00:00:03:03 - 00:00:23:00

Thank you so much for joining us. Thank you to everyone who is joining us online, as well as lots of people in here without just now this afternoon. My name is Alia Hall and I work for the UK government here at the British Consulate General, just out on the road at Kendall Square. And I'm the US director of the Global Talent Network.

00:00:23:02 - 00:00:47:19

So I'm here to try and convince you all to come to the UK to live and work. But I'm also here and delighted to be able to introduce today's discussion on frontier AI and safety. I want to start by saying a very quick thank you. To CSAIL to Glenn in particular, associate director for Global Strategic Alliance, {indistinguishable}

00:00:47:19 - 00:01:15:13

{indistinguishable} MIT international science and tech commissioners program delighted to be here today with our panel of representatives from the UK's AI and Safety Institutes, Queen's University Belfast. Robin AI and Harvard. Thank you again to everyone for coming. If your here in person please fell free to take a look at some of the merch in the back we got from pounds governments finance.

00:01:15:15 - 00:01:32:02

And those postcards, rates and information about living working in the UK. Some of the options for you there. Without further ado, I'm going to pass it over to Kwamina who's going to be moderating this discussion. Thank you so much.

00:01:32:04 - 00:01:56:20

And thanks very much. And and also a big thank you to CSAIL for, creating this opportunity. I want to thank you to all of you for for coming. It's really quite motivating to see this amount of enthusiasm and support for AI safety. I think before November of 2022, it was all just kind of pipe dreams and, and ideas about what I would be capable of.

00:01:56:22 - 00:02:22:17

And I think with the release of ChatGPT, the whole world's kind of paid attention to this and started looking at this much more seriously. There's been a lot of explode about where we're going from here. There's a lot of optimism about what frontier AI will be capable of next year. In five years and ten years. But one thing that's quite clear is that for those for those benefits to accure we need to make sure that we have the safety in place to to get there.

00:02:22:19 - 00:02:43:10

And so everybody here on this panel has done a lot of really interesting research and work around, dominance in AI safety and so I'll go around and introduce them. It's seated order which isn't quite the order. But start with, I Dylan , Dylan Hadfield an assistant professor at Harvard-at MIT. I'm sorry. Oops. They were all squished up.

00:02:43:10 - 00:03:04:19

So we understand We want to give you a quick, please. So. Hi, everyone. My name is Dylan, and I run the algorithmic alignment group here at MIT. I'm also a part of CSAIL. My research really started thinking about AI and alignment and AI safety. Sort of in 2014 or so. And my advisor came back and said, hey, I think this is an interesting problem.

00:03:04:19 - 00:03:24:14

And I looked around that everyone else had a thesis topic. And lo and behold, a decade later, here I am, with in my group, we do a lot of work around trying to manage downside risks for many AI systems. We think a lot about policy implications of our work, and we try to find, technology interventions that can facilitate, the kinds of policy goals we might want to hit.

00:03:24:20 - 00:03:44:20

We're also just interested in studying these systems and building them out. We do work in sort of preference learning and Bayesian methods. Now we also think about adversarial settings. And we've been looking at interpretability and, broad sort of safe agent designs. So if you're interested in chatting about any technical things, you can either come find me afterwards or I'll highlight

00:03:44:20 - 00:04:04:03

That like this over here consists of, like, a whole bunch of my students. If you want the real story, just go talk to them. HIma Lakkaraju an associate professor at Harvard. Got it right this time, Right. Hi everyone, nice seeing you all here. And thanks so much for joining us today. I'm Hima Lakkaraju

00:04:04:08 - 00:04:27:16

as Kwamina said, I'm an assistant professor at Harvard. I work on similar topics to Dylan. I would say all of us here are working in some capacity or the other on AI safety. My, sort of tryst with this topic started, I guess, almost a decade back when, the sort of the turn interpretability meant rule based ordiance

00:04:27:18 - 00:04:54:12

I think we have come a very long way since then. Until today, you know, we are dealing with interpreting models like LLMs where as past parameters as, like a small language model, right? So, yeah, I've seen the gamut of the evolution about these topics over the years. But in addition to interpretability, I work on {indistinguishable} privacy, and also the intersections between these topics.

00:04:54:12 - 00:05:16:06

And more recently, me and my group have been focusing on understanding the gaps between AI regulations and research, identifying them and finding approaches to bridge them both in a technical way as a and also, in a way that policymakers can sort of address some of these challenges. Right. And so one of the folks from our group are sitting in that row.

00:05:16:08 - 00:05:37:22

So, yeah, you can also talk to them or talk to me after this. Thank you. Thanks HIma. I will go to Joseph Enguehard, LLM researcher at Robin AI. Hi. Nice to meet you, I'm Joseph Enguehard and I live in London. Actually used to live in Boston a few years ago. And then I made a move to the UK, and I've been in London for a few years now.

00:05:37:23 - 00:05:58:01

I've worked in a few different startups. Currently I'm working at Robin AI, which is a legal tech company. So we work around legal contracts for various clients, which can be either law firms, or finance firm, for instance. And my interest in AI safety comes from the fact that we we need to make sure that this system is robust.

00:05:58:03 - 00:06:20:07

We are particularly concerned about statisticians, and related to this, as we really want to make sure that the contracts that we are editing with AI are drafting, searching for them, not, done the wrong way, because this can have quite serious consequences. So my interest in AI assisting is more on an apply point of view but, which is also very important.

00:06:20:09 - 00:06:51:14

Thank you. And last we will move over to Vishal Sharma. Thank you so much. Kwamina. I'm Vishal Sharma I am an associate professor in the School of Electronics Electrical Engineering at the Queen's University of Belfast
{indistinguishable} know that and that motivated me to work in the field for AI Safety

00:06:51:15 - 00:07:13:13

{indistinguishable}

00:07:13:15 - 00:07:40:19

Thank you so much. Nice. And then on last time my name is Kwamina Orleans Pobee. I am the head of engineering at the UK AI Safety Institute and we usually call AC for short because quite a mouthful And the AI Safety Institute is doing it's also our mission is to, provide information, necessary to evaluate and understand large language models on frontier frontier AI systems, for governments both UK and and world wide

00:07:40:21 - 00:08:02:03

As engineering-I'm head of engineering and also head of platform our internal tech team and which is responsible for all the various builds that we do to, to for that purpose that involves welding the station and scaffolding, the infrastructure to do the large scale, LLM evaluations. But, you know, we can do interactions to, to open into probing, into teaming, and any other techniques.

00:08:02:05 - 00:08:19:00

I'm also going to be both a panelist and the moderator for this. I'm going to do the awkward thing at some point of asking myself a question and answering it. So first I will not do that. And I think that's all of the introductions. So we'll we'll jump into the panel then. So the first question, I'll start, I think it's for everybody.

00:08:19:01 - 00:08:48:15

So I'll start here and go around. I want you to talk through what technical advance over the past year that has made you most excited. In terms of, avenues for AI safety. I'll start with you Dylan. So I think one of the problems that that I think we have not done a great job as a research area of really exploring and contributing to, is figuring out safety interventions that support open, sort of sharing of models and things like that.

00:08:48:16 - 00:09:16:05

I think the AI safety community very quickly went to like an anti open position. And so pretty recently there have been a couple of, works that are starting to take that idea seriously. And some of the, paper from, the center for AI safety, looking at tamper resistant fine tuning or know the idea is, can we take models and make it so it's hard to fine tune them or modify them to do specific things as a really challenging research area.

00:09:16:11 - 00:09:35:16

Really, really hard to really hard to see whether or not it will be successful. But these results were actually more promising then-I then I had hoped, and we're starting to see a bit more attention and effort in this area. And I think that's just frankly, really hopeful to see a positive result. I wasn't expecting to see that.

