HOUSE OF LORDS

Communications and Digital Committee

Corrected oral evidence: Large language models

Wednesday 8 November 2023

2.25 pm

Watch the meeting

Members present: Baroness Stowell of Beeston (The Chair); Lord Foster of Bath; Baroness Fraser of Craigmaddie; Lord Hall of Birkenhead; Baroness Harding of Winscombe; Baroness Healy of Primrose Hill; Lord Kamall; The Lord Bishop of Leeds; Lord Lipsey.

Evidence Session No. 8 Heard in Public Questions 64 - 72

Witnesses

I: Dr Moez Draief, Managing Director, Mozilla.ai; Irene Solaiman, Head of Global Policy, Hugging Face (formerly of DeepMind); Professor John McDermid OBE, Chairman, Rapita Systems, and Professor of Safety-Critical Systems, University of York; Dr Adriano Koshiyama, Co-Chief Executive Officer, Holistic AI.

Examination of witnesses

Dr Moez Draief, Professor John McDermid and Dr Adriano Koshiyama.

Q64 **The Chair:** This is the Communications and Digital Committee, and we are continuing our inquiry into large language models. We have four witnesses today. We will be exploring the implications of open versus closed-source models and, I hope, understanding the open-source proponents' concerns that moves to introduce safety and testing requirements might inadvertently or otherwise introduce barriers to new market entrants and entrench big tech. I hope we will also hear in today's session some practical options for dealing with some of those challenges. Can the four witnesses introduce themselves and the organisation they represent before we get going?

Dr Moez Draief: Hello, I am the managing director of Mozilla.ai.

Dr Adriano Koshiyama: Hello, I am the co-chief executive of Holistic AI.

Professor John McDermid: I am from the University of York, where I direct the Assuring Autonomy International Programme.

Irene Solaiman: Hello, I am the head of global policy at Hugging Face.

The Chair: I thank all four of you for giving your time to join us today.

Q65 **Baroness Fraser of Craigmaddie:** Welcome. This is a really important session today to give us an idea of the debate on open and closed-source models. I will start with you, Dr Draief. We know that it is a sliding scale, but can you summarise the implications of the policy in this area? As regulators, does it have to be either/or? How should we view this?

Dr Moez Draief: It is clearly important not to see this as an either/or. Before going into the debate itself, I want to tell you a bit about what open source is and where it exists today. Open source enabled us to arrive at the innovation that we see in large language models through open data, open science and open libraries that have been used, and without those we would not have large language models.

There are other technologies that underpin the digital world that are based on open source, such as cloud computing, with Linux, and the 5G infrastructure. I represent an organisation that came up with the first open-source browser. It was considered radical back in the day, but now most browsers run on open source. Open source is not the problem in the space of AI but part of the solution, so it is important to consider it, as you were saying, as a gradient of openness that will enable different industries and individuals to choose whether they want proprietary or open-source models to power their solutions.

Baroness Fraser of Craigmaddie: Coming back to the view of the regulator, regulators often try to be technology neutral. Is that sufficient? Also, it would be an awful lot easier for regulators to deal with closed-source models, because the guard rails are up.

Dr Moez Draief: Are you alluding to safety?

Baroness Fraser of Craigmaddie: Yes.

Dr Moez Draief: Open source provides an opportunity for many people to examine those technologies, to test them in a variety of settings and to provide fixes to problems that arise. From a practical perspective, it is very useful to have open source as a means of creating transparency around the technology. If we were to rely on a few engineers in a certain part of the world to define safety, that would limit the opportunities to uncover problems. Security through secrecy and obscurity is not the way forward when it comes to understanding this technology. So I urge regulators to think carefully about the impact of open source on transparency, competition and access, which I am sure we will talk about in detail.

Baroness Fraser of Craigmaddie: Just to push a little bit, how open does open source have to be to fit into that definition? What is the minimum on that sliding scale to be classed as open, in your opinion?

