, the products here are not, AI is not

capable of making a picture that a human could not make. Conceivably, it doesn't work quickly. But you know, we have Photoshop, we have been uploading billions of us, I've been uploading things onto the internet for a decade. So functionally, all of the photos that can exist have been on social media. And so in that sense, yes, they're very much our lessons to learn. And there has been a movement of trust and safety professionals with social media backgrounds into the AI space, I think, because the companies do do realize this. There are other parts of it that are unique, and Red Teaming is actually one of those areas where it's particularly unique. And that's because the the question with red teaming, which is the process of sort of prodding the model to try to get it to do something that it shouldn't do is a multi turn conversation. It's not just an interception, or a single sort of past question. You're, you're essentially trying to convince slash bully a robot into generating something for you that its creators trying to teach not to create. And that process is very different than sort of testing social media systems, because that means evasion in those systems. But there isn't that sort of convincing process in quite the same way. And that is, is interesting. I do think the AI companies, so it's hard to generalize about them, right, Google is a very different thing than open AI or anthropic as an organization, because it's a massive behemoth that directly owns a bunch of social media to open AI. And anthropic are both very much startups. And so have they learned lessons and social media is sort of a weird question to ask about that group. But I do think that we are not starting over in quite the same way. There's still an organizational capacity building question, particularly on the startup side, but there has been a lot of learning. In terms of low resource languages. It's actually every question that I skipped over my notes, but I'm

49:34
supporting low resource languages and cultures with traditional moderation solutions is really hard. There's, there's as many people in Massachusetts as there are in Serbia, and that's difficult to staff for. And Serbia is frankly, like a fairly big example, right, as you get to smaller minority languages, that becomes really, really hard.

49:54
LLM 's aren't going to magically solve that. But I do think that if we intentionally train

50:00
For them to be capable of it, they will eventually prove better at load balancing, because you can support a capacity to understand those languages in detail better than you can, when you have to get a group of humans where if you want 24/7 coverage, 365 days a year, you're talking about at least 21 people to cover, you know, like a decent a decent amount of decent content. And so that gets difficult to sort of justify if you're not seeing enough content in a particular language to actually justify that many jobs. So I think there's hope there as well. But it will require focus, it's not just gonna magically happen. Just another round of online questions, and we'll move back to the room to about sort of the commercial implications of what you're describing. One question is asking for your bowl and bear case for when a company could implement 100% ad powered moderation that's fast, cheap and accurate. And another question is asking you about the way in which content moderation has become a competitive advantage for some platforms.

51:02

Do you think the development of AI moderation models will commoditize content moderation, for better or for worse?

51:10
Taking the second one, first, it's already commoditized. In a lot of ways, the fact that it is a mass produced commodity is sort of the thing that I'm pointing towards. And I think a lot of our problems, I do think we are going to see trust and safety as a field and moderation as a part of that move in a direction that is more similar to cybersecurity over time where there are more external vendors that are providing solutions, instead of everything being homegrown and roll your own, which is the case at today's very large platforms,

51:40
mostly for historical reasons, because the external tooling and solutions just like didn't exist, so we had to build them because there was no one to buy them from. I suspect that will change over time. And I think you're already seeing some of that, in terms of the bull and bear case, my, my bull case is like now, there are startups that will do this for you. And at least if you were focused directly on text, and you have a

52:05
flexible definition with the or flexible relationship with the definition of cheap, they can do this pretty well, with automation, using a lens today, once you talk about sort of images, that is also possible, but probably not even any reasonable definition of cheap,

52:23
bear, it's less about possibility, actually, then adoption by bureaucracies are slow to change, the uptake of new technologies into the survey processes is slow. It's always slower than everybody thinks it's gonna be. So I've been using five years as a number, but that's kind of because I don't have any idea what the future is gonna look like more than five years from

52:45
now for a really good father reason.