00:09:35:18 - 00:09:57:09

Go Hima, to me to say that, the most exciting thing that has happened in the past couple of years is probably the figuring out of putting a lot of building blocks together. Right. So I think, you know, all the basic building blocks that underlie models, that the whole world uses now, right? Like, you know, ChatGPT or GPT based models.

00:09:57:11 - 00:10:26:14

But building blocks were kind of there for a while. But I think the way in which those came together, those were put together, of course, you know, some a lot of it coming from the industry, of course. And then made mainstream, I think that transition, the way they sort of went from research labs and, you know, these are just for tech companies and tech organizations to now my grandmother getting excited about using, these kinds of models every day when she wakes up in the morning.

00:10:26:16 - 00:10:48:02

To me, that transition has been the most exciting development over the past couple of years. But that said, that has also, you know, raised a series of problems and questions that we don't have great answers for and we are trying to grapple with, on one side, you could look at it and say, well, if there's more bread and butter, we now have lots of interesting problems to solve on the other side.

00:10:48:03 - 00:11:09:10

It's also scary because, you know, you never know that the impact of these kinds of risks would be especially on a technology is being used at mass, right? Definitely. I'm going to jump in. But because, I think what I see as the biggest, very positive advances is quite similar to that. We're starting to see much more kind of standardization and, uptake of, of safety.

00:11:09:15 - 00:11:39:01

I think precaution. So, I think two or so years ago, you know, evaluation was in a very nascent stage now. Most countries and most companies have their eval teams out there doing those kind of evaluation, I think that kind of development of standardization and adoption is quite important. Let me jump over to you. Joseph and yeah, I think from my side

it's about, capability of reasoning of this model that are able to do very complex task, where before they would be able to just answer a simple question.

00:11:39:03 - 00:12:01:04

But it's also a challenge, in my opinion, because, the way we use this model, we can use, for instance, to find some specific, piece of the contract of a legal contract which sometimes for example, for NDA works very well, but a very similar task on another type of contract would work much worse, worse. And it's really hard to like guess in advance whether it would work well or not.

00:12:01:08 - 00:12:28:24

So I think one of the interesting challenge is robustness to make sure that this model is able to do similar tasks on similar data in the same with the same performance. So I think we've seen a lot of improvement there. There's also a lot more we can do. Right. I am going to talk a little bit if that's okay. {indistinguishable}

00:12:29:01 - 00:12:55:01

{indistinguishable}

00:12:55:03 - 00:13:19:02

{indistinguishable} fascinating stuff.

00:13:19:03 - 00:13:37:06

Well, and something which also escapes me {indistinguishable} So the other half of this question, which is, what do you think the biggest challenges that you know, facing from your area of work around AI safety.

00:13:37:08 - 00:14:14:04

{indistinguishable}

00:14:14:08 - 00:14:31:14

{indistinguishable} Joseph, I'd love to hear your side. Yeah I think I could just say about the changes we have around robustness, which is actually very similar to your problem, but on legal contracts, which is very different application. But we can implement that sometimes. And with that we work as expected.

00:14:31:14 - 00:14:47:23

And sometimes it won’t. And it's really hard to know when either way happens before using it. And I of course when you go to contract you can have severe consequences. If we send the wrong contract to a client. I can just imagine for autonomous driving, it can have severe consequences if the vehicle doesn't behave the way we want it to be.

00:14:48:00 - 00:15:20:12

So I think, robustness, yeah, it's the main challenge. I think it's important to make sense. Go ahead. So from our side you're working on like a few different problems of course, related to the directions. You know, Vishal and Joseph already spoke about I'll start with one problem, which is since we think a lot about interpretability, improving the faithfulness with which these models explain themselves has been something that we have been thinking a lot about, but also trying to come up with strategies.

00:15:20:12 - 00:15:42:08

For that to happen. I think that is turning out to be a very challenging problem. Been more than we anticipated, which is these models explain themselves, but usually what the explanations that they're providing don't really capture what they're doing underneath. Right? The explanations look plausible, but that's not what's really happening. What's happening under the hood. So that has been a big challenge.

00:15:42:08 - 00:16:06:10

And that's been a problem that we have been working on and on a couple of things is, to give a shout out to, Anna and and David with them. We have been working on this problem of, identifying how elements can be manipulated to enhance certain kind of product or content visibility. For example, positive articles about presidential, certain presidential candidates versus the others.

00:16:06:16 - 00:16:28:02

Can we game these models for manipulating content and what they show? When you ask questions and turns out you can. And they have a very nice work which demonstrates this. And the last one that we have been focusing on, as I mentioned briefly in my introduction, is understanding the gaps between regulations and research today. And finding out ways to fix that gap.

00:16:28:02 - 00:16:54:23

And Alex and Oceana written a very nice paper about that recently too. Alright, so I think, I'm a little I'm a little torn on we're going to take this and maybe or point this out like less. So I think first off, there are tons of technical challenges. I think if you if you put me to like

name one that I think is is particularly difficult, I would say uncertainty estimation and in particular epistemic uncertainty and inform- estimation.

00:16:54:23 - 00:17:15:00

So what is unknown, given the things that are missing from your data is this critically difficult problem. But I think it is. It is just a key ingredient of a whole bunch of successful systems. So that's I see a little point too, on the technical side, on the policy side, I think there is a I think there-

00:17:15:00 - 00:17:40:11

So I want to highlight, like I think we're now starting the conversation and this is a really, really good thing. And and so from that standpoint, I, I'm sort of encouraged, I think some of the ways that the conversation has advanced does make me discouraged. In particular sort of the impact of polarization. And a lot of the discussions that I see is really disappointing.

00:17:40:13 - 00:18:01:20

And I think there is a broad question that we are starting to struggle with, which is who's- like we know that there are benefits or we think that there are benefits. We hypothesize there are benefits. We're also pretty convinced there are some downsides. We've already seen substantial evidence of of fairly large externalities from systems that are out there.

00:18:01:22 - 00:18:25:08

I would highlight sort of like rampant cheating, creating these huge amounts of extra work for teachers, that is either being done for free by them or is leading to increased costs in schools, is like a large societal externality, and we're just not really talking about too much, like freaked out and now we're just, well, you know, let them solve the problem and we'll work on, you know, move on to the next one.

00:18:25:10 - 00:18:56:12

And then I think, you know, nonconsensual and, pornographic images of people. I think it is like that's just a new feature of society that's out there. And we have a bunch of people being hurt by it, too. I think it's possible for externalities to get worse. But the problem that I see is we're not really having a discussion about who should pay for these, that that conversation just hasn't really happened in society yet, like people have talked about liability, but it's mostly from the same standpoint of how liability would affect developers.

00:18:56:12 - 00:19:23:04

And fine. I think we could say developers shouldn't be the ones to pay, but if you're going to say that, you need to say who should pay. And right now, the answer seems to be whatever random people seem to be on the wrong side of, what these new systems bring into society. And that, to me is a is a concerning path that I think would be very easy for us to sleepwalk down.

00:19:23:06 - 00:19:47:13

Yeah, I think that's that's really true. From, from AC's perspective, we're also thinking quite similarly to that. We do a lot of work around, misuse. So we're worried about people taking these models and doing horrible things around *ineligible audio* child abuse, that kind of, material. And one of the things technologically that we think is going to make that a lot worse is that as the, agent environment or the agent.

00:19:47:13 - 00:20:09:22

Yeah, the agent landscape changes and becomes much more advanced. It's gonna become much easier for people to build out much further human abilities. And these abilities built quite quickly. And it's very difficult for, just one central kind of evaluator, government or anybody researcher to figure out what will be what will people be capable of, what will people be doing a year from now?