Dr Moez Draief: The technology has different components to it. There is the data that the model is trained on, the process of training the model, the model itself, its weights, and the evaluation processes that are used to decide whether something has good performance. The sliding scale has many dimensions to it. This will depend on the context of the application and on the jurisdictions, the industries, and how to protect the population. It is difficult to define the scale. It can go from "everything open" to "everything closed", and, depending on the capabilities and the confidence we have in these models as we develop them, we may open any of the dimensions that I described earlier.

Baroness Fraser of Craigmaddie: So you would not point, for example, to whether the source code and the training data are freely available as two specific areas without which you could not call yourself open source.

Dr Moez Draief: You can open them without making them available. We can have APIs to examine them, and we can describe them in detail. I think Irene will speak later about some of the work that Hugging Face has done in this space by creating more transparency around the data, the needs, and the processes for cleaning them. I know you may expect a number with regard to openness or closedness. I cannot give that number, but I feel that we are working as an industry to come up with better ways of opening the model. More open is better, but there are limitations to how much we can open things.

Baroness Fraser of Craigmaddie: Irene, do you have anything to add?

Irene Solaiman: I have so much to add.

Baroness Fraser of Craigmaddie: We are under a slight time limit, so if you could help us by being succinct, that would be appreciated.

The Chair: It is helpful to us if you have things that are different from what we have just heard or if you disagree, rather than simply repeating or restating what has been said.

Irene Solaiman: I strongly agree, and I have published a piece on a gradient of options that I am happy to share after the session. To be specific about the key components, exactly as my colleague at Mozilla is saying, there is no specific definition for open source regarding language models. There are a lot of parallels that we can draw with open-source software, but when we get into components such as training data and technical papers that share how the model is trained—these look specifically at language models that can be extrapolated to different types of generative systems—there is a level of access that researchers will need in order to improve systems, look at the data and look through the code, depending on the risk and the use case.

However, what is really important in openness is disclosure. We have been working hard at Hugging Face on levels of transparency such as model cards and datasheets for documentation to allow researchers, consumers and regulators in a very consumable fashion to understand the different components that are being released with this system. One of the difficult things about release is that processes are not often published, so deployers have almost full control over the release method along that gradient of options, and we do not have insight into the pre-deployment considerations. We are very transparent at Hugging Face—we have, for example, a big science project releasing BLOOM and documenting everything—but there is no set standard on how to conduct that process.

Professor John McDermid: Very briefly, to complement that, policymakers and regulators have no choice but to deal with both open and closed-source models. The Law Commission's work on autonomous vehicles, for example, talks about a requirement to collaborate, and I think that will be needed where there are closed models to deal with problems and rectify them. That is one way to think about it: having more access in the open source, but in both cases the requirement to collaborate will be needed to deal with problems.

The Chair: Not to pre-empt anything that you might say later, but does regulatory capture worry you in this context?

Professor John McDermid: I think regulation will be very difficult. It will be a real challenge for the regulators, in resources and in skills. The Government need to help the regulators build that skills base. I may say a bit more about that later. Briefly, it can be done, but there is a really big road for them to go down, and they need help in doing that.

Q66 **The Chair:** I want to zoom out a little before we move on to other more specific questions. On the debate between open and closed source, I have a question for Dr Draief—I am happy for others to add to this as well. Let us cast our minds back to when the internet first arrived. It has already been said that the internet is open source. Could you remind us how that has remained open source? Was this decided through a battle in the courts, or was some policy decision made by a Government somewhere that protected this as an open-source technology?

Dr Moez Draief: I am not an expert in the history of open source, but I can give my perspective on it in various industries. In general, the industries have found benefits to open source, because it enables the innovation to be shared among the big players, be it in the internet age or today in the AI age. Once this matures, further downstream, the less savvy companies that may not have expertise in AI or the web can use open source as it becomes more mature and can build on top of it.

This is the virtuous circle with open source: it enables further access. The more people are interested in that technology, the more they make it easier to deploy, maintain and generalise, benefitting many more people. I presume that would happen for the web. I do not think there have been major policy discussions. I am sure that my colleagues in the early days of Mozilla, who were part of that debate between closed source and open source and the browser system, can provide more information later, if you are interested in that topic.