52:49
super interesting. Thank you. So

52:53
you said that you think that one of the problems with these AI is that they're relying on religion? And so there's kind of like a cap, which is the cap of the highest performing crowd workers, right. And then many of the benefits that you talked about for using gender divide in this context were more of like, they don't have the same problems of exhaustion and lack of memory and stuff as humans. Yep. Do you think that the first set of problems of the capitalists like the best crowd workers will also be removed such that AI is able to exceed human performance even in like the most focus and best human context? Or will that always remain, we only will see the benefit from the kind of soft replacing of the soft? Because as

53:32
far as I suspect, it's going to exceed us for these kinds of tasks, even under our best conditions, I am positive it's going to exceed us under the actual conditions. is I think the way I will, I would think about that.

53:45
But that is mostly a statement about how difficult and mediocre the actual conditions are. I don't think it's a super high target. Let's put it that way.

53:59
Another example, I have had it changed my mind in working on policies with one of the systems that I was using to label content and asking you why it made a choice that I thought was the wrong choice. I've had it respond to me with an answer that was better grounded in the policy text and the reasoning I was using, such that it was incorrect, if correct is defined as fidelity to the document I gave.

54:23
Hi, I'm from the University of Toronto, and I'm a law professor there. I'm super sympathetic to everything you're saying about some of my colleagues at UT in the tech space have been telling me how good things like GPT four are at classifying content and to

54:39
sort of say that because I want to kind of push a little bit on on some of what you're seeing, not to undermine kind of direction going back to maybe open up another problem. So you gave some examples where people just like that the amount of disagreement on human labelers is huge. And that's a problem.

55:00
problem. And you also talked about what we have values in the policy. But the problem is at the implementation level, and one of the things that you know, is proposal does is it turned out implementation into sort of this sort of human problem into an engineering problem. So here's the perspective from law. While we deal all the time as analogy, we hear all the time with these kind of broad policy, like statements, rules, wherever. And we have to apply them in the real world. And the application is deeply normative all the way down, right. So there's values and then there's implementation implementation simply depends on do we have enough information and other things like that. And so we have these in law is our offense, like deeply contested. So it's not just that we can't agree, and there's some ground truth, we can't agree. And it's normative all the way down. So what law does is it doesn't give you the right answer, it settles the matter for a period of time, right. And that's open to change. And in certain amount of for a period of time, we agree to that if we agree that the process is legitimate. And once that trust goes away, that we've got lots of problems as we can all do. So what interests me about your proposal is, you know, there's always problems with the human moderating and have a technical model. And we can learn how that model settles the matter.

56:21

But there's no ground truth. So we still need to answer the question about the legitimacy of that. And so that seems to open up kind of interesting possibilities, like how could we do that? And I'm just kind of curious, wait, wait, are you eating good? Yeah, actually, this there was a question that you asked online that we skipped when I asked you for recall around the sort of policy standards being very vague. That is true of the standards that are published. And I'm gonna never navigate this and answer your question. That is true of a public standards that most companies

56:55
publish, the actual written standards that are being used to make these decisions. Those things are basically like PR for what the actual rules are, the actual rules are very, very, very, very detailed, because they are trying to solve this problem of inter human coordination. So they get extremely explicitly exhaustive. And you do need to write in that explicit and exhaustive not high level way to get good performance out of these LM so that it is a very sort of very, very concrete

57:23
writing process. You're totally right about legitimacy. I think there's some and this is not an area where I'm an expert. But there's some really interesting experiments going on around using the models themselves to gather people's preferences, and how to sort of combine them to create proposals for policies of various kinds, and probably expensive, interesting experimentation here of using sort of an input process to figure out what those rules should be. And a lot of what is cumbersome about a, and again, a non expert understanding, but a lot of what is comprehended about those kinds of processes today is because of how manual they are. And actually using sort of AI assistance systems to sort of gather those things and more dialogue based way might actually make doing them at scale, more practical, which is really cool. So there's some there's some very interesting sort of directions there. We're gathering input as to what the goal should be, as a way of establishing legitimacy might also become easier under this system, not my area of focus or expertise. But I think it's a valid problem. And I think a really, really interesting one.

