



HOUSES OF PARLIAMENT

Joint Committee on Human Rights

Uncorrected oral evidence: Human Rights and the regulation of artificial intelligence, HC 1262

Wednesday 25 February 2026

2.25 pm

Watch the meeting

Members present: Lord Alton of Liverpool (The Chair); Baroness Chakrabarti; Baroness Hamwee; Lord Murray of Blidworth; Lord Rook; Lord Sewell of Sanderstead; Peter Swallow; Sir Desmond Swayne.

Also present: Dame Chi Onwurah.

Questions 89 - 98

Witnesses

I: Rob Sherman, Vice-President and Deputy Chief Privacy Officer, Policy, Meta; Ginny Badanes, General Manager, Tech for Society, Microsoft.

USE OF THE TRANSCRIPT

1. This is an uncorrected transcript of evidence taken in public and webcast on www.parliamentlive.tv.
2. Any public use of, or reference to, the contents should make clear that neither Members nor witnesses have had the opportunity to correct the record. If in doubt as to the propriety of using the transcript, please contact the Clerk of the Committee.
3. Members and witnesses are asked to send corrections to the Clerk of the Committee within 14 days of receipt.

Examination of witnesses

Rob Sherman and Ginny Badanes.

Q89 The Chair: Good afternoon and welcome to the 45th meeting of the Joint Committee on Human Rights in this session. I would like to particularly welcome Dame Chi Onwurah, who is the Member of Parliament for Newcastle upon Tyne Central and West. Dame Chi is the Chair of the House of Commons Science, Innovation and Technology Committee and she is guesting with us today, as she has done in an earlier part of our inquiry into AI and human rights. This is the sixth and final public session of this inquiry since we closed our call for evidence in September last year.

Members of the committee have noted that earlier this month Zoe Hitzig, a researcher with OpenAI, announced her resignation with an essay in the *New York Times* calling “deep reservations” to her aid, worried about OpenAI’s future plans, warning that ChatGPT has the potential to manipulate people and that their data is not being properly protected. We have previously noted that the Nobel Prize-winning scientist Geoffrey Hinton, known as the godfather of AI, left his role at Google, who were witnesses before our committee quite recently, to warn that AI poses an existential risk to humanity. These are important, disturbing and troubling questions for the committee—our mandate is the upholding of human rights—to explore properly and competently. We are, therefore, extremely grateful to our two witnesses who have joined us today.

We are going to hear from two of the biggest names in the artificial intelligence and digital technology sector, Microsoft and Meta. They will share their corporate approaches to respecting and protecting human rights. The committee will then welcome the Government Minister for AI, Kanishka Narayan MP, from the Department for Science, Innovation and Technology. He has just returned from the AI Impact Summit in India, which he attended in New Delhi. The Minister will be given the opportunity to update members on the progress that the UK is making towards things like the ratification of the Council of Europe Framework Convention on AI and how his department is working with regulators to ensure that safety measures are effective within AI systems before they are released to the public.

First, following our recent hearing with Google, let us turn to Microsoft and Meta. We are going to hear from Ginny Badanes, general manager of the Tech for Society programme at Microsoft, where she leads a global team focused on protecting communities, strengthening security and building trust in technology. Her work spans safeguarding elections, supporting law enforcement and advancing responsible innovations to ensure technology serves the public good. She is joining us online.

We will also hear from Rob Sherman, the vice-president and deputy chief privacy officer, Policy, at Meta. Mr Sherman leads the company’s efforts to develop and implement policy solutions across its products and services, with a particular focus on data protection and emerging

technologies like AI and mixed reality in collaboration with Governments and global experts. Since joining the company in 2012, Mr Sherman's work has influenced nearly every major Meta project like Meta Quest and Ray-Ban Meta Smart Glasses and Meta's business product suite. He is a frequent speaker on innovation policy and privacy and has represented the company before Governments on six continents.

Before turning to my colleagues—and we will begin after I have asked the initial question to you by hearing from Sir Desmond Swayne MP, who wants to ask you about transparency—I will open with the question that sets the scene for the others, which is about the balance that has to be struck between innovation and human rights. In your view, does the UK's approach to AI regulation, as outlined in the *AI Opportunities Action Plan*, strike the right balance between supporting innovation and protecting human rights? If the answer to that is yes, are current UK regulations adequate to address the novel, evolving and emerging harms from AI, some of which I referred to in my preamble? If no, what needs to change in the UK to address the most important and urgent harms and risks from AI? Ginny, would you like to speak first?

Ginny Badanes: Sure, I am happy to. First, thank you again for including me here today, especially for accommodating me when I was not able to join in person.

That is a long question and I will try to keep my response short, as I know we will get into the details as we go on later this afternoon. First, our approach is that the UK has had a sensible start to this question and you are building on a really strong foundation. Essentially, as you all know of course, the UK already has a strong foundation of human rights throughout your law, whether it is the Human Rights Act or the data protection laws, but you have also already started regulating on AI. One of the highest harms, which is AI-generated CSAM, is an area where you have already started a strong foundation. We think that that is the right approach so far.

In addition to that, we appreciate that you have made a distinction in the regulation between the model level and the use case level. For us, that is an important distinction and one that we agree with as well. You have also started a strategy that builds on something that we think is foundational to the tension between innovation and the regulation. You are starting from a place of building trust with the public, which is critical because public trust will be the foundation for widespread adoption of AI. People will not embrace and use a technology that they do not trust. In large part, some of that trust lies on us as the companies that are building and deploying this technology, but foundationally we believe that strong regulatory frameworks will increase that trust as well. When people trust the technology, they will use it, and that is critical for us as well.

I can say that in Microsoft's 50 years of operation—I have only been around for a portion of that—we have seen first hand how trust has been a critical component in every wave of innovation, and AI is not an

exception to that at all. Then, of course, trust grows when regulation is proportionate and risk-based. We will get into more of that as we go on today, but that is our foundational comment as you are looking and considering what you might build within the UK. It is critical that you have a risk-based approach. The highest risk harms, the most likely products to create harm, are where you start. Then it is proportionate, meaning that we are looking at how we support those who are at highest risk while also making space for innovation and growth.

I will stop there as the initial comments but I look forward to the additional conversation on this topic.

The Chair: That is a very good opener. Thank you. Let us turn to you now, Mr Sherman.

Rob Sherman: Chair and members of the committee, thank you and good afternoon. I really appreciate the opportunity to be here to talk with you about AI and human rights. There is probably not a more important topic to think about as we are building this technology for all of Britain but also all of the world. I appreciate the committee's attention to this important topic.

When I think about the role of AI, and particularly the way the UK has approached it through the opportunities plan—which as the committee will know was recently updated—it is a thoughtful and sensible approach and in some ways a global model. If you look at the way that the opportunities plan is approaching AI, it is recognising that we are at an inflection point where the technology has the opportunity to be pivotal and enable access and voice in ways that just were never possible before, and economic growth as well. here in the UK. It is an important opportunity to leverage the technology to enable people across the country, at the same time recognising that there are real risks associated with the technology and we need to be thoughtful about it.