The Chair: Do any of our witnesses want to give us a view on the extent to which the business models of Google, Microsoft and Meta might influence their views on open versus closed source?

Dr Moez Draief: These companies contributed a lot to open source in the past. Currently, the competition that started last year is getting everybody to become more closed. These companies used to publish many papers and frameworks used by academics and other professionals. I used many of the tools of some of the companies you are talking about to conduct research and develop AI products for UK businesses. There has been a recent shift from being open to being more closed. I hope this does not become the norm, because we will all suffer from it. A country like the UK benefits a lot from open-source technology in the absence of big technology players in the country. It would be extremely beneficial to the UK if open source continued to thrive with the right guard rails and a community that is ensuring the safety of this technology.

The Chair: Can you give us an illustration of why that would be particularly beneficial to the UK economy?

Dr Moez Draief: In practice, a lot of the tools that are used currently in the UK rely on classification algorithms. We are not talking about LLMs; we are talking about older generations. A library called scikit-learn has been developed by a consortium of American and French academics that is very popular, extremely rich and very safe. Many businesses and government agencies in the UK are using these technologies. I can tell you about a number of other technologies on the code side or on the data aspects that have been very practically useful in academia and industry in the UK and other countries.

Q67 **Baroness Harding of Winscombe:** Following up on the question about regulatory capture and spelling it out more simply, how much should we be worried that the arguments in favour of closed models are entirely commercial self-interest on the part of the owners of existing large language model technology, and vice versa? How much should we be

worried that the arguments in favour of open source are also entirely self-serving? Is that what is really happening, rather than a policy debate? I am purposefully being quite provocative to get a reaction.

Dr Adriano Koshiyama: I remember a client, a big IT corporation, that was trying to make a case for why closed-source cloud AI solutions are probably better than open-source ones. I was trying to think of how it could couch an argument for that. One argument that unfortunately happens a lot in practice, and we have seen this with clients and vendors, is nothing to do with language models, because it is a new phenomenon: facial recognition software, which is widely available on GitHub, open source.

Academics have done so much research over the years on models that are available on GitHub, and there has been so much evidence of bias with respect to race, gender, age—you name it. Those models have not been fixed; people are still using them in production, and unfortunately things have been perpetuating since then. Even though there is transparency around the risks and academics have done research, it seems like no one has managed to go there and say, "Let's create a new open-source version in which those problems have at least been able to be fixed somehow, if possible".

The argument for the closed source was to say "Well, we could have a facial recognition software that's closed source, but then we use a third party or some other organisation that can test this, so that when we provide our customers with this, we are ensuring that the system is safe before deployment". It is a company, so it would take liability in case a problem emerged. So the argument for closed source is that we can take liability and responsibility and bring a company to test. When it is open source, who knows what can happen. That was the argument on the closed-source side. Open source also has its own arguments, too. It depends on the application and on the case we are discussing.

To finalise, in the case of large language models, it is not a new technology, but it has become quite popular and has really spread itself out. It is very difficult to see whether the solution is closed or open source at this point in time.

Professor John McDermid: I have a slightly different point of view. If something is open but is so complex that you cannot make any sense of it, I do not think that actually helps very much. What we really need to say is, "What information do we actually need?" Do we need something about the distribution of training data and whether that is relevant for the application domain? If it matches the distribution of that domain, it should be relatively free of bias. If it does not, that is where there will be risks of bias.

We need to be able to ask about openness at the right level and of the right information. I do not want to look at 2 million or 2 billion neuron weightings that will not tell me anything. However, I can help people to understand what things they need to extract from the models or the development process in order to make sensible judgements about risk. That is where we need to work.

Irene Solaiman: It is important to recognise that language models are still an evolving research field and researchers need access. I am speaking from industry, and there is always some level of scepticism that regulators should take coming from industry. When looking at conferences like the ACM Conference on Fairness, Accountability, and Transparency and RightsCon—I am a researcher myself—I am seeing, frankly, fear from researchers that they are unable to conduct especially the more social side of research, even through simple APIs.