58:29
And also, your point is totally correct about sort of there not being a full difference between values and implementation. Don't please don't read me as saying there is it's more that the act of performing the instructions, whatever they happen to be, has a technocratic component, which we can be good or bad at. And also there are these values things intersecting with that. And our technocratic inability

58:56
is a big part of the problem as the system exists today, particularly because at least at the mass scale we're talking about, which I can't figure a way for us to get out of, and some of the processes that law uses to make these decisions simply will not scale up. And so we're sort of stuck in this more technocratic mode. Thus, the focus on improving the technocratic mode.

59:22
Just another question online, but I think there's a few more there as well. One question is that you seem to imply that as content moderation improves, the worker trust and safety will be focused on updating content policies, what would be your guest, who decides these policies are related to what you're just

saying, will be platforms themselves, users or platforms? We were talking earlier about platform assemblies or citizen assemblies. So I guess the question more broadly is about mechanisms for deciding those values as implementation becomes increasingly automated is potentially I think it's going to depend on so so a couple things. One, even with more automated implementation, you're going to need human oversight over

1:00:00
He's sort of frontline decision clouds, decision making agent clouds to make sure they're still dialed in. So there is very much going to be work, I think that work is going to move from being frontline labeling work to sort of quality oversight work, in addition to the policy work.

1:00:16
I also in terms of who that decides the values, it's a really interesting and difficult question, because this is an adversarial space, right. So at a very high level, it's pretty easy to get, like, what a sort of have buy in systems to create legitimacy. And we should do that. And that's great at this specific level of exactly which things violate, which rules and rarely need to tweak language, that becomes harder to run consensus processes for simply because the amount of stuff that happens every day on our internet of three and a half billion people means that we're constantly responding to attacks, and you do need that sort of flexibility. So I think that's going to remain contested. I think the obvious trend has been towards more government intervention in this space. And I don't think that's gonna change, because it's a sign of power and will become more accessible. But I don't think it's going to like signs of one neat, great resolution, wherever it is happy. On that point. There's another question from online about how the regulatory space seems to be moving towards requiring some sort of human in the loop or human reason or explanation or explanation for content decisions. So the question is about how your proposal or what you see as the future, how does that intersect or interact with regulatory movements? Yeah, I think that this, I think the rise of AI problematizes, a bunch of the European moves here, because they assume that a human answer is going to be better, more fulfilling and more correct. And my basic thesis is that's wrong.

1:01:44
And that we are very, very quickly going to end up in a world where some of these large model powered systems produce better and more fulfilling, feeling more detailed, more accurate, with accurate and consistent answers, and probably have to change the law or your we'll just have a weird version of the internet, probably some next.

1:02:06
Like, we'd start with weird version of the internet. Now.

1:02:10
I have one last question. I thank you for your time. So it's super interesting. Do you have any concerns about the idea that AI tools are also used to generate images, text or otherwise to confuse the system or overwhelm it simply for the purpose of moving the needle on the norm. So you could think about it, you can basically keep gaming it until you push the norm to the place that your particular thing you want to say, or way that you want to articulate an idea or misinformation, it becomes allowable, if you know

what the rules are, and test the system until it forces it. So I guess I keep thinking, yeah, it's faster. But what happens if you also have the creation of text that's faster? Learning faster, figuring out what's, you know, what's being rejected? What's acceptable, keep moving the line of what's acceptable to where they want to go? So a couple of things, one, most of these systems don't do online learning. So simply interacting with them isn't going to change what they do. And you can decide as a system designer, am I going to allow that or not? And you wouldn't for for that reason. The other thing, though, is I flip it around, they're gonna do that. So we need to learn to use the robots to help us because we're not going to be able to do this quickly enough, given the scale of what is going to happen, and given the existences. So it's like, less, will you create that arms race and more like the arms race is upon us better start building battleships that's never ended badly? What could go wrong?