00:20:09:24 - 00:20:35:05

I will talk to the next questions here, which is, all of the panelists here have worked both on, techniques that are effective for safety on closed models, but you just have API access and also, techniques that, require you to have access to the way to open white box access. And I'd love to hear about people's perspective on how important that black box level of access is, and what kind of techniques they're looking at.

00:20:35:07 - 00:21:02:01

Okay. Right. So I can give one succint example, where {indistinguishable} but also a lot of the loss in machine learning model.

00:21:02:03 - 00:21:23:18

{indistinguishable}

00:21:23:20 - 00:21:50:11

So that's something we're trying to understand. And we are trying to see if, like you just automated {indistinguishable}

00:21:50:13 - 00:22:20:17

If we can't verify that system, {indistinguishable} Yeah, absolutely. And you, Joseph. Yeah, to answer your question. So, white box or black box? I, as I mentioned, to like more pro white boxes. I think, the research we are doing now, where we are now could not have happened if we did not have access to the way these models worked before.

00:22:20:19 - 00:22:39:01

And that's both for the like, from, like, a theoretical point of view, but also particular point of view, because we have access to like platform, like, or frameworks like PyTorch, which are all open source and freely available. And I think that was a great on contribute into the research. So yes, I'm very much into white book, white boxes.

00:22:39:03 - 00:23:02:05

I understand there are challenges with white boxes because then it makes, accountability a bit harder. Who's responsible? If some of that is just released, and anyone can use it to take your that your, example Dylan, things on nonconsensual like images. I do think that users should be blamed if these images are created, or released.

00:23:02:07 - 00:23:25:09

So I would probably. Yeah, I would I think actually in the UK it is the case that if you create those images is the user is taken is accountable and not the one creating the model. Yeah. So that's neat. Yeah. So the AC idea is, is perspective so much what we're doing in the getting pre-deployment access to model before it's publicly released and trying to figure out what's capable of.

00:23:25:11 - 00:23:46:24

And that kind of, that kind of testing is almost always black box because, the labs aren't usually able to give you full model access. Not at the time that they're releasing models. Actually, we've done a lot more focus on black box evaluation, but that is really good for assessing what it. But. Well, it is it is sufficient, perhaps, but efficient for assessing what a particular model can do.

00:23:47:01 - 00:24:06:05

But in order to get ahead and understand what's coming down the pipeline, it really requires a much deeper understanding of what's going on. And so some of our research, focuses and finally, quite useful to get open white box access and to do white box research and try to find that balance. It's quite hard and had a couple of points of discussion.

00:24:06:05 - 00:24:27:21

One is, as you rightly pointed out, I approached this question known as though is white box better than black box for your research or not? Not, not so much along those lines, but more along the lines of as we see more and more of these models becoming proprietary. What is our best bet in terms of potential auditing?

00:24:27:21 - 00:24:51:06

These models are seeing these black boxes from an external site and still figuring out if there are issues with it. Right. So that's a I think it's a very sort of simplistic high level. It is reasonable to assume that the more you know about some black box or some model, right, some some entity, the more we would be able to understand how it's functioning, what are its failure modes, and so on.

00:24:51:06 - 00:25:30:02

So you're at a very high level white box is preferable. But given that the world we are living in is forcing us to think more about auditing models with, let's say, query access, for example, I think developing approaches that can be extremely useful. Another thing I've been thinking about when we compare these two types of approaches, people can be a lot of post-hoc approaches that we, you know, operate on black boxes with, can, I think, help us get some good sets of hypotheses about model behaviors that if we really had white box access to, we may be able to verify in a more stronger manner.

00:25:30:04 - 00:26:01:10

Right? So I often kind of I get excited about this perspective of thinking about these problems as what if we developed this sort of, you know, sets of approaches which can operate at different granularity. So. Right, so where if I have black box access, if I have gradient access, if I know everything about the model, like, can this approach operate across these different levels and do better and better, and can we come up with new approaches that can seamlessly, go across different levels of accessibility to the model?

00:26:01:10 - 00:26:21:08

And there's also something that I think about quite often that, so, absolutely, I think I would I'll say that, you know, I was a part of a large collaborative paper titled Black Box Access is Insufficient for Rigorous AI Audits. So I'm a little bit, already out there at the position on this one and committed, I'll say it's a good academic.

00:26:21:08 - 00:26:44:05

It's a great paper. Credit to the coauthors and students that largely lead it and you should all go read it. But one of the things we document in there is, several different cases where increased privileged access, lets you get a better picture of what's going on inside of models. I think this is not surprising. It's in the category of things you expect to be true.

00:26:44:07 - 00:27:02:03

But now we have it all collected in a place. So when someone asks, how do you know that's true, you can point them there. I think one of the main things that we identified in that paper and that I think is, is a really useful takeaway is one I think methods that work across the different levels of access clearly needs to be part of the story.

00:27:02:07 - 00:27:32:24

We're not going to always get the level of access we want. You got to work with what you have. But to you, there are some mechanisms that allow us to meet the IP, primarily IP and competitiveness concerns of companies that still allow us to have very high degrees of access externally. So, for example, you can imagine setting up a sort of a secure site or a kind of a room that you lock down in some way and you can guarantee that sort of any information about what audits will run is wiped from that afterwards.

00:27:32:24 - 00:27:56:08

And then you model on things that are included in there are wiped afterwards. And this way you set up some privacy guarantees, we're actually protecting the model developers from more information being shared broadly. And we're actually also protecting the auditors from needing to share the deep details of their techniques, which is something that, we also haven't we haven't talked about quite as much is also a potential concern.

00:27:56:10 - 00:28:17:21

Right. Because you do want to maintain some level of, distance. So it's harder for companies to gain these kinds of, mechanisms. And that's a really interesting point there. With its a pilot with, topic around secure enclaves and that kind of, privacy preserving evaluation. I'm curious if you have any thoughts on that Vishal of, of info security.

00:28:17:23 - 00:28:44:17

So, on a flight to Boston, I was reading a paper which is {indistinguishable}

00:28:44:19 - 00:29:21:19

{indistinguishable}

00:29:21:21 - 00:29:43:02

{indistinguishable}

00:29:43:02 - 00:30:14:03

{indistinguishable}

00:30:14:05 - 00:30:45:19

{indistinguishable}, and that makes sense if it makes sense, pivoting a little bit. I'd love to hear from you, Vishal, as well as are we in this about what kind of advancements do you see in the UK that have been, really exciting? The best thing I would say, of course, {indistinguishable}

00:30:45:21 - 00:31:09:06

They have started asking the right set of questions. {indistinguishable} From the perspective of technological development {indistinguishable}

00:31:09:08 - 00:31:36:10

{indistinguishable}

00:31:36:12 - 00:31:57:10

{indistinguishable} Yeah, that's a lot of really interesting work. And it's often a bit sensitive because you don't necessarily want to release these to the wider world. Exactly how to make they can do all sorts of horrible things. So we are actually doing a lot of like there was a request for {indistinguishable}

00:31:57:14 - 00:32:13:00

They went out I think last week or maybe two weeks ago t{indistinguishable}. So yes. Yes it is to I'm going to talk about something totally different. I was going to say that you mind sending that my way after this? Definitely. We've got one here in my back pocket for you {indistinguishable}

00:32:13:02 - 00:32:32:02

And I'll link up about more on the infrastructure side. One of the things that's been really exciting is that, there's a new kind of supercomputer cluster, called Isambard that we've been talking about for a while in the UK, and it's now come online and it's really been, instrumental for supercharging the amount of the kind of research we can do.

00:32:32:04 - 00:32:50:09

Obviously, as we all know, anything that computationally intensive is very computationally intensive, that these models scale up and be able to research on the frontier really requires, dedicated facilities and infrastructure for that. So some very good to see the UK building up their capacity.