This is a distinction between openness and access. A larger language model, which will often require more compute to run, may not be accessible to many researchers unless they have that compute infrastructure, and even basic query API—just being able to ask a language model for a response—may not be enough access to conduct the research they need to do to improve a model along some axis, such as bias.

The Chair: Okay. If Baroness Harding is happy with that, we will move on.

Q68 **Lord Hall of Birkenhead:** Can we talk about safety testing systems? Irene Solaiman, perhaps I could start with you, although I would love to hear from all our guests today. Should there be pre-release safety testing systems that are mandatory for models that meet a certain set of criteria? If there should, what are those criteria? We saw President Biden suggesting some of those criteria in his executive order last week. If you believe that there should be mandatory safety testing, what should those tests examine and what are the trade-offs? I suppose this is at the back of the mind, because we have been talking about open versus closed. Could this be to the detriment of open-source models? Irene, I would be grateful if you would kick off that answer.

Irene Solaiman: Yes, absolutely. Pre-deployment risk assessments are necessary and one issue is that some risks cannot be assessed until deployment. We see this, for example, even in the labour impact. We will not understand fully the labour impact of language models until they have been integrated into the economy for a set period of time. This is where post-deployment assessments, iteratively, are also important. Under that, what should be tested under the safety umbrella?

I am working explicitly on the social harms side, so we are looking at biases and environmental impact. We are really seeing how difficult, and frankly impossible, complex social issues are to quantify or to robustly evaluate in a technical system. We have a lot more literature for language than we do for other modalities. For example, for image-generative models, which are not the focus for today, we just have far fewer evaluations that we can run throughout the deployment process. Again, this is an ongoing research question that will need more researchers to develop this assessment suite throughout the process of deployment.

Lord Hall of Birkenhead: What are the trade-offs, then, if you can begin to get together a set of criteria that you could use?

Irene Solaiman: Some of the trade-offs are about how much insight an evaluation will give you into a model. Some evaluations that are popularly run, for example around bias, are about specific protected classes and are then quantified. The Wino bias, for example, quantifies gender and is very binary. It gives you somewhat of an insight into a model but will not give you full clarity on the biases of a language model when deployed in a contextual setting. Ultimately, the best assessment happens contextually, and more tailored evaluations will need to be crafted for that specific risk.

Lord Hall of Birkenhead: Are the criteria for those testing systems best worked through by government or regulators, or by government and regulators? By whom?

Irene Solaiman: This is where external expertise is necessary. No one organisation, regardless of how large or well resourced it is, can have all the possible perspectives and expertise to robustly analyse a system. This is where we see that red teaming is incredibly popular, for example at Defcon in August, where community college students were trained. Many people can be trained to bring their life perspectives and social science backgrounds across different axes. I think government needs to increase capacity by risk area to be able to evaluate systems, which is where third-party auditing comes in too.

Lord Hall of Birkenhead: Thank you. If I could turn to our experts here, who are not in the warmth of the Bay Area, is there anything that you would like to add, Dr Draief?

Dr Moez Draief: I would add that the industry has not come up with standardised ways of evaluating these models. We still have a long way to go. Irene talked about biases and privacy, et cetera, but a lot of risks can happen in deployment, especially in specific industries that we do not yet understand and have to evaluate for. The industry has to come together and figure out what needs to be evaluated, so that regulators know what to evaluate and test.

Another thing that would be extremely helpful is if the people who are using the technology understand what the technology is or is not capable of. This was prior to LLMs: when we deployed AI in any settings, we did it with the people who would use it, so that they would understand what its limitations were and would be able to report them—and this is live testing. We had constant feedback on the limitations of this technology. Then, the developers could improve on it.

Going back to your question about open source, it contributes to that. Irene referred to red teaming or looking at the data that is being used to understand the different components and test them, not as a one-off exercise but as a continuous effort. This is extremely important, and the more people look into it, from users to developers and regulators, the more we can take advantage of this technology in a safe way, rather than accepting that tests are passed.

Some companies are training these models in such a way that they appear to be safe, but the safety guard rails get broken very quickly

once they are deployed and we do not know why. There is a lot of work to be done. I think it is doable if there are many more people involved in the pre-release testing, testing on deployment, auditing, and then having access to this technology to be able to know what its capabilities and limitations are.