There are two things that I think are particularly helpful about the approach outlined in the opportunities plan. The first is, in contrast to some other jurisdictions in the world, that the approach is to leverage the expertise that exists in the Government and government agencies to regulate. AI will become a part of almost every sector, every technology that we use, and so making sure that we are empowering the government agencies that have expertise in those areas to enforce is one important area.

The Chair: Mr Sherman, that is really helpful. On the reference that you just made to other jurisdictions, is there a jurisdiction anywhere that you would point us to and say it is doing exactly the right thing in protecting human rights?

Rob Sherman: I certainly would not say that human rights is a soft problem anywhere in the world. It is an important area of focus for us all. While I am quite impressed with the approach that the UK Government

have taken so far, we need to continue to collaborate on this and I do not think there is a single model that I would point to.

The other thing that I will also mention, which I think is an important attribute of the UK's approach that is important for protecting the public and for human rights, is the UK's AI Security Institute. It is a global thought leader at this point in thinking about how to build AI systems responsibly and particularly building technical governance to make sure that, as systems are built, they guard against the risks that we all worry about.

The Chair: Thank you very much. Perhaps when Ms Badanes comes to answer the next question she can also tell us if there is anywhere at all in the world that we should be looking at as an exemplar. Let us go to Sir Desmond and after that we will go to Dame Chi.

Q90 **Sir Desmond Swayne:** I have two questions. The first is I want to know, and I believe that I have a right to know, if the results of which I am the beneficiary or the victim have been delivered by processing using artificial intelligence. Is that a legitimate expectation?

Rob Sherman: I think that transparency is a critically important part of people's experience with AI, in part because they will not have confidence in using the services that we offer if we are not transparent about how they work. When people interact with AI on our services, whether through a chatbot or other experiences, we make sure that it is clear that they are interacting with AI.

It is important to note that there are other forms of AI that are used in other ways on our products and services. For example, AI is used to help decide what content you see in your news feed on Facebook so that you are seeing things that are interesting to you and that will be different than what is interesting to me. Similarly, we use AI as a part of our content enforcement and our integrity efforts to keep people safe on the platform. In those cases AI is part of the puzzle but there is also human involvement. Typically, when we take those actions, we also try to be transparent about the reasons why the actions were taken and give people recourse.

Sir Desmond Swayne: Do you have anything to add?

Ginny Badanes: I think it breaks into a few different categories when it comes to what your rights are in that situation. The first is the question of when somebody should know if AI materially shapes an outcome that affects their rights; for example, decisions about opportunities or jobs. You should know that AI had something to do with that process and it is something that you deserve. To the point that was just made, when you are interacting with AI in a way that could be interpreted to be a human, I think you have a right to know that in fact you are interacting with an AI.

It is also important to talk about what transparency actually looks like. This is a complicated situation but disclosures are critical. There are a lot

of details to be worked through about when disclosures should exist and in what way they should be labelled, but a lot of work has gone into this area and there are a lot of opportunities for in-product notices and limitations and citations where you can. I feel that I should add that there is a critical component of content provenance and watermarking that should be considered here as well. This is, of course, more for videos and images and audio files, but there is some technology out there that enables technology companies and the platforms that are serving it up to essentially label when something has been created by AI. That is an area that we should be into as well.

A final point is the who: who is responsible for that? As a reminder, there is a very long supply chain in the AI life cycle. There are responsibilities at the developer side, the deployment side and then at the platform side. I will finally add that there is always some responsibility with users, particularly those who are attempting to manipulate the information that they are using those tools for.

Sir Desmond Swayne: You have touched on what would have been my second question. First, how do we enforce this but, second, how easy would it be to conceal with watermarking, for example?

Ginny Badanes: There is no silver bullet to this question, which is why it is complicated. Yes, watermarking can be manipulated and if it was the only solution to this problem then I think we would be in real trouble. The reality is that there is a defence in-depth approach that is being recommended as a best practice in this case. That is that you should have invisible watermarks, visible watermarks and content provenance applied to it. There is even a case for fingerprinting in certain cases as well. When you put all those things together, if someone chooses to maliciously strip content provenance or does it accidentally, you will have some resilience there because the watermarks still exist, and vice versa.

We need to get to a place, to your point on enforcement, where certainly as the tech industry, but more broadly in collaboration with Government, we come up with these best practices and enshrine them in a way that we are all following the same principles so that we have resiliency in our system.

Rob Sherman: I certainly agree with the importance of transparency. One thing that I will add is the importance of multilayered transparency in communicating with people who use our services in a number of different ways but also recognising that there are different people in the ecosystem who need different kinds of transparency. For example, Governments will be looking for different things than users, and experts will be looking for different kinds of transparency. I agree with the point that was just made on transparency.

When our AI models generate photorealistic images, there is embedded watermarking within the image that identifies it as AI-generated. When those images are shown on Facebook, there is what we call an AI info tool that lets you understand that they are AI-generated and get

information about that. In addition, we try to publish other kinds of information about other AI uses. I referred earlier to AI when it is used to show you different things in your news feed. In that case we have a feature called “Why am I seeing this?” that tries to explain to you the AI-driven decisions about what has caused us to show you that. There is an additional layer for disclosure to Governments and experts.

We publish transparency reports about our overall human rights approach, which includes obviously our work on AI, but also when we release a model we publish details about the specific evaluations that we did, the specific risks that we looked at and how we mitigated them.

Finally, I think there is also a concern about AI being a black box. One of our approaches is to open source our models. I like to think of open-sourced models as a glass box instead of a black box in the sense that even if you do not necessarily trust all the information that we put out—we hope that you will—in addition to all the disclosures that we have about the work that we have done, because models are open source anybody who wants to can access them and understand how they are working.

Q91 Dame Chi Onwurah: Thanks to the committee for allowing me to guest. It is a great pleasure to be here. Meta will be before my Select Committee on 11 March. I am not sure whether that will be you, Mr Sherman.

Rob Sherman: I am not aware of it, so it would be a problem if it were me.

Dame Chi Onwurah: We will be looking at the update to our social media and algorithms inquiry. We have already heard about rights and the multilayers of technology that go into artificial intelligence. An important point for clarity here is responsibilities, particularly when it comes again to the question of regulation. Could you say to me what your company—and I will start with Mr Sherman—sees as your responsibilities when it comes to the impact and outputs, two separate things, of artificial intelligence produced via your AI services and shared on your platform?

Rob Sherman: Thank you for the question. It is an immense responsibility that we feel as a company, and that I feel, as you point out, building technology that enables people to create and share using AI and building platforms on which they can share. That is a real opportunity and I am excited about the value of it, but we also feel a real sense of responsibility for how that works.

I will talk just for a moment about how we approach that. When we are talking about building our models, I mentioned earlier that we do reports when the models are released but those reports entail a significant amount of work before the models are released on what we call alignment. It is basically a process where we identify the ideal behaviour of the model and we work to ensure that the model is aligning to that

behaviour. That can be safeguards around things like privacy and safety. It can also be behaviours that we want around non-discrimination, just as an example.