00:32:50:11 - 00:33:12:11

On the up side for, American side of the panel, the American base side of the panel. I{indistinguishable} most unique about collaborating with UK collaborators versus how do you see American collaborators looking at the same issues. If there's a difference? I'm not sure I actually see a difference there.

00:33:12:11 - 00:33:37:03

I think there's maybe differences in types of resources and and the things that people can get access to. But it it feels to me like, I don't know, I'm trying to think of my UK collaborators. Like in many cases, I feel like we're sort of we're sort of headed in similar directions. And then, I don't know, as far as access to resources goes.

00:33:37:05 - 00:33:59:23

Yeah. I'm sorry that I don't have a better answer for you on this one. It's good we are on the same page. Hima, so I am it is the difference, I mean, and talk about two types of collaborators on the UK, so. Right. One is that academic collaborators, they, they don't notice so much of a difference. I think, you know, the academic collaborators are thinking about a lot of these topics in a very similar way.

00:34:00:00 - 00:34:25:15

They're working on problems pertaining to adversarial safety and, you know, interpretability and so on. I have however noticed a little bit of difference when I talk to policymakers on either side, though, in the UK and the US, I think in the UK, I see a little bit more focus on the...maybe I'll use this phrase like the extended risks associated with these kinds of models.

00:34:25:16 - 00:34:46:21

Whereas when I have conversations with the policymakers in the US, then I think the focus is a bit more on the immediate risks, like the hallucinations, the fairness, the discrimination issues and so on. So that is it a difference that I note, but it's a difference that I observed

between the two sides. Yeah. I am not necessarily saying one is better than the other or if one is important more on than the other.

00:34:46:21 - 00:35:11:08

But I think I was just kind of mindful of that difference when I have these conversations. Yeah. And I sort of sense and the next question we have here is, just toward me. So I ask myself this question, but, how the UK is positioning itself to act regarding AI governance? Especially with relation to how other countries and geographic areas like the EU, or US are pushing it.

00:35:11:10 - 00:35:37:14

And I'll caveat this whole thing by saying I don't set the UK's policy on a lot of different things. So this is just my opinion. But I do think there are some differences in kind of the focus. And so one of the things that the EU has always been on the forefront of is, that the rules around privacy, GDPR, that kind of work and try to take a look at the EU AI act, they were really strong on that right out of the gate on cyber attacks and for individual consumers, individual privacy rights, which I think is hugely important.

00:35:37:16 - 00:36:00:00

And then I think if you look at like you're saying, with the US as was probably much more of a focus on immediate misuse harms, and actually, in the UK governments and think the way that the ACs are going to most effectively interact with someone, like with specialization. So rather than the UK saying, well, actually you only care about the A or B, I think it's actually quite useful to have different ACs focusing on different parts of the problem.

00:36:00:02 - 00:36:10:13

And so I think you make some of that kind of, specialization, even if it's not necessary, saying one is more important than the other.

00:36:10:15 - 00:36:50:07

I'll do the next question here. Yeah. Which is, what are the most exciting research funding job, opportunities you see available within your areas. Well, in the UK or externally. Okay. {indistinguishable}

00:36:50:12 - 00:37:12:08

{indistinguishable}

00:37:12:08 - 00:37:35:10

Are we really. You can you know, so normally research women limiting those people. And I want to say a little bit also when I say partner and manuals which we can solve. So instead of you who can back to can soon solution for a lot of stuff was well actually it was a little problem. But that's at all points of things.

00:37:35:10 - 00:38:03:10

{indistinguishable} Definitely. I Joseph I'd love to hear your thoughts on opportunities within, whether it's legal space or {indistinguishable} or more broadly. Yeah. Yeah. I mean, I think UK is quite interesting for that. {indistinguishable}, I think what's interesting about the UK is that's because it's part of Europe.

00:38:03:10 - 00:38:24:05

You have a lot of people coming from all over Europe, and also it's part of the Commonwealth. So it's a very diverse set of people who come with very diverse mind and, thinking and I {indistinguishable}AI research in general. So we have a very good ecosystem. I mean, I work in London, so I know quite well for that.

00:38:24:06 - 00:38:46:09

That's we have a lot of startups, companies, working on this kind of topics. Also you should mention we work hard with our relationships with academia. There is a lot of links between academia and industry in the UK. Well, a lot of students are encouraged to create their startups and they do. I know a few, I know a few who did this already {indistinguishable}.

00:38:46:11 - 00:39:05:06

And also as a like a worker in the company, it's also possible to have links with academia if you are doing some projects, so to do a PhD while being at the company. So there's definitely a lot of links here, which makes it a very, very interesting place. And that's what it is to work. I can just say as well

00:39:05:10 - 00:39:26:04

working for Robin we are growing up we doubled the size in last year, and we are looking for {indistinguishable} So if you are interested. Yeah. Of course I will put my plug for the UK, see as well. If you're interested. Please come talk to me. I think one thing that came through from both of the responses is that there's a lot more flexibility and opportunity than maybe seems obvious.

00:39:26:04 - 00:39:48:07

So I'm just talking about people who are doing a PhD while also working at a company or working doing safety research. I think that's something that we're seeing a lot because the space is expanding so incredibly rapidly, and there's so much important work to be done. The people are kind of able to start in in places you wouldn't think, in UK AC, I think version of this is that we are obviously a government institution as part of the UK government.

00:39:48:09 - 00:40:04:10

But we have gotten along with people who are, you know, American, German from wherever. Actuallu its not as-, I think maybe 30 or 40 years ago, you'd imagine it's only British citizens doing British research. And now we're in a place where anybody is doing good work. There's places for people to do that. So that's something I think is it's quite exciting.

00:40:04:12 - 00:40:25:07

And I'll tell you from more research side, what you guys are seeing is the biggest opportunities are places where we can make a difference. So I mean, I think I have to put in the obligatory I'm looking for PhD students and application season is soon check my website for for details. I think, you know, I think well, there there's a question.

00:40:25:07 - 00:40:46:16

I'll say I'm, I'm not totally sure about this, but but I think we might be like one of the questions I'm looking for is like when what are the startup environment around responsible AI and AI safety really kick off? I think we're waiting to see when private investment starts to, to really go to solving and working on these problems.

00:40:46:18 - 00:41:06:07

I, you know, I thought maybe we were there in like 2020, 2021. And, it seems like we probably weren't yet, but but I think one of the real questions as to what extent do companies see these, these risks and concerns as things that they have to spend money to to fix. And once that happens, we'll start to see the startup ecosystem really grow.

00:41:06:09 - 00:41:23:23

And so I think if you if you want to take a swing at a really high impact move, kickstarting that ecosystem with your own company, that you figure out how to get a good business model in place and get off and running, I think that could be massively impactful. I wish I could tell you how to do it.

00:41:24:00 - 00:41:52:18

I don't know too much advice there. So given that Dylan has talked about the, startup side, doing, I think the biggest opportunities also seems like a stronger collaboration between the government and academia. Because, you know, of course, buying this fact, but are like, keeping sites a lot better pieces that, well, lots of economics around so associated with, like different companies and organizations.

00:41:52:18 - 00:42:25:08

But if I keep that, aside for a bit, you could incorporate {indistinguishable} was probably among the impartial entities that is approaching these topics with bias. We had these external people that are trying to audit a lot of the models that are coming from different companies. So given that perspective, it makes a lot of sense for academia and government to collaborate potentially share resources, whether it is computers or, other kinds of resources, and come up with things together.

00:42:25:08 - 00:42:49:04

And that could lead to a lot more push, both in terms of, making sure that policy is catching up with technological advances rapidly, but also ensuring that there are no gaps between policy and technology. Like, we are not asking for something in policy documents that looks like what technology today would never do. Right. Bridging those gaps.