Lord Hall of Birkenhead: That is so interesting, because you are saying that the process of testing is continuous—I think Irene made exactly the same point—and that it gets broader and broader, because the more people are using whatever is developed, the more they will have to have testing systems themselves for whatever they are doing, and we not know what those things are.

Dr Moez Draief: Absolutely. If there are some layers of openness to be able to troubleshoot—to look at what causes problems—the closer we get to the models, the easier it becomes, and the more everybody benefits. However, something to keep in mind is that if this becomes an extremely onerous process, it will have the opposite effect of enabling only the few people who can afford to pass these tests. There are trade-offs to be found in how much pre-release testing will cost, who conducts it and what the responsibility is downstream. We need an understanding of this complex value chain to know where to intervene, and how to guarantee competition and diversity of offers and choice to consumers when we allow these technologies to be deployed in the market.

Lord Hall of Birkenhead: Thank you, that is really interesting. Dr Koshiyama, is there anything you want to add?

Dr Adriano Koshiyama: I agree with both witnesses about conducting an impact assessment before you deploy a system, so that you understand the specific risk you should bear in mind. What is different from a customer service chatbot when you are trying to build an AI assistant? They are very different applications, so you need to bear that in mind.

The second point is safety testing. When we think about the LLM context of safety testing and look more broadly at where the regulators are pushing, here in the UK, in Europe and in the US, they all talk about the same four major buckets of risk: robustness in performance, privacy and security, transparency, and a kind of fairness—or bias, whatever you want to call it.

When we talk about the LLM ones, they are holistic.

When we talk about robustness and performance, we are thinking about how good the system is for common-sense reasoning. Does it have word knowledge, for example? Can it prevent or have some mechanism to avoid hallucination? That is in the bucket of robustness.

In the bucket of bias, we should think about toxicity testing and the form of stereotyping that is emerging from the language model. When you are talking about privacy and security, you should definitely think about PII leakage and prompt injection.

On transparency, there are all the elements that were mentioned before about the model cards' documentation data.

As you can see, there are a lot of methodologies, which were developed by the open-source community and academic community tests.

Lord Hall of Birkenhead: You used the word "toxicity". What did you mean by that?

Dr Adriano Koshiyama: Toxicity is an inclination of language models sometimes to throw up some form of foul language or profanity. One of the things you can do is red teaming. Even if you do not want to red team, you have to think about whether you want to stimulate the language model through prompts so that it can attack you back or maybe throw back a profanity towards you. If you want to make a customer service chatbot, ideally you would like to avoid that, so people will test for that before they deploy it. That is one of the tests that you would want to do for that kind of context, but every context would have a different type of safety test.

Professor John McDermid: I have a few things to add. Pre-deployment testing is necessary but far from sufficient. I will not repeat the discussion we just had about openness and evidence, but we need other things as well. The space of behaviours of these models is so complex that you cannot hope to get confidence just by testing them.

Dr Koshiyama mentioned some common areas of concern. I have personally worked for many years in areas where we are concerned with physical harm, either directly through physical systems or indirectly through misinformation. You need to be able to analyse and test in that context. For example, if I build a model for recommending the treatment of sepsis in adult patients, that advice will not necessarily be good if I apply it on a paediatric ward or for children. We need to really understand the context in which these are used and to analyse the requirements in that context of use. A lot of this comes back to the regulators, which have the knowledge of the domains to be able to ask for the right information to make judgments about deployment.

I reinforce what was said before about continuing to monitor these systems in operation. They are very complex, and we have what we have call emergent behaviour, which is basically things happening that we did not predict before we used them. It happens with all sorts of systems and will happen more with these sorts of models, particularly with LLMs. We need to be able to observe those in operation to say, "Does it matter?" If it does not, that is fine, but if it does, what do we need to change in order to get rid of that behaviour in future? The recent removal of licences from Cruise in California is an exact example of that phenomenon. We need regulatory practices in place that make it possible to do that.

Irene Solaiman: Clearly, we need more specificity about what needs to be tested. "Toxicity" is a very fuzzy term, and it has a lot of red flags in the way we measure it.