A big part of the work that we do is building diverse training sets to make sure that the models behave in a balanced and non-discriminatory way. We do that work when we are building the models, we are transparent about how the models are built and evaluated, but then once we ship them we are also doing work in looking at how people are using the models and looking at our platforms to understand how people are sharing and what content they are generating. In that case, AI is a particularly useful tool for ensuring that we can spot harmful content on the platform, whether it is AI-generated or not, and help us before things even get reported to us to take them down. We feel a lot of responsibility for building safe and responsible models and making sure that we have a safe and responsible community in how people use our platforms.

I agree with what Ginny said earlier that, in addition to the role that we play, we also try to build the capacity in the people who deploy our AI models to make sure that they can deploy responsibly. For example, we also open source technology that helps deployers of our models build these protections into their software as well.

Dame Chi Onwurah: You are saying to the committee that you see yourselves as responsible for the impacts and outputs of the artificial intelligence. That is what your answer is. Ms Badanes?

Ginny Badanes: Thank you so much for the question. We have a series of things that we call responsible AI principles. I will start by saying that they are grounded first and foremost in our human rights principles as well. We try to make sure that we do not just have nice principles that we put into place so that I can stand in front of you and say we care about responsible AI. The reality is that it is critical for us to take them and not have them abstract but turn them into concrete obligations that go across the full AI life cycle. In fact, one of those AI principles is accountability, saying that there are accountable people at the end of every decision—that we should not point just to technology because we know that behind that are people and institutions and they are the ones who carry those obligations. We take that responsibility as an end-to-end approach.

We look at things like how models are designed and trained, all the way to how the systems themselves are deployed and how they are used in the real world—how people are interacting with these products. The reason that is important is because responsibility does not stop at launch. We do not just put a product out into the world and say, “We did all our work, we are good now”. We recognise that it requires ongoing monitoring and feedback channels. We need incident response and we need to be flexible and nimble enough to make changes when things do not go the way that we had intended. We need to have that responsibility ourselves so that we can course correct.

We focus on the most salient human rights risks. There is a lot of risk out there to manage and it is important that we start with what is most important. We did an independent assessment a couple of years ago to help us identify the five primary, most salient human rights risks with AI specifically as Microsoft. We identified those as AI development around bias and hallucinations, risks from misuse and unsafe deployment, working conditions in the AI supply chain, environmental impacts and, finally, inclusion and access to AI. Our approach focuses first with those as our most critical human rights risks and we govern around those.

Finally, I will say, and I have already mentioned this, we believe that responsibility follows meaningful control. Essentially, when you look at the different layers of the life cycle, the developers, deployers and platform providers all have roles to play. Where Microsoft has control, for example through system design, our own policies or access controls, we put binding requirements on ourselves for those to ensure that we have, for example, things like heightened review for sensitive use cases or high impact uses rather than relying just on the voluntary principles that we put out there ourselves.

We do not believe that companies can do this alone so I just want to reinforce that we support clear and proportionate regulation and international standards that will make this accountability possible.

Dame Chi Onwurah: Thank you. As a quick follow-up, I think that both of our witnesses are taking responsibility for the impact and outputs of AI. AI is evolving and those impacts may be foreseeable or may not be foreseeable in the long term. What do you see as your responsibility for managing and responding to future foreseeable harms to human rights? Just answer briefly, please.

Ginny Badanes: I will quickly add that we do our best to identify what is around the corner and a large part of that is working with external stakeholders. We do not believe that we contain all the information inside Microsoft to know what the next risks might be, so we have a variety of councils and consultants and people we engage with, the communities that are impacted by AI, to ensure we are getting the signals back in so that we can anticipate and do our best with what might be coming next. Then, of course, we plan to have responses when we miss the mark.

Dame Chi Onwurah: Have you published the future risks?

Ginny Badanes: Yes, we have within our transparency reports.

Rob Sherman: At Meta we have a corporate human rights policy that sets the company-wide standards under which we protect and uphold human rights. As a part of that overall human rights programme, we have built what we call our Frontier AI framework, which is our policy for how we look, as we are building other novel models, at what the risks are of those models and how we guard against them. We publish that on our website so you can see what the process is, but briefly, for the frontier models that we build and any changes that we make that involve AI, we

have an internal risk process where we evaluate the thing that we are proposing to change. We look at the risks, including human rights risks, of any proposed change and we evaluate the benefits of the change and the risks and adopt mitigations before that goes out.

That is a routine part of our process and something that we take very seriously. It has become foundational to how the company makes decisions about changes to our products.

The Chair: Thank you very much indeed. We will move on to Dr Peter Swallow now and after that we will hear from Lord Murray.

Q92 **Peter Swallow:** Ms Badanes, one of the features of using AI, which I am sure you are aware of, is you see a lot of em dashes. There are lots of different theories as to why this may be. Some have suggested that it is because of the information that is plugged into models to help create them, a lot of 17th, 18th and 19th century literature. This is a form of bias, a bias for an American em dash over a good old British en dash. There is also, on a more serious note, a lot of concern about how bias may creep into AI models because of the nature of the evidence that they are being built around. Literature from an earlier period is not known for its modern way of looking at social issues. There is a lot of concern about other AI models spewing very alarming racist, offensive content, based on what it has been trained on. How does your company mitigate and address unintended bias and discrimination in AI model training data? How are you monitoring the outputs to ensure that that bias is not there?

Ginny Badanes: It is difficult to ever ensure that bias will not exist. We are humans and it is inherent in all things that we do. We have an obligation, though, to do our best to mitigate that bias and to have processes in place that will improve that situation. We start by embedding it across the entire AI life cycle.

I will start with the model development. Teams are required to assess the training data carefully and then identify how representative it is, if any gaps exist in the training data, looking for things like groups that might be underserved or misrepresented in that data, because it is not something we can just assume is not there. We have to actively test and look for it because again we are human and it will not immediately be apparent to us.

We then apply what we call systemic evaluation, which means that models and systems are tested and we are using what is called disaggregated analysis, which is our way of saying essentially you cannot just test one way one time and say "This model is good". You have to have different context, consider different groups, have multiple different tests with the different lenses to get an accuracy score that gives you any confidence that the bias has been considered.

Finally, it is not just at the model level because you can often see bias emerge through how the AI is deployed and used. That is where we bring in our high risk or sensitive uses to subject those uses to additional

human scrutiny to ensure, through an internal governance process, that we have not inadvertently introduced bias into a highly sensitive use case, and we have escalation processes around that as well. There is no perfect solution to ending biases in the systems and part of it is being cognisant of that and putting processes in place that can get us into the best place, so that particularly those who are most underrepresented and vulnerable are represented in these models.

Peter Swallow: You are able to have that level of control even though Copilot is powered by other models?

Ginny Badanes: It is an excellent question, because when we have our own models—we do have smaller models—we have more control, but we still run those that are presented within our systems through similar processes as well. That goes to the deployment stage, so that is not the training side, that is not the development side, it is the deployment and then the platform side as well. We uphold, even in those areas, the same principles.

Peter Swallow: Mr Sherman, did you have anything to add, on em dashes or anything else?