00:42:49:04 - 00:43:08:07

But also like just quickly catching up with technological advances. I think both companies could see improvement if that partnership happens in a seamless way. Which it's I don't see it happening that seamlessly. And I wish it was on the near future. Yeah, yeah, that makes a lot of sense. That's the thing that we're trying to think about, and it's quite hard to do.

00:43:08:10 - 00:43:25:09

Well, with clouds programs and things like that. But I think what you're talking about is maybe a deeper or a stronger connection and landing that I think would probably be quite important for this to go well. I think we should not start that collaboration just at the moment of, you know, we have a grant and would do some research and write a paper.

00:43:25:11 - 00:43:56:01

I think it should go beyond that, because that already kind of happens in different ways, like NSF funds, a lot of research and so on. But I think it's a bit more than that, which is, you know, there should be, like whether you want to do a grant or some kind of farm formal partnership where the implications of your research will directly and from policy, and there

will be more intimate dialog between policymakers and academics, and a more regular discussion and ongoing discourse between them.

00:43:56:03 - 00:44:20:02

Absolutely. We've got a few minutes left. So what I'll just do for the last question, just kind of open up to every one of these panelists to lead the room with anything they'd like to like to talk about. I'll start on the side tomorrow. We'll sweep through. Okay. {indistinguishable}

00:44:20:04 - 00:44:46:12

{indistinguishable}

00:44:46:14 - 00:45:10:12

{indistinguishable}

00:45:10:12 - 00:45:43:02

{indistinguishable}
Joseph. Yeah. I mean, this is very know speech, but I very much encourage you to, to come to the UK if you could I actually made the move a few years ago from Boston to London and been very happy by that.

00:45:43:08 - 00:46:04:17

I think there's a lot of opportunities there. And, yeah, it's a very nice place. And it's still. What it's nice is that it's. Yeah, it's like close to American way. I mean, English speaking country. It's not that far. And it's, Yeah, a lot of liberties. I mean, in general, we. Yeah, we are one of the leading, legal tech companies.

00:46:04:17 - 00:46:28:00

So we are and we are growing a lot. So, we are interested into {indistinguishable} That would be interesting working with us. So reach out to me. Okay, so I'm going to use this platform from England for quite a bit. I mean, we are talking about, you know, different countries and attracting workforce and talent.

00:46:28:00 - 00:46:55:15

And I do indeed believe that for any country to become a leader in AI, having amazing workforce, passionate people who want to work on these topics is like the most critical thing you need, right? And one of the easiest ways that you can get that is by making your

immigration processes efficient. So, somebody who is been {indistinguishable} challenges, but I have known some amazing people struggle with this waste

00:46:55:15 - 00:47:18:18

Well, energy, time and resources. I would love to say to other countries or anybody else who is listening that, the best way to become a powerhouse or leader in AI for any country is make your immigration easy and efficient. Yeah. So I've spent way too much time fighting with things like that. Yeah, I, I won't give you guys the gory details.

00:47:18:20 - 00:47:40:11

We I guess I have the last word in some sense, which is- I am going to go after you.. Sorry. Oh no no, then all the pressure is off. No, I think against the the thing that I want to leave people on as is something that I've been musing about recently, and I once has gotten into my head up and having trouble getting it out, and it's.

00:47:40:13 - 00:48:10:18

I think as, as academics, we are not being ambitious enough from, from the people outside of companies and in particular, I think the fact that we're we're largely not really touching pre-training, except to prove that we can do it, seems like particularly problematic and potentially a real issue for, for AI safety, if that's the place where it's easiest to to make changes to, to models and model behavior in reliable and systematic ways.

00:48:10:20 - 00:48:35:04

So, so one of the things that's been on my mind is, well, why why haven't we tried the really ambitious things? And I think it's because our model of how we share resources is we have a big resource that everyone then has, like access to in a fair way we do job scheduling, compare this to what scientists do, and we're figuring out how to send off a space probe like the one they just launched towards one of the moons of Jupiter.

00:48:35:04 - 00:49:02:02

Right? There's negotiation. We're figuring out what do we put into this big, expensive thing so that we can later on answer a bunch of different scientific questions. And I would love to see us sort of doing something like that, but with a regular schedule and cadence of pre-training, runs on a whole bunch of different scales. And I think we haven't done it yet because it's too we've got too easy of a way to waiting for us to share resources.

00:49:02:04 - 00:49:19:20

And this this is going to be really difficult to coordinate. But I just I wonder if that's sort of the level of ambition we need to really be thinking through. And it's it's more of a social problem and a technical one. Yeah. I think one of the things we find is that a lot of problems seem like technical problems, actually social problems.

00:49:19:20 - 00:49:38:21

when you get down to them. Yeah. I'd like to add to that piece. I think the two biggest things I've taken away from being at, at AC, the, importance of ambition and immediacy. I think a lot of when I'm talking to candidates or people are thinking, well, I can do this other thing, but this might set me up to do something five years down the road, ten years down the road.

00:49:38:23 - 00:50:09:09

I mean, there's a lot of kind of playing it safe, which I think makes sense in a lot of different situations. But I think we we are at a point where people can have a really great impact by doing something now and by doing something that they're not sure if it'll work or not. So I think something I've talked about, they're doing and trying, like swinging for the fence in a way that doesn't feel safe and feels a little bit scary, and taking bets on doing something right now rather than trying to layout impact, I think, is a important ironic to hear that from the AI institution ha ha ha.

00:50:09:11 - 00:50:35:05

Socially, it was like, I mean, Glenn I think we can begin to open up for questions and answers, I think have a mic that will be floating around the room and there'll be some zoom questions as well. So. Like, yeah, right there.

00:50:35:07 - 00:51:10:12

Hello. I first I just want to say thank you. My question is about where we're headed. So there's often the idea, for when we're talking about AGI which I'll define as, general artificial intelligence system that is, as capable as most people are, most things. So I'm curious about your thoughts on whether or not we should expect such models the next few years and how these models or models are even just smarter than all people, which means how you think about safety.

00:51:10:14 - 00:51:30:01

Then maybe I'll jump in first. I'll say, I don't know if expect is the word that I would use, but I think be prepared is is sort of right. I think, you know, that it's unclear is sort of like the AGI point to me feels like a somewhat arbitrary dividing line. There are there are new capabilities coming out and things may go fast.

00:51:30:01 - 00:51:52:21

There may be reinforcing components. We're just not sure. We've just got a lot of uncertainty about how quickly things will move. And so I think the prudent move is tend to be ready for for those possibilities. And I think I'll throw in and, you know, a lot of people talk about, sort of how much should you worry about that future problem versus how much should you worry about the things today?

00:51:52:23 - 00:52:09:19

I think if you're actually serious about managing and being prepared for that problem, you're doing a lot of things that are really, really useful today. And I, I guess that to me means that I that's why I don't spend as much time thinking about that part as much, because I feel like I want to be prepared for that.

00:52:09:21 - 00:52:48:02

If that doesn't happen, I feel like the the consolation prize will be that we'll get a really robust ecosystem for the technology that that's safe for the technology we will develop. And that's sort of my that's one of my core beliefs, actually, in doing this work. So, yeah, adding to Dylan's point a comment I think at this point during sort of an ambiguously defined end goal, I believe that we might want to think about this as we are likely to see more and more heightened capabilities under different, tasks and different like aspects that humans can, the different tasks that humans can do.

00:52:48:04 - 00:53:06:09

We can see more improvements on those in the coming years, how quickly they're likely to happen, where it's going to take six months or like one year. I think those things are quite hard to predict, but it doesn't hurt to be prepared. But I want to touch up on the second point you made, which is what is the implication of all of this for AI safety?

00:53:06:09 - 00:53:32:24

Right. Are you thinking about it in a couple of things. One is that, you know, and I'm using this ambiguously defined term for now just to capture the question, but, let's see, we get more closer to AGI is going to raise more AI safety problems or permanence. It kind of is omni intelligent in some way. Is it going to come with solutions for safety problems?