1:03:30
Couple more from online, we have some more in the room as well. We're technically at time, but I think there's generally no leave until tomorrow. Perfect.

1:03:38
All right. Well, a couple from online then. One is, is anything that you're saying relevant to data annotation, and get to solve some of the problems there? Yes. Yeah, I mean, what I'm proposing is a text directable classifier. It's just the sorting machine for content that sorts based on a document you wrote at route. So if you have a problem that's about sorting content, and you can describe your feelings about how you want it sorted as a set of instructions, this is useful to you, for a lot of different activities in the same way that it's applicable to both AI alignment.

1:04:14
Question. So I have a general question. So what is the difference between the monitor content moderation and the candidate censorship? Because I think I'm a little bit

1:04:26
confused with that. Do you have any, like dividing line 10 misleading or? I don't think there's I don't think there's a technical dividing line to me. I think we are careless with our use of the word censorship. I think censorship is one of government does it and maybe we sometimes even sometimes do overlap. You know, the, the issue that covered I think,

1:04:49
Tariq. The techniques overlap because again, it's just a question of sorting. But Mark Zuckerberg does not own any prisons, and has never put someone in them and I think that is as important

1:05:00
distinction, right? So from a technique point of view, yeah, I sort of flagged at the end of the talk that these techniques are definitely usable for malicious purposes by governments, whether well intended or ill intended. That is not disentangle from the existence of the technology from a technique point of view. If you're sort of hinting more towards jawboning or the pressure that gets put on companies, by governments, that will remain a problem. It's a problem today independent of the existence or non existence of these techniques, and is a thing we need to continue to work on, as a society.

1:05:35
Have more questions online? One is about the future of volunteer content moderators potentially, on places like Reddit,

1:05:45
do you think that those sorts of roles will still remain those sorts of spaces will continue to exist? With the sort of developments that you're describing, I think they'll change I think your role will move from being the actual direct content grader to being someone who's dialing in and overseeing a version of these systems may be provided by Reddit themselves, to actually help you do this such that you can scale that up with less direct human labor, which if you want dispersion of power at who was overseeing subreddits is a good thing, actually, because it'll allow normal people in sort of their spare time to viably moderate larger communities without having to professionalize.

1:06:25
I really liked your comparison between any development of industrial machine versus an artisanal machine in the case of content moderation. You've been in and out of the of the platforms and the platform and i Is it safe to assume that

1:06:46
these companies are understanding on content moderation, based on

1:06:53
the legitimacy question proposed here?

1:06:57
I think that's a version of the

1:07:02
benevolence.

1:07:04
thesis, I sort of hit on the front, right? Like, I, maybe some of them are, some of them aren't. I don't, I don't think there is an amount of money you can spend with prior technology that would produce a result we all really loved. Like that's sort of the crux of what I'm trying to get across is like bad content moderation, because we're bad at content moderation, not because we're secretly good at content, moderation and mailing it in to be jerks.

1:07:34
Also, yes, probably there should be more investment. And I specifically think more investment in learning how to do what I'm talking about. So we stopped being bad content, moderation, is very important. And there probably has been an under response on that score, simply because a lot of large companies are giant bureaucracies that change slowly. And so I think a lot of this will get figured out at startups and smaller companies, and then percolate its way up, or from like special labs within larger,

larger companies. I suspect you'll see interesting things in this vein from Jigsaw within Google, how quickly that translates into all of the giant processes. Google runs as a bureaucratic question, more than anything else. Perhaps a fitting question to end on from online if if what you're describing it becomes the brave new world of trust and safety and content moderation, what are the AI skills, you recommend someone interested in entering the field developer build, what sort of skill sets you having a schematic understanding of how these models work, so that you, you don't have to be able to build them, but at least being able to understand how they work so you understand what their sort of tendencies and properties and meanings are, is really, really important, understanding that they're still ultimately just prediction machines, but they're predicting word outputs is helpful. Honestly, in the policy space, the same level of clarity of writing that is helpful for writing for like 10,000, guys that have EPO translates pretty well to writing clearly for these sorts of machines, because a lot of the places where policy writing fails, is when it moves into questions that are not discernible to the person who has to use the policy. Right. So moving into things like, oh, what were their intentions when they uploaded this, which is like doesn't exist from the point of view of a content moderator. Um, so a lot of those skills actually translate super well.