Rob Sherman: I am afraid that I am equally perplexed by the em dash situation, so I cannot assist the committee on that particular aspect. More broadly, on non-discrimination, it is a critical priority from a human rights perspective. It is also a critical business priority, because if our models are biased, if they are behaving in ways that underserve some segments of the population versus others, or do not fairly characterise certain views, that makes the product less useful to people; people are less likely to adopt them and choose our products. For both reasons, we spend quite a lot of time doing that. I agree with everything that Ginny said in terms of the training data being critical, making sure that we have broadly representative training data that represents different views, different subsegments of the population, and having technical evaluations to make sure that we are guarding against it.

The one other thing that I would add is the use of both human and automated red-teaming. One of the things that we really tried to do in looking at how the models can be used—and because bias changes in society over time, the way that people may prompt our AI may change over time—was to look at adversarial testing, where we hire people specifically to go and look at whether the model is giving biased answers in some way, but then we also look at how people are using it in production, online and if we see problems coming out. We can score those things quantitatively but also look at them qualitatively to understand if the model is aligning with what we want.

Peter Swallow: What would you say are the core risks to human rights posed by AI?

Rob Sherman: I think that it is a topic of much debate, as I am sure in this inquiry you have heard a lot of different topics. The ones that I think

most about are non-discrimination, which we have just talked about, and the right of privacy, the right of free expression. I think that those are often top of mind when we think about human rights and AI, but there are also a whole lot of others. Keeping people safe is obviously a fundamental priority, so I do not think that I could identify one or a small set of human rights priority, but I know that it is a big topic of discussion in this space.

Peter Swallow: You do not specifically reference there risks to democracy and the democratic process.

Rob Sherman: We certainly do think about that, yes.

Peter Swallow: Are you satisfied that Meta is currently doing enough to challenge the use of AI by foreign actors on social media, where there have been concerns raised about how AI and social media more broadly are being used to undermine democratic rights and freedoms?

Rob Sherman: I would certainly never suggest that the work to do that is done, in part because some of the people who are trying to engage in those attacks are adversarial and they continue to evolve their tactics. They continue to behave adversarially, so we have dedicated teams at our company that focus on this issue specifically. That involves a number of different efforts, one of which I alluded to earlier. To come back to it, it is the role of AI in helping us to identify and combat, for example, co-ordinated inauthentic behaviour on the platform. If people are trying to use our platforms to spread disinformation, we have actually been quite successful in using AI to detect those activities and deter them. We published a transparency report on that activity as well, and I am happy to share it with you if it is useful. Broadly—

Peter Swallow: Specifically on that, because this is something that I wanted to raise—

Rob Sherman: Of course.

Peter Swallow: —obviously, there is a huge role for Meta to spot some of these activities. It is also fair to say that individuals need to get better at being able to spot it. As politicians, we see this in our day-to-day interactions on social media. In recent months, there has been an increase in the number of groups on Facebook where anonymous posting is becoming increasingly the main way of posting into those groups. Anonymous posting means that it is impossible for an individual who sees a post to be able to do any due diligence into the person, or bot, posting it. Are you comfortable—are you content—that this change in the way that people are using your social media platform is allowing them to safely interact free from bots, free from AI, free from foreign influence using these tools to undermine our democracy?

Rob Sherman: It is an important question, and you asked earlier about how different human rights play into our calculus. One of the things that

we thought a lot about—and as you will know, by default Facebook is a real identity platform; you identify yourself—

Peter Swallow: It is not by default, though, is it? In all these groups, you have people posting as anonymous users and there is no way of qualifying for an individual who is responding to those posts, seeing those posts, whether the person at the other side of it is their neighbour from across the road or is somebody in Russia or is from a bot farm anywhere around the world.

Rob Sherman: The design of the service is when you create your account in the first instance—when I create an account, I have to say that I am Rob Sherman, and our policies say that you have to identify yourself by your real name. One of the things we also wanted to do, recognising that people use Facebook groups for lots of different purposes, some of which are sensitive—for example, getting support for emotional challenges that they are facing in their lives, health conditions and those sorts of things—is that we wanted to give people the space to be able to engage in those communities without putting their real name attached to it. In some specific groups, where the group administrator has enabled this, we allow people the option of being anonymous within the group.

It is important to note that in those cases those people are not anonymous to us, meaning that they are still posting under their real-name account. We have their identity and have access to the information, so to the extent that there is criminal behaviour or things like that, we can work with law enforcement. In addition, we have some of the measures that I talked about to guard against inauthentic accounts, bot farms, things like that. We have been pretty successful—I would not say 100% successful—in addressing that problem.

Regardless of whether we are talking about anonymous speech on Facebook or whether it is anonymous speech anywhere on the internet, I would encourage, and I think the company would encourage, people to be thoughtful about the sources of the information that they consume. If you are getting information that comes from somebody anonymous in a health support group, for example, that may be a positive source of emotional support for you. I would not, for example, conclude that that person has medical expertise. Making sure that we are clear with people about what level of confidence they can have in who is talking to them is another important piece of this.

Peter Swallow: I would suggest there is work to do. Ms Badanes, very quickly, because I think the final question is important to be asked, how would the development and implementation of universal rules for AI risk management impact human rights?

Ginny Badanes: It is hard to know exactly how those are going to be laid out. I would say that it is important for there to be some coherence across multiple jurisdictions. It is challenging for both the technology companies to comply, but also for the actual desired impact of those rules

to be effective if they are fragmented across jurisdictions. I would essentially say I am hopeful that we will get to a place where there is more coherence and that the application of universal rights could be a platform for that.

The Chair: Thank you very much. Mr Sherman, in answer to Dr Swallow, you said you had some more information that you could share with the committee. I think we would be very pleased to see more of that policy background.

Rob Sherman: Absolutely. I am happy to do so.

The Chair: Thank you very much indeed. I am always the enemy of the committee, because I am always looking at the clock. My colleagues know that I have to try to move us along. We are going to hear from Lord Murray of Blidworth in a moment, and then after that from Baroness Chakrabarti.

Q93 **Lord Murray of Blidworth:** Yes. We are going to move on now to discuss accountability and AI. I have to put to you some easy questions, and then Lady Chakrabarti will put the hard questions. The easy question is this. If I ask a large language model a question, will it tell me the truth, Mr Sherman?

Rob Sherman: Well, it is certainly designed to tell you the truth. I talked earlier about the alignment process, and one part of that is teaching the models to give factual information wherever they can and to be transparent if they lack confidence in an answer. There are some things that are not fully knowable. I certainly would not say that AI will always give a 100% accurate answer. There is a challenge that exists for our models, and for all models in the industry, called hallucination, where models can give incorrect information. That is why, among other things, it is important, as I just talked about with your colleague, to be thoughtful about the sources of information. It is also why, when our AI gives responses, it cites sources, so that if you want to know where a particular piece of information came from, you can do additional research.

Lord Murray of Blidworth: So between yes, no, and maybe. you said maybe?

Rob Sherman: The goal is for it to be factual and to be clear when it does not have confidence, but that is not always consistently something that the model does.

Lord Murray of Blidworth: Thank you. Ms Badanes, what do you say to that question?

Ginny Badanes: It is incredibly difficult to ask a large language model to consistently provide you with the truth, in part because of some of the inherent flaws of the way the systems are designed. I do expect they will continue to get better, but also because truth is at times subjective it is a challenging environment to guarantee or ensure anything.