00:53:32:24 - 00:53:53:06

Right. To which we are going is something that I often contemplate about. But, given that we don't have answers to that at this point, maybe it's better to say, let's be prepared for

safety challenges and try to address it, from our side. Right. Like as opposed to assuming that the AGI will have solution states on safety challenges.

00:53:53:07 - 00:54:05:06

Right. But maybe there will be such an AGI or we can hope to see something like that at the future. You never know. Yeah. Any.

00:54:05:08 - 00:54:32:01

So on it has a lot of question which my postdocs ask me {indistinguishable}

00:54:32:01 - 00:54:59:20

{indistinguishable}

00:54:59:22 - 00:55:49:22

{indistinguishable}

00:55:49:24 - 00:56:42:13

Question here {indistinguishable}. Yeah. Hi. I will introduce myself, and said so as I'm {indistinguishable} fellow in technology, policy and management. So I'm a technical engineer. {indistinguishable} empowerment in stand and provide the digital literacy and allow some I'd like to hear you know, thoughts about this AI bias specially for black skin color for people with black skin color, like me, and {indistinguishable}

00:56:42:13 - 00:57:04:11

{indistinguishable} do you think? Like, what is it about all of these behind that?

00:57:04:11 - 00:57:28:21

Like, the developers also like, kind of biased into when developing for doing that. {indistinguishable} my When I tried to unlock my iPhone. I use like the face recognition, but sometimes it does fail. I like, yeah, I'd like to have more from you. Yeah. Thank you. Well, I think I can I can sort of jump in here to your first maybe, which is.

00:57:28:23 - 00:57:52:21

Yeah, I think there's a lot of different, sort of different sources where, where things like that might come out. I think one of the ones, the one that's, that's been discussed a lot is the sort of biases in the data pipeline where you have either either less representation of

different skin colors or genders, or the representations that they show up in our are heavily biased towards certain types of outcomes and possibilities.

00:57:52:23 - 00:58:14:12

And this is like a clinical driver of it. I mean, another one which I think plays a big role in the broader AI ecosystem, has to do with the fact that a lot of the companies building this don't have good tests for these problems or just developing them. And so the one of the things that I think about a lot is we don't start off knowing everything that matters about a system, right?

00:58:14:12 - 00:58:45:04

We have to decide what we want to measure and evaluate about it. And this is one of the biggest places where biases from the people building systems come in, which is to say that now they're choosing what to add in. And build and use measurements and where we allocate that effort. And my guess is that a lot of people that are building the systems are well, you know, well, we learn things about their demographics and so there, I think, just much less likely to stumble upon finding this problem and then recognize it as a problem and go fix it.

00:58:45:06 - 00:59:11:18

And I think that question of how do you find the things that you've missed? This sort of relates to my point about epistemic uncertainty early on. I think it applies to AI systems, but I think a lot of this looks like it sort of failing at the level of AI developers and sort of not doing a good job of, or it being very difficult to to find the things that you've missed and not just kind of falling out in the system.

00:59:11:20 - 00:59:37:16

So like I think we talked earlier about so this is the white box. I think it's at a different level of white boxes. And so where you don't have access to the model and it's it's it's also interesting to have access to its data, which is another level of white boxes because I assume like one of the problems with skin color is that these models have been trained on a lot of data from white people, and a lot less with black people, and that's something that you can easily check and detect if shown access to data.

00:59:37:19 - 01:00:11:11

Someone that has been trained on. So I think, {indistinguishable} it would be very good to have these kind of data available. So my perspective on this is it's clearly a big problem by the way. And it has been debated on various stages, for very long time to but something that would be very useful. But I, I wish there was more activity around something like this is see,

there is always going to be a difference between somebody who has never been at the receiving end of this problem, talking about this problem versus

01:00:11:11 - 01:00:34:15

So somebody who has been at the receiving end of it, talking about this problem. Right. So can we develop more frameworks or tools that can allow end users, even if they're not experts in machine learning and technology, to potentially identify the blind spots or pain points like this, as external users kind of tinkering with the model. Right?

01:00:34:15 - 01:01:04:17

I would love to see more and more of those tools and more and more of those frameworks, because the passion with which somebody who was affected by the problem would go at it and try to identify these things would be a whole other way of approaching this problem. I think we need to bring that energy more into solving this problem, and build frameworks that make it easier for end users to be able to tinker with these models and say, oh, here I'm able to find these instances when I see a clear bias problem, right?

01:01:04:19 - 01:01:25:02

On that front, I also want to talk a little bit about {indistinguishable} recent work, which is we approach these problems from the lens of interpretability, where we look at it as, you know, the more you're able to see what sorts of concepts are being captured in different layers of the model, you have some hope for intervening in them.

01:01:25:05 - 01:01:43:18

Right? S{indistinguishable} built this nice framework, which kind of, takes the explanations that some of the other methods give and say, oh, this layer is capturing, let's say a skin color, for example. Right. And let's say it should not. So her tool analysis, the ability to edit that out, or intervene and say this should not be captured.

01:01:43:18 - 01:02:04:09

Like now how do I invite that into the, layers of the model? Right. I think tooling like that. But also with an interface that an end user could tinker with, is supremely important for addressing these problems, like right from the ground level. I think that passion with with somebody who's affected by this problem needs to have a say in addressing these problems.

01:02:04:11 - 01:02:25:22

If you can, I can I oh, sorry. Go ahead. So I think that's that's absolutely right. And I think that there's a technical element to that and a social element to it. And the technical element is do we have the tools for somebody a user is not technical to actually have an impact on how these models are being, you know, developed or being edited or being modified and I think there's a social problem of who is deciding that, who is making sure that happens.

01:02:25:24 - 01:02:43:24

And I think that's something we haven't yet answered as a society, because right now a model gets released without releasing the model, and that's kind of it. Until there is the next model. And so right now it is the developers is these labs. And as you mentioned, the labs have certain demographics and they have certain, incentives and they have certain constraints that they're operating within that society I think isn't.

01:02:44:01 - 01:03:04:16

And I think that's where you should see you should go and you should see academics should see. Yeah. Yeah. I think just to summarize, I think it should this bias should become addressing bias should become more than a research problem. And I think we should see that there we should meet and users participate in addressing this problem.

01:03:04:18 - 01:03:23:12

It should not just be something that is being limited to academic labs and, you know, companies, research labs. It should become way more than that. I think that's when we really address this problem. I just wanted to say that I wanted to the point you made about the social challenges here is, I think, one that I wanted to echo.

01:03:23:12 - 01:03:47:07

I spent a fair amount of time thinking about platforms and recommendation systems. And in that case, it's very, very clear that recognizing the problem in the system and advocating in a way that gets someone to make a difference when they're building the system are two massively different problems. Like when I looked at that problem and said, like recommendation on algorithms are misaligned.

01:03:47:09 - 01:04:13:15

We need technical tools here. Like, turns out, yes, we would like some technical tools, but the big problem is just that getting that information back into that, like even convincing a platform to measure a specific thing that's related to bias is politically a huge lift, much less getting them to do anything about it.

01:04:13:17 - 01:04:44:17

Sorry. Hello everyone. {indistinguishable} at Northeastern University with David Bau. Thank you for the discussion. I was curious, you talked about the startup ecosystem briefly. And, I'm right. And so it's the intersection between governance and research. So, you know, I'm curious about an inside view of, like, writing, I think the regulation will create {indistinguishable} yeah, we're creating a business.

01:04:44:20 - 01:05:10:24

And how do you plan on collaborating with, like, a lot of third party startups and {indistinguishable}? You know, that's a really interesting question. I think the, the short answer is, we don't know yet. I think that's actually one of the really big benefits of the startup scene is that it's like what it goes where, it's most useful, or the path of least resistance to the path of most utility, I guess would be the, the Adam Smith interpretation with that.