1:09:27
All right. Well, unless there's any more questions in the room. I hope everyone will. Oh, yes. Okay. Well.

1:09:35
I'm literally not leaving till tomorrow. I also know that

1:09:39
well, I have questions about some of the

1:09:43
let's see, I have a few questions, but I'll try to synthesize into one.

1:09:47
If you were to, if you can evaluate the up and coming startups that are entering this space. How would you evaluate essentially, the viability and pedagogy

1:10:00
going forward. And then also, of those, there are many of these problems. And just to give you a background, I also have 17 years of experience and interest in safety and overseeing teams, I am

1:10:12
concerned about some of the problems that I don't think we should be applying towards AI. And part of it is I worry about some of the issues of bias, inequity, calcification of the model itself, and not adapting to the space. And so I'm curious, from your perspective, what are the problems that you think we really shouldn't be applying AI moderation to? Or that we maybe need to do an adaptive hybrid approach? Like AI assisted but certainly not AI? deferred to?

1:10:40

And, and therefore, also, what should startups in the space be avoiding? Because they all say that they're trying to solve it all? I don't actually think that's

1:10:50
a safe way to progress. Yeah, in terms of evaluating startups, that would evaluate them the same way you evaluate a video, right, I would actually ask them to do the work, and see whether or not they're producing

1:11:04
the outcomes that you want. And you shouldn't be thinking through how you really put them through the paces in terms of what the actual set of stuff you send them is not just a random sample, the hard, weird edge cases kind of stuff. I prefer, I would prefer personally, startups where they're giving you a toolset for you to do it yourself. Not saying like, hey, plug this in, and we'll make all the bad things go away. Because that feels like pretty much giving away your your sort of control. Whereas a lot of the better ones aren't. They don't do anything on their own. They're like, here's a way to write a document. And we'll fire this at the outline for you in a structured way. And you can set up your rules engine. So it's a it's a system for doing the trust and safety work within not just like a plugin thing that solves all problems, which keeps you in control of any animal to assess whether or not it's working well, which is killing it before we want. As an insider in terms of where automation becomes dangerous. i There's an interesting question that you sort of got to in the second half of that, around how much this becomes pure substitution versus like side by side working? That's actually what I think that's an interesting question from like, does it go to full replacement? Or does it go to like, the actual interface moderators are working in just becomes profoundly different. And AI enabled, that could be a direction it goes, I suspect, it's a little of both. But it's just hard for some percentage of decisions, like, we're already in this world, right? Where classifiers make the very easy and very hard decisions. And then humans make the middle decisions, I suspect, like you're gonna get another band of the AI make decisions, and then the coal is going to be this AI assisted side of human decisions. I'm not proposing taking humans out entirely. And I do think human oversight over sort of the instrumentation of the model decisions you have going on will continue to be important for any foreseeable future.

1:12:53
That's a little like in the weeds for half an hour talk. But I do think that's pretty important. I don't know that I think there are any categories of decision that should never be made by a machine. I think the question there is who will achieve better results for the people impacted by the potential harm?

1:13:15
And so I think bias is an obstacle there that we need to account for a very important one. But to me, the question is, what is the most effective method to create that both like intercepts the horns were trying to deal with and creates the least amount of suffering for the people who are part of a process? And if AI wins, that man wins that and if there are situations where it does it, it does. You showing us the thing that is best.

1:13:41
All right, well, I hope everyone will join me in thanking Dave

1:13:50
and we'll be back here next week for another speaker series event. So people will join us either online or in person, but thank you again, everyone.