All that being said, it is important to focus on how we use these models, what we are using them for and what our obligations are as technology companies to educate and to work on things like media and AI literacy around these, as well as to make it clear what the purpose of a model is. This is where model cards and system cards are effective tools for transparency. If you are going to a model or a system and your purpose is to write a fan fiction novel, then it is maybe not the place where you are going to necessarily expect the most truthful answer; whereas if you are going to a system and the purpose of it is to get the latest on the news of the day, then that is an environment where you have a higher expectation to get truthful information. In that case, you would expect things like citations that link back to real articles that have real, accurate news reporting and information. I do think that we all have a lot of work to do in this space. Part of it is the expectations of what these systems should provide, and part of it is making sure we are working on things like media and AI literacy for those who are consuming it on the other side of the screen.

Lord Murray of Blidworth: As far as trying to generate trust in the answers given by LLMs, you both point to extensive citation by the machine to indicate where the sources have come from. Are there any other steps that you would suggest could be taken to encourage trust, in a way that is obvious to the user of the LLM?

Rob Sherman: I would suggest two things. The first is, as we talked about earlier, making sure people understand when they are interacting with an AI and understanding the context in which the AI is operating. That is something that we try to tell people and give people information about, for example, the fact that AI may not always be accurate. Educating people about that is important.

The other piece, which is less for individuals but is an important accountability mechanism for society, is the technical evaluations that I alluded to earlier. Among the technical evaluations that we run, there are some industry standard ones that look at helpfulness and accuracy and alignment of the model's responses with factual information. You do not have to simply rely on us saying that our models are reliable, although we hope they are. One of the things that independent people can do is they can compare our models with other companies' models to see how well they stack up around this. Again, nothing is 100%, but it is certainly a goal. The fact that there is data published around it is a helpful accountability measure.

Lord Murray of Blidworth: Would it be possible for the model to generate, together with its answer, an assessment of the likelihood of the reliability of the answer it has given?

Rob Sherman: We do try to do that. Again, with the caution that nothing is perfect, one of the instructions that we give our model is that when it does not have confidence, when somebody asks a yes/no question or a factual question and it is not sure of the answer, it can qualify and give a sense of the extent to which it thinks that answer is right, or it can say,

"Based on this factor, it is probably this answer". Similarly, one of the things that you can do, and which I actually do quite frequently when I use our AI, but other companies' AI as well, is ask the AI specifically to respond with a percentage confidence in the answer. AIs are pretty good at understanding what their level of confidence is. If you simply ask, you can get a pretty good sense of that.

Lord Murray of Blidworth: Thank you. I will have to try that. Ms Badanes, have you got anything to add to that particular answer?

Ginny Badanes: No, I agree with what was said there. I would say one area where we have seen that trust is lost is when the system does not answer a question. One thing that we have seen as successful is, when the answers are complicated, giving various context around why. I think that that does build trust, essentially not just saying, "We cannot be 100% sure on this one", but also providing almost the other approach where other people might differ on the answer here, and then sources for that as well, and allowing people to go to those original sources and investigate more thoroughly themselves.

Lord Murray of Blidworth: Thank you very much.

The Chair: Let us turn to Baroness Chakrabarti, and then I have a question for you about age restrictions.

Q94 **Baroness Chakrabarti:** Given the answer that nothing is perfect, if an individual relies on incorrect, fabricated or even manipulative outputs from models and suffers human rights harms as a result—for example, incorrect legal advice—or worse, or find themselves in a suicidal position that they have been encouraged towards, who should be held responsible and who can be held responsible?

Ginny Badanes: I am happy to go first. I have mentioned before that, first of all, accountability is a critical component. It is one of our responsible AI principles. We think that it should be attached to where there is meaningful control. In hypothetical situations, it is difficult to say where the responsibility lies but, if you get into specific examples you will understand, "Was the mistake that essentially led to this at the actual model level? Was it at the deployment level? Was it where it ended up on a platform somewhere? Or was it a manipulative and malicious user on the other end of it?"

A good example of that is we know that AI is being used to defraud people in a lot of different ways. Fraud is nothing new. AI is supercharging it, making translation and access to personal information easier, and therefore using it to create better and more effective phishing tactics. In that case, we would ultimately say the actor who manipulated and used that information should be held responsible. It does not mean, though, that those who are on the product side should not identify gaps where that may have been too easy for them to access and make improvements there. So it can be a shared responsibility, and in some cases it is very clear.

Rob Sherman: As I talked about earlier, I think making sure that our models are safe and responsible is an important, critical priority. It is something that I spend the majority of my time on. Among other things, that means making sure that if somebody expresses intent to harm themselves—that was the example that you gave—we are giving thoughtful, safe, research-backed responses to that and referring people to human experts who can help support them. I think that it is important that the AI not step out of its lane, so to speak, when it comes to that. Similarly, for something like legal advice, we would not want people to rely on an AI model as their lawyer. Similarly, we would want the AI to give people information if they are asking for information, but also encourage them to seek this advice.

Baroness Chakrabarti: That is your approach. You are both describing a best practice approach, but to return to the fundamental question, where the model is giving bad legal advice or is supposedly giving some emotional support but ends up manipulating somebody, who should be held responsible?

Rob Sherman: I would not feel comfortable expressing a view on what the right legal framework is. I think that is a hard challenge for Parliament, and I am glad that that is not my job to decide. I do agree with what Ms Badanes said earlier, that there are different players in a particular fact pattern, and a court would need to look at the relative liability of different people and who was in the best position to guard against that. In most cases, I would imagine there are multiple players there.

Baroness Chakrabarti: Okay, that is interesting. That is multiple players at the corporate end, but my next question is about multiple players at the regulatory end. I think that this was touched on a little bit earlier—that you are huge global corporations, you are operating all over the world, and of course there are different rules in different jurisdictions. I wonder how workable that is from your perspective and how you deal with it. Do you end up complying with the more onerous rules because you have a global product, effectively complying with the more onerous rules even in jurisdictions where they do not apply? For example, the EU AI Act is one of the more robust approaches. If you are complying with that, do you end up complying with that all over the world?

Rob Sherman: There are a couple of pieces to the answer to that. The first is, obviously, we operate in most countries around the world, and we have responsibility to our community in each of those countries. Our human rights policy particularly applies globally. It is a single policy for the company. What that means is that some of the programmes and processes that I talked about, including doing risk assessments before we make product changes and those things, apply globally. We also have what we call a regulatory readiness process, which is a comprehensive process for the company to look at new legal obligations that are introduced in different countries around the world. The approach tends to be that we try to take a thoughtful and responsible approach to everyone

globally, but we also recognise that different countries are going to have different obligations.

Just to give you one example, here in the UK, the UK GDPR gives people a right to get access to information that we hold about them. That is not the law everywhere, but because we did it to comply with the law here in the UK and some other jurisdictions, we extended tools to give people that access globally because we thought it was a responsible and good thing to do.

There are other requirements. For example, you mentioned the EU. The EU GDPR has different disclosure requirements that do not apply to British citizens. It would not make sense for us to apply the formalities of the EU GDPR to people here in the UK. What that means in practice is we generally try to take a broadly responsible and safe approach globally, but then also recognise we have specific obligations in different places.