01:05:11:01 - 01:05:29:00

But I think there are a lot of things that we, from the AC perspective we think we can be setting direction, but we shouldn't necessarily be doing all of the work. So. Right now, evaluations happen in the major labs and by 1 or 2 major, maybe 2 or 3 third party evaluators. AC Apollo, meta, a couple other big, big players.

01:05:29:02 - 01:05:52:14

But it doesn't necessarily have to be that case. If you look at the airspace industry, there are a lot more people who are doing this kind of assessments. And so we think that that's the, low hanging fruit for much stronger third party startup ecosystem for doing assessments. I think as you get into, I mean, start building out more technical, intervention type approaches, you can start seeing more and more people who are building out companies around providing those services.

01:05:52:16 - 01:06:00:06

Which is we I think, but hope that some of that question.

01:06:00:08 - 01:06:21:18

{indistinguishable} I work at a lab here called MIT future tech. I also work on this project called the AI Risk Repository. I just wanted to know if I have something that I hear that was sort of raising your eye, which is sort of suggesting many that there was this, gap are we were blind spot around how we managed biases, with AI models.

01:06:21:18 - 01:06:44:10

So I was wondering with those two. I think we're also on sort of a sense, a collective blind spots maybe in the community, things that they think, you know, will be useful to point out that maybe are being missed at the moment. {indistinguishable}

01:06:44:11 - 01:07:21:19

I{indistinguishable}

01:07:21:21 - 01:07:48:15

{indistinguishable}

01:07:48:15 - 01:08:18:23

What so what are political {indistinguishable}

01:08:19:00 - 01:08:47:19

{indistinguishable} are suggesting some sort of actions for your supply chain and that basically make school conscious and unconscious bias being a system.

01:08:47:21 - 01:09:13:05

So, so those kind of things and workflow sort of as a part of governments. And so we're trying to explore, you know, better also, but you know, and also I think it's a lot it's sort of an abandoned role that governments brought back into the stifle to the social model. Yeah. I'm just going to turn it over here and we can get a question from the audience, and then we'll come back.

01:09:13:05 - 01:09:48:12

And when. Thank you. The question was given, the panels focus on frontier AI safety. I'd like to explore allocation of safety resources across style and landscape. So while frontier models arrived from the central concern, trailing edge models in the hands of bad actors could present substantial risks in the short to medium term. How do you view the balance between frontier focused safety research and addressing the more immediate risk posed by trailing edge models?

01:09:48:14 - 01:10:25:03

What criteria really should guide our priorities in this regard? One thing I can say from the AC's perspective is that the frontier isn't necessarily just a single number. It's it's more of a,

it's a curve. And so our threat modeling has how well resourced are the actors and also how, how much capability to get from the AI models tooling that somewhat addresses what that question is asking, because we don't necessarily think that, a disgruntled teenager sitting in their room is going to be fine tuning, you know, GPT-6 or something like that.

01:10:25:05 - 01:10:46:08

But we do think there are things that people have access to. And so we're doing in that threat modeling is doing a number of different, levels of capabilities or levels of an assessment to try and understand what even are actors can do and what's the combined harm or risk from those two actors. I think I'll I'll chime in here to say that I think there are not enough people.

01:10:46:08 - 01:11:17:01

I think a lot of people look at this as like, should I do A or B and I don't think there are enough people looking at how can I do A and B, because I think there really are a lot of things that are valuable in that intersection. So. So for example, on the aspect of sort of edge models being being problematic, one of the ways in which that could be an issue is actually that you have this diffuse set of problems that happen in society and like where the big kernels is just detecting that that's going on, and then marshaling a societal response or figuring out how to respond.

01:11:17:03 - 01:11:42:01

I think this type of problem of detecting that something's going wrong is going to be critically important to have at the frontier, as those types of things happen, because we'll be looking for new problems that have occurred that that we don't necessarily see. That would be a sort of being clear, a point of like, oh, yeah, definitely something happened, but this sort of diffuse, sort of slowing the on a whole bunch of changes distributed across a ton of people.

01:11:42:03 - 01:11:56:17

I think actually, to get to the question of, blind spots, that that's my that's my blind spot. And I think we exercise that well by thinking about those systems that are going out there now.

01:11:56:19 - 01:12:29:07

Yes. Hello. And thank you for your discussion. I'm, {indistinguishable} MIT student. And I just, have a question. I felt like going to one of the biggest takeaway from this discussion is that AI safety is not a, technical problem. It's kind of, to understand those concept of

interpretability and explainability, kind of {indistinguishable} of complexities in both social and technical aspects.

01:12:29:09 - 01:13:02:01

So I mean, online, {indistinguishable} that individuals like small academics. Are trying to developing both, developing expertise in both sides in technical and in social aspects, like is I know that in my world that people can really work together, instead of just sharing, knowledge. But really work together to understand the problem and to, yeah, work out the problem.

01:13:02:03 - 01:13:28:00

Yeah, that's my question. I can start I just wanted to say by making this one correction, I think it's not that interpretability or for that matter, AI safety is not a technical problem. I would say it is not just a technical problem. Right. So I just want to make that distinction very clear. So it's both a technical problem but also something that interfaces with society.

01:13:28:02 - 01:13:50:16

And I agree with a broader point that I think the way we look at a lot of these problems as, we sit in our labs, and then here is a nice metric that I can come up with and I can prove something about it. And, and, it gets in the paper and, like, look, we can solve the problem of any topic, right?

01:13:50:16 - 01:14:20:15

Like fairness, interpretability, bias, anything. Right? I think {indistinguishable}not coming up with solutions that are actually useful in practice, but we are telling ourselves into thinking that we have solved this problems just because you have seen 100 papers out there, right? So I think that's exactly why we need collaboration from a broader set of actors who are not just people sitting in research labs, but especially people on the receiving end of the consequences of some of these systems.

01:14:20:15 - 01:14:41:17

Right. And I'd just like to add one last thing to your point. But the fact that I think it becomes very clear right at the definition, let's say an explanation, right? Like an interpretation of a model, just by definition, you know, it has two sides of the coin. One side of things, this interpretation should be faithful to a model.

01:14:41:19 - 01:15:02:24

Otherwise it is not a really an interpretation of the model on the other side, that should be understood by somebody or intended to be understood by right. So there is only two sides of the client or the human side. What is the {indistinguishable} side? I think both sides are important. Therefore we need that interaction.

01:15:03:01 - 01:15:26:10

I can also speak from the industry perspective. Is that Robin AI don't three quarters of the employees are lawyers who do not have any technical background. And it is a daily challenge to work with them. And somehow we have a way we have to have, like a way to understand each other because they have to understand the matrix that we are using and the way we are building these models.

01:15:26:10 - 01:15:43:12

And we have to understand their needs and, which also makes part of this, work very interesting because we have to have this collaboration working if we want our product to be successful. And that's, that's a very important challenge in our company.

01:15:43:14 - 01:16:12:19

We're going to take three more questions one, two, three. And then we'll wrap up there are refreshments and stuff. So three more questions. We can keep the conversation going. Just not in this format. Hi. I'm once a student here in {indistinguishable} and, one question that I could have been thinking about is, we talk a lot about AGI in the future, but, some of these models are, where they're taking commands over the last few decades.

01:16:12:21 - 01:16:44:24

Are there points of inflections, say, the coming of the internet, social media system, then, in each of these additional inflection points with different challenges, like, for example, the internet is, not like a corporate owned entity. Then social media systems handling mistakes made in like, in the media landscape. So I guess, what lessons can you, AI safety landscape take, from these, you know, just these, are the inflection points in the past, and.