The last thing I will say is that the details matter quite a lot. We have talked about different roles that companies like Meta or Microsoft play. Sometimes we are building an AI model, sometimes we are running a service like Instagram. There are lots of different things in between. Dealing with regional complexity is easier in some places than others. For example, we are not building a separate AI model for the UK and a separate AI model for France and a separate AI model for the US. We do need to take a consistent global approach to doing that, and the specifics do matter. It makes our lives easier if there is consistency across borders, but we also have a process to deal with the fact that different countries have different expectations.

Ginny Badanes: I agree with a lot of that. Right now, we are not looking at a lack of activity in the AI governance space. I would say the bigger challenge we are seeing is fragmentation. Certainly, the fragmentation is complicated for technology companies, but the reality is that we are big companies; we will figure it out. I have concerns about creating actual risk. Essentially, I worry that at times when we have this variety of approaches, we are not actually addressing the broader safety or human rights risks that are at the centre of what everyone is trying collectively to solve. There is just a reality that all this stuff is crossing borders. Essentially, that is the reality of what is trying to be governed here.

Everything about advanced AI is transnational by design. The systems are developed and tested and deployed in a variety of places across borders and within multiple supply chains, and then integrated into products that are used at a global scale. This is a global challenge that I understand people are trying to regionally approach. I mentioned before that the more we can align ourselves to international standards and norms and have, essentially, a base layer of agreement—whether that is the Hiroshima approach, which we have mirrored a lot of our work around as well—that is a strong place to get us all out of these fragmented models and into a place of agreement for foundations and then build from there.

Baroness Chakrabarti: I am getting a sense that you would actually welcome a greater harmonisation at the global level in terms of regulatory standards.

Ginny Badanes: Absolutely. We would like to have some cohesion or some coherence across the different regions. That would be welcomed from just an ease of use perspective, certainly, as a technology company, but also to have more impact of what we are trying to achieve.

Rob Sherman: One thing I would just add is that the Chair, in his introduction, mentioned that Mr Narayan was just at the India AI Summit this past week. I was there as well. One of the goals of those summits is for Governments to come together to talk about some of these challenges and to talk about the approach. Year over year, we have had four of these. The first one was here in the UK. Year over year, the technology is vastly different each time. Giving Governments an opportunity to talk, certainly to us but also to each other, to make sure that there is consistency in the way that they are approaching it, that they are sharing information and that we are working together as a community, is a critical piece.

It is a little bit different from your question around legislation, but I think it is a really important area and probably one that we could do as a global community to make more investment in. The technology is moving so quickly that this informal collaboration will be critical.

Q95 **The Chair:** Thank you very much. In fact, I would like to build on that question before I turn to my colleague, Lord Rook, for the following question. I would like to ask you specifically about how you are collaborating with others within the world. I have in mind particularly Google here on the issue of age restriction and the parental controls that it seems to be breaking by saying that it would allow the turning off by young people of parental controls. We followed that up as a committee when we had Google before us. I want to say on the record that I am extremely grateful for the long statement that it has now sent us answering the questions we raised.

Let me put a couple of points that it has made to you to see whether you are taking the same approach. It says that, "We are continuously evolving our products to provide age-appropriate experience and enhance parental controls". It admits that there needs to be a change in this idea that you might enable children to turn off parental controls. It says, "Under our planned policy update, any supervised minor will have to get parental approval before they can turn off supervision". It says, "Our focus remains on empowering families with the tools they need to navigate the digital world safely". Elsewhere in this long statement, it says, "This year, we began testing a machine learning-based age estimation model in the US ... We believe actors across the digital ecosystem"—ie, Microsoft, Meta and others—"have a role to play in ensuring age-appropriate online experiences".

Let me ask you this. One way to protect children from adverse outcomes

might be for the fundamental human rights to be protected by age controls across products and services that use AI. In what circumstances should AI systems and products be subject to age restrictions? Are you also, along with Google, looking at ways in which that can be managed? Mr Sherman, we will start with you.

Rob Sherman: This is a critical area of focus. It is something that is essential to us as a company. Making sure that young people are safe on our platforms is critical. I should also say that I am, myself, a parent. I have an 11-year-old and a 14-year-old at home, so I am in the age range where these questions are particularly salient for our family and we have lots of conversations about them. I think about this a lot. This is a thing that companies like Meta—Google and Microsoft I am sure would include themselves in this—have a responsibility to do work around.

We have taken a number of steps to help keep young people safe on our services, including within our AI. I should say that a lot of this has been informed by the work that has been done here in the UK. Obviously, the Online Safety Act, but before that the Age Appropriate Design Code, was very influential in our thinking about how to build multi-layered protections for youth online. For example, we rolled out teen accounts on Instagram. If you sign up for an account on Instagram and you say that you are 16, by default you will start with a more restrictive experience. If you are an adult, your account is—

The Chair: How do you verify that?

Rob Sherman: We have a number of measures in place to do that. The first is, obviously, we ask. You mentioned the use of machine learning. This is something we have been doing for several years at this point, using machine learning models to predict if somebody is likely not telling us the truth about their age. We can use those additional signals.

The challenge that we face—and, candidly, that I face as a parent—is that each app is doing that individually. I think that it is certainly a responsibility of a company like Meta to make sure that we build age-appropriate experiences into our app. The teen accounts that have a more private, more restrictive experience are a critical example of that. We also have a responsibility to build parental control so that parents can make the choices that are right for them. I think that this point you are raising about how we validate age is a challenging one, and it is an area where—

The Chair: I am sorry to interrupt, but is it true that even things like library cards are used as a form of verification? That does not seem very robust to me.

Rob Sherman: We have a number of different methods that we can use for age verification. Typically, we use birth certificates or government identification. We also have some AI-based validation methods that we can use. One of the core challenges in the industry, which is not a Meta-specific challenge, is, as a parent, I have to go from app to app and do

different verification methods in various places. Some of them are more reliable than others. Some of them require me to give my credit card. Some of them require me to give my personal information or my child to give her personal information. It is just inconsistent. For me as a parent and for us as a company, where if it is an app on Android or if it is an app on iOS, that would be done once at a platform level and then we would just be told, "This user is 16" or "This user is over 18", or whatever it is, and then everybody would build the appropriate experiences based on that. That is not something that exists in the ecosystem now that I think would be valuable.

The Chair: Okay. That sounds as though we want to explore that a little bit further. Ms Badanes, could you perhaps build on that for us?

Ginny Badanes: Sure. This is foundationally important when we talk about human rights, to talk about the most vulnerable and the highest-risk communities, and I am not sure that you can get much higher risk than children and youth, so it is critically important. The other thing to note is that with AI you will have a variety of different experiences. There are lots of different programmes. This is getting to the previous point. Their experiences are going to vary wildly depending on the system that they are using and the context of which they are using it in the first place.

I am going to sound like a broken record, but it is why we talk about the risk-based approach rather than trying to apply something like a single age threshold across all AI tools. We think that we have stronger safeguards to apply where the risks to the children are higher. There are higher risks in certain scenarios than in others. For example, a chatbot where they are forming relationships will be a higher-risk scenario than, potentially, a tutoring app. That does not mean that there should not be restrictions and mechanisms around the tutoring app, but you do need to take the context into consideration because some are higher stakes and more sensitive than others. From an operational perspective, it is not just about restricting access. We also need to build these age-appropriate design and safety guardrails. Again, I am just echoing what was said previously, creating clear boundaries into the system from the very beginning.

I do not know that I have much more to build on than that. It is a complicated situation, but if we start with a high-risk user and then a high-risk scenario and work from there, I think that we will find that we have some effective strategies.

The Chair: To help the committee, can you tell us what products you have that do have age restrictions on them? Are there any negative aspects for human rights about imposing such restrictions?

Ginny Badanes: We, of course, have Xbox, which is a gaming platform where you do have a lot of younger people. Going back to the point of having children, you can maybe see from behind me I have three boys, and they game and use Xbox, so this is something that is important to us

as well. It is important regardless of whether I have children or not, but I do experience it first hand. Yes, we do have age verification measures within Xbox, in large part in response to the Online Safety Act, where kids are directed to age-appropriate experiences based on the age that they have validated when they went into the product.

You asked about trade-offs, sorry. There have been trade-offs that we have seen so far. I think that we are still gathering our key lessons. It is an area where I am not as directly connected with our digital safety or Xbox teams. That is something I am happy to give you some lessons learned back in writing if that is helpful. I do know that we have experience that it can create some friction, both from the parents and the kids. Of course, we want to value their experience, but we also want to make sure that the kids are safe. I do think that we have some lessons learned in that area.

The Chair: We would be very interested to see more on that if you could share it with us. Thank you very much indeed. I would like to go now, if I may, to my colleague Lord Rook.

Q96 **Lord Rook:** Ms Badanes, I am particularly glad to ask someone from Microsoft some questions. We had a general election here in 2024, and a few hours later I found myself on a plane to Hiroshima for a meeting convened by the Vatican about the ethics of AI. I got to sit next to your president, Brad Smith, for the day, and I thought, "I am going to ask him lots of intelligent questions", but I soon found out that the combined effects of jet lag and election lag had turned my brain to jelly. So I am going to ask you the question that I probably should have asked him, which is: should there be an opt-out for individuals who do not wish to be subject to AI in the services, products and technology they use? If there should, how are you progressing that? If not, why not?

Ginny Badanes: This is an area in which there will be additional exploration, and there is a lot of opportunity here, which means essentially it is not perfect yet. I would say that the best place to start is at the specific user and feature levels. There are a lot of those features currently available where you can opt out. It is hard at this moment to know from a data perspective and every single thing that you do where you can opt in and out. It is something I would love for us to explore with you all, as you have thoughts on it.

There are some good practical ways that you can opt out of engagement. There are also ways that we have seen from an opt-out perspective—I recognise this is a little different from your question, but I just want to address it—which is around the use of data, for example, from organisations. There are features where a website can choose to opt out of having its content made available for training of models, that kind of thing. There is progression in the space of opting in and out. I will be completely honest; I think that it is an opportunity for additional growth.

Lord Rook: Mr Sherman, I am interested in your thoughts on that. Are we at the point whereby AI is so embedded in some of these systems

that it would be impossible to opt out because it is just part of the mesh of how different technologies that we are using work?

Rob Sherman: When you look at services like Facebook and Instagram, AI has been a part of them since the beginning, pretty much. I have been at the company for 14 years, and AI has been a part at least as long as I have been there. We talked earlier about some of the wide range of ways that AI is used: choosing what you see in your newsfeed or helping to deal with spam and do spam filtering. I should say that if you project forward a couple of years, I think that it will be even more true in more segments of the population. AI will be embedded in technology in the way that internet is embedded in a lot of our lives today. I do not think that opting out of AI as a technology is probably realistic.

I do think that it is important for people to have control over what experiences they have and how they engage with technology, which is why, for example, you may choose that you do not want to use Meta AI or a chatbot when you use our services. You can still browse Instagram and you can still engage with the people who you are friends with and do all that and then just not choose to use that part of the service. I certainly think that that is available, but I would caution against an idea that we can wall off AI from the rest of technology.

Lord Rook: Moving to a slightly more apocalyptic zone, in 2023 a statement from the Center for AI Safety, signed by many leaders in the tech industry, warned of possible risks of extinction from AI. I am interested to know how you are managing existential risks to human rights, such as human extinction, for instance.

Baroness Chakrabarti: For instance?

Lord Rook: Yes, I know, just as a starter, then you can develop this from there. If not foundation model developers, who should be responsible for addressing some of those serious risks that this powerful technology could create?

Rob Sherman: It is certainly a discussion that we have in our company, in the industry, and it is a discussion in the AI space more broadly. One thing that I should say before getting to the specifics of your question is that my company and other companies often talk about super intelligence or AGI or some of these big terms as though they are a binary—that today we wake up and there is no AGI and tomorrow we wake up and there is AGI and things have vastly changed. The reality is maybe a little bit less exciting and a little bit more mundane, which is that the technology will continue to iteratively improve, day over day, week over week. Along with that, the possibility and the opportunity of the technology will increase. The risks also will increase, and our ability to use technology to guard against those risks will all increase. So, I do not think that we are in a situation where we are going to wake up one day and the world is vastly different. I do think that there is time to look at technology and evaluate it to make sure that we understand what the

risks are as things are being developed and as they are deployed and to build safeguards. In general, I think that is important.

I will say that for us, or any foundation model developer, thinking about the full range of risks is important. We have teams at our company that literally just think about this. I think that I referred to it earlier, but we have published our Frontier AI framework, which talks about the programme that we have in place to evaluate models before they are put into the world for things like chemical and biological weapons risk, cybersecurity risk or these kinds of things. It is an important part of our process, both before and after we ship models, to look at those things. In addition to us doing that, we are working with Governments and we are working with experts in the private sector that we hire to help advise us on these things, so that is critical.

There are two things I would leave you with. First, I referred earlier to another adversarial threat. I think that this is a thing that we need to continue to work with Governments on. Governments have intelligence that we do not always have. Making sure that we are collaborating is an important priority. In addition to this, I worry sometimes that, while the risks that you are talking about obviously are extreme, they are also remote. When I think about where companies like mine should spend their time, they need to think about those risks but they also need to think about some of the more here-and-now risks that we have talked about earlier today. We feel a responsibility to think across that spectrum.

Lord Rook: That is really helpful. Ms Badanes, are there areas where Microsoft is concerned or working around the potential for existential risks from AI and how you combat those and put safeguards in place?

Ginny Badanes: This is a primary, key example of where these international gatherings and frameworks are so critical and important, and they bring multiple stakeholders together. While this may be a low-probability example, it is still extremely high risk, as I think we have all acknowledged. That is why the Seoul AI summit that brought together a framework around how we would respond to this existential risk statement was so critical. A lot of what was just said by my colleague at Meta is what I would say, too, and that is in large part because I think that we are all building from that same framework.