01:16:44:24 - 01:17:05:06

Yeah. Well, what can we learn from that? I think the AC Institute pretty explicitly as a reaction to some of that. So I think if you just talk about social media, I think there's a pretty strong sense that governments were behind, behind the ball and that, like, we didn't really realize how important or how impactful this is going to be until it already happened.

01:17:05:08 - 01:17:25:14

And I think that was part of the explicit pitch for why we need to be making sure government has an understanding and were good technical understanding of AI before we get to AGI, whatever the, whatever the delineation it it's used to make their yeah, I think I would be the biggest answer that I think the details actually can look quite different.

01:17:25:14 - 01:17:46:12

I think the internet for social media versus AI, are such different systems that, the direct lessons maybe don't apply as much. I think that, situations the way that will come in. So one example is, social media. One thing that, we saw as the company completely we were banding together trying to determine this, like child sexual abuse material.

01:17:46:14 - 01:18:07:00

We had crush company repositories of things like the hashes of dangerous photos, that kind of thing we have not yet seen, I think, at the level of here are techniques that we should all be using or looking at. And I think that creating spaces and creating opportunities for that kind of collaboration between labs is something that's that's the lesson we could take on social media.

01:18:07:02 - 01:18:34:11

{indistinguishable} Right. I often say, sorry, it's going to be a little bit philosophical, but that's how I put that question to myself you know, {indistinguishable}

01:18:34:13 - 01:19:10:19

{indistinguishable}

01:19:10:20 - 01:19:45:20

{indistinguishable}

01:19:45:22 - 01:20:15:07

{indistinguishable}

01:20:15:09 - 01:20:38:00

{indistinguishable} start taking shape. And a lot of those questions, some of them demands, will kind of be logical to confront, at least to some extent. We have some answers to it. So just adding to both of those, great points.

01:20:38:02 - 01:21:14:07

So I have found that a lot of my questions whether it is broadly about the pace of the technological advancements of AI, AI, safety policy and so on, can be answered by thinking about all of this in terms of incentives. Right? So I think what we might want to learn whether it's from the social media era, or the era before or even, you know, like prehistoric agencies when somebody is building something, what incentive would they have to self check in an objective way.

01:21:14:07 - 01:21:55:07

So I think kind of let's say if you, start a company and that you're building a model and so on, and if you advertise it as this has all possible safety checks, it's fair, like you retested it. I think if I, if I were like an external user, it is in my best interest and that of the best interests of several other people like me was not closely associated with you to double check your claims in as many ways as possible, because I think they're just simply not incentives aligned for you to make really honest gains and be critical about what you're putting out, when on some level, when you're also running a

01:21:55:07 - 01:22:20:12

business right. I feel like that part of something that should inform a lot of how we think about this landscape yet again. {indistinguishable} but we're going to wrap up pretty quickly. So thank you. Yeah, I'd like to, go back to something that Dylan raised earlier.

01:22:20:15 - 01:22:51:02

Regarding liability. It's a great segway from talking about incentives. So, yeah, I would love to hear your discussion on, who should we assign, that, you know, should we assign my ability to. For what externalities. And if you want, like, is something to start with. SB 1047 in California, which was recently vetoed, prohibited developers from keeping or deploying frontier AI systems.

01:22:51:04 - 01:23:32:02

If doing so, posed an unreasonable risk of, essentially causing mass casualties or harms in excess of 500 million. I'll give you one other point to like, play around with. I know scholars who are, who've proposed essentially treating AI systems like employees of developers or providers, where if the AI, commits a crime, then the, the, the, developer or provider, is is liable, and where the user, couldn't foresee that the model was going to commit the crime.

01:23:32:04 - 01:23:53:20

Yeah, anyways. You guys can take that wherever you like, but. Yeah. So so I think the point about incentives that that Hima made is drives a lot of how I think about this and to, to

make the point, you know, a useful thought experiment for me is let's say there is some safety intervention that adds a 10% increase to the cost of pre-training.

01:23:53:22 - 01:24:21:06

Do we expect that to happen? And when is it a good idea for that to happen? I think are two separate questions that we can think about societally and and my claim is that, liability of some kind for developers feels at least my belief is that that's probably necessary for these types of things. And I'm not sure why there that type of intervention is necessary, but it seems like we should be acting as if it might be expensive to to solve some of these problems.

01:24:21:06 - 01:24:51:10

And we should expect someone to foot the bill, and we have to figure out how to pay for levels of society. In terms of the breakdown of of who's responsible, I think it doesn't necessarily have to be an either or. I think you can both say that, the person who intentionally misuses a model to cause harm is liable for the consequences their actions that I think nothing that we are proposing here should reduce that.

01:24:51:12 - 01:25:18:19

On the other hand, if that's the only thing that if that's the only person who's liable, you might miss out on a lot of opportunities to, sort of intervene and reduce the number of people who do go misuse the models. So why in terms with allocating liability? I think the right way to do it kind of depends on how expensive to rethink those pre-training interventions are.

01:25:18:21 - 01:25:46:18

And do you think they're necessary. And as a sort of to make the problem even harder, to some extent, if we this is all assuming we know which interventions are important. One of the real challenges is like we kind need the companies to do the research on how to figure out avoiding these problems. But then, and I'll give a shout out to a paper I wrote called the Penalty and Default Approach to AI regulation.

01:25:46:20 - 01:26:09:22

But, one of the things we analyzed in there is a mathematical model on these incentives. And basically, as a developer, if you've got some potential harm that could go wrong, you've got two strategies. You think they're trying to fix it and make it so it doesn't happen. Or you could let it happen and take the gamble, that people won't find it and then only fix the things people find.

01:26:09:24 - 01:26:25:05

And that ends up being sort of mathematically the right thing to do under a whole bunch of reasonable, reasonable assumptions. And that thinking really drives a lot of a lot of my thoughts here.

01:26:25:07 - 01:26:52:20

{indistinguishable}. I think we went also on to think, on your point, and I was just thinking about that as you're talking about this, demonstrating about, even, like, complete recklessness. Right? How do we do that once you

01:26:52:22 - 01:27:14:11

{indistinguishable}. The limit is going to be an extremely challenging problem. Let's think of health care, for example. So if there's a surgeon or a healthcare provider who is done something with an intent to harm, how do you prove that intent? But now you have added another complexity into the process, right? There is a machine kind of doing some recommendations or part of the process.

01:27:14:13 - 01:27:38:10

Now, how do you demonstrate like intent to harm or if not, but the other end of the spectrum is how do you, demonstrate real negligence in the process? Right. Like you're you just {indistinguishable} group even putting the most basic safety checks and that you're {indistinguishable}g one possibly could.

01:27:38:10 - 01:28:03:04

You know in today's knowledge but still went wrong. Right. So how do you demarcate the situations is an extremely challenging problem in this context. Say also one thing. My {indistinguishable} how do you do an attempt? I don't know the answer, but I think if you don't document in reasonable detail what your model supposed to do, we should assume that everything it does is what you meant for it to do.

01:28:03:06 - 01:28:26:14

So that's my my high level answer to that puzzle thing that I've said, jump in and we've had some amazing questions online. I mean, we've got some real questions and we we didn't get a chance to get to you know who the panel are and who we are stay connected with us. Let's keep the conversation going. It's fascinating. And it's vital and important stuff that everyone is working on.

01:28:26:16 - 01:28:44:14

And I want to say a quick, personal thank you to {indistinguishable} who really put this event together {indistinguishable} in Cambridge, so be here and be there for the event. And to say a huge thank you to the panel, if we can all give them a round of applause

01:28:44:14 - 01:28:47:19

Just so.

01:28:47:21 - 01:28:57:16

And there are refreshments for those who are here and {indistinguishable}. So let's keep the conversation going and thank you so much. Enjoy. We're evening. Thank you, thank you.