We also put together a frontier governance framework, which is what we call ours, where we look over the risks that are coming from chemical and biological risks, cyber operations, advanced autonomy, and loss of human control. These are the categories that were identified as the most existential risks. We have also put together our own testing, red-teaming, efforts around what those would look like. That is similar to what you are hearing from others in industry, which is, I think, a promising approach that says that we are all adhering to this international framework that included expertise from a lot of different areas. As part of that commitment, we are also engaging with independent experts—people outside of our company, people of expertise in these areas—to ensure

that we are informing the way that we are approaching our own governance framework. More than anything, it is a strong endorsement for why we see value in creating these broader frameworks that we can all align to and then execute, of course, independently, relative to our own products and services.

Q97 The Chair: Before we go to Sir Desmond Swayne for the final question, I want to go a little bit further on what Lord Rook has just asked you about. I mentioned Geoffrey Hinton's remarks at the beginning of our proceedings. He is the godfather of AI after all, and it was he who coined the phrase about existential risk. Some of us have also read, as a cheery thought over the Christmas period, the book, *If Anyone Builds It, Everyone Dies*. Now, we are not apocalyptic here, but we are here to defend the human rights and the safety of our citizens and to recognise that there may be real threats. When we see people leaving the industry, including people from your own companies and those associated with others I have mentioned in the course of the proceedings, because of their concerns about what AI represents, and when we are well aware of the malign actors in the world who do not maybe share the ethics that you have been expressing here today, what more can we do to ensure that there is a robust international architecture to defend us as best we can from the potential threats that others have identified? Ms Badanes, would you want to go first?

Ginny Badanes: Sure. It is never ideal when someone leaves a company where they feel like for some reason they were not able to be heard in that environment. While this is a very small thing, one thing that we have implemented and put into place because of this concern is a very specific, simple tool where employees can go and if they would like to express concern about the way our products are being either built or implemented or even used by our own customers, where they believe that it goes against any of our human rights commitments or our responsible AI principles, they can do that and they can present evidence and information they have. This is, again, a small tactical thing, but it is important that we make sure we do not close out dissenting opinions within the companies, because as we are building these technologies it is critical that we identify where the concerns are coming from.

If people feel like they are not being heard at the company and they have to go externally in order to do that, that is not ideal. As far as ways that we can improve on it, there are a lot of complicated challenges that we—and I say “we” as a big, large society—have been able to resolve that have had similar roots. I would look at how we addressed nuclear challenges initially. I would look at maybe less existential risk but still clearly complicated challenges like airplanes. How have we ensured that an airplane can take off in one country, fly over five others and land in another country, all committed to similar standards and agreements of safety and what that looks like from an interoperability perspective? We have frameworks we can build on; I do not think that we are starting completely from scratch. This is not the first time we as a society have

addressed these existential risks, but of course there are going to be new components that we need to incorporate into them.

The Chair: That certainly gives us a more optimistic view and a rather more realistic one as well. Mr Sherman, do you want to quickly add to that, and then we will go to Sir Desmond.

Rob Sherman: I obviously cannot speak to why individual people left other companies. There are three godfathers of AI, and Yann LeCun, who until recently was our chief AI scientist, is also one of those. He recently left the company, not to leave the industry because of concerns but to go and build a start-up focused on a different AI development. One of the things that he has said is that actually the concerns that you alluded to from some of his colleagues are not going to come from the current generation of models at all. Again, I am not one of the godfathers of AI, but I think that what Yann would say is that actually the current generation is relatively nascent relative to where we are going and the technology. I think that he would say that this is not the thing that he is most worried about right now.

That said, you asked the question of what more we can do. There is a critical answer there. I alluded earlier to the UK AI Security Institute, which is a leading voice in the global debate on these issues and is the thought leader globally when it comes to this. What it has been focused on is technical evaluations. How do we not only identify the risks that we would be worried about, but actually build technical measures where we can look at a particular model and see whether it presents the kinds of risks that we are talking about and, if so, take action to mitigate those risks, withdraw the model or whatever the case is? One of the things that it has started doing through a network of safety and security institutes globally is collaborating on doing these evaluations. For example, this past year, the UK AISI, along with some of its colleagues in other countries, used our Llama open-source model as a test, in addition to some other models, for cross-country collaboration on testing. They were able to build a methodology where Governments could collaborate with each other to do the risk evaluation that we are talking about. I think that is an important area to build on.

Q98 **Sir Desmond Swayne:** Imagine that you are in the shoes of Lord Alton, the Chair of this committee. What would be the two most important recommendations that you would want to see in our report?

Rob Sherman: I appreciate, first of all, the work of Lord Alton, but also the broader committee. This is, as I said in my introduction, a critical area of focus and I am just grateful that the committee is spending time on it.

I have two suggestions, one building on the point that I just made about the safety and security institutes globally. I think that there is a real opportunity when we think about the risks that we are talking about here to build consensus across Governments and with companies on the technical solutions, either to evaluate whether risks exist or to mitigate

those risks when they exist. There is a real opportunity to invest not only in identifying those risks but in building technology. I think that will help companies like ours to align our approach to global expectations around human rights, safety and other considerations. So, that is one.

For the second one I am going to the bias question that your colleague asked earlier. I alluded to this, but it is an important point that often gets overlooked. AI models are trained on a large volume of data, and inherently they know what they are trained on. One of the things that we spend a lot of time thinking about is how to make sure our datasets are broad, and broadly representative, so that they can serve not just the majority in a community but all the different parts of a community. That is something that is primarily our responsibility but, in thinking about the equities that the committee identifies and where public policy can also go, I would encourage you to think about how to encourage that.

Sir Desmond Swayne: What would you like to add, if anything?

Ginny Badanes: I have two. One is a bit high-minded and not too different from what you just heard, and then the second is a bit more tactical.

This will sound like a broken record again, but I would encourage you all to adopt interoperable risk-based obligations for high-impact and high-capability AI systems.

The second one, which is a bit more tactical, is focus on content provenance. I know we talked about this earlier, but I will come back to it. When we are able to embed content provenance and watermarking and AI labelling experiences—once we all start doing this in a more collective way—it will really have an almost immediate impact for end users who are trying to make sense of what they are engaging with online, in a world that is quite confusing to them right now. I think that is one fairly easy way, if done collaboratively, that we can start to have impact for real-world people.

The Chair: Ms Badanes and Mr Sherman, thank you very much indeed, on behalf of the Joint Committee on Human Rights, for giving us such an interesting and lively set of answers to the complex questions that colleagues have posed. Our clerks may be in touch with you afterwards because this is always the beginning of a conversation rather than the end of one. We have told you, and Sir Desmond has just mentioned, that we will be now in the throes of writing our heads of report and then the full report once we have met the Minister, who comes in next after we conclude this part of the session. We would always be pleased to hear from you if there is more that you feel you would like to say that we did not cover sufficiently, or some things that were perhaps off-piste but nevertheless ought to have been on-piste and that you would like to include in the evidence that you have given us. So do not see this as the end of the conversation, see it as the beginning. Dame Chi mentioned to you also that her own committee will be looking further, as other parliamentary committees across the House are. For now, thank you very

much indeed for giving us your time, your expertise and your knowledge. It has been greatly appreciated. I bring this part of the proceedings to an end.