Q69 **The Chair:** I do not know whether you are the best person to direct this question to, but how can we have pre-deployment testing if models are being made open source and then being fine-tuned by private actors? I am keen to understand that.

Linked to that, yesterday we took evidence from a copyright lawyer and legal expert who gave an example of how, once models are released, if they include false data and someone wants to challenge the consequences of that false data and asks for a modification to the contents of that machine, it cannot be modified. You have to wait until the next model is released and, even then, the previous model does not evaporate. It is not like software, which gets updated and overrides the previous version. I wonder whether you can comment on that.

Irene Solaiman: I will go to the point about what is available and then come back to how open source is released. What is available depends on what we are updating. If we are updating the training data itself, that is likely to be incredibly costly, depending on the size of the model, and will take time to update.

There is a distinction between access and openness. Having model weights that are more accessible and more available means that we can take down some model weights from the Hugging Face platform, for example. However, to be clear, there is a level of availability that can change.

For pre-deployment risk assessments for open source, an incredible community norm has developed over the past few years in disclosure when models are released. This is not regulatory. Almost all popular models, and all models on our platform, are released with a model card. Model cards do not have explicit requirements for what is fleshed out but will generally come with a set of evaluations. When models are released openly, they will have the evaluation that was done early in the release process documented in a very consumable manner for many types of technical and non-technical audiences. Then there is a process of opening to researchers. Again, that provides more access to get that perspective by different safety categories.

The Chair: I have one final question for Professor McDermid before we move on. What is your view, as an outsider rather than someone who is part of the tech world, as to whether pre-release testing entrenches incumbent advantage or whether it encourages greater competition to build safer models? Do you think there is a likely outcome from it, and which of those would it be?

Professor John McDermid: Independent testing should encourage better outcomes. It is in people's interests to do well.

The Chair: With more competition?

Professor John McDermid: Yes. An obvious example would be NCAP testing for cars, where people strive to get better ratings. If independent testing were done and results on an intelligible scale were published, I think that would encourage constructive competition.

Q70 **Lord Foster of Bath:** We have just been having lengthy discussions about safety testing both pre and post deployment, and we are already discussing what buckets of risk we should have and what issues of further specificity there should be. Here in the UK we are looking at how we can develop our own safety requirements and our own liability rules for the development of AI models, but we in the UK have already done that for other complex systems, such as software development in the airline industry, which I know you know a bit about.

My question is a simple one: can we learn from other sectors or, as some have argued, are the differences just too great, not least in complexity? There is a subtext to all this: is it even worth us bothering to do it, or should it be done only as a result of international agreement?

Professor John McDermid: My very simple answer is that we can learn from what has been done. It will be a challenge to translate it into the world of AI and LLMs, but it is worth doing. I shall try to pick up a number of more specific points that I think would move across. In many domains we have the idea of a safety case. This is an argument, in the sense of a rationale, and evidence that the system is acceptably safe to operate in some given context. It is a widespread idea whose origins go back to serious accidents such as Piper Alpha. There is work on applying this to AI, and I can give the committee references and so on as appropriate.

One reason why I really do not like the idea of just dealing with testing prior to deployment is because most of the systems I have seen that have bad safety properties were bad from the concept stage. What they envisaged building they built, but it was not a great idea in the first place.

In a number of sectors, particularly aerospace, they now have a standard in place that says that when you propose an initial concept, you evaluate that to see whether it has desirable safety properties, but also whether it is possible to realise that system with available technology and whether we know how to assess it and ensure that it is safe. It is taking a very early stage to look forward and say that this is something we should be carrying on with, and/or making modifications to, to make sure that as we develop it we end up in a good place. That is difficult, but it can be done.

Another crucial thing is system architectures. The simple thing to draw out of this is that we design systems so that no single point of failure can lead to dangerous behaviour. That means that we use redundancy, so that we have something that does the function several times. We use diversity: things that do the same function in different ways. We compare them, and if they agree, fine. If they disagree, we take some other action. This is deeply ingrained across all the safety-critical industries I know. Can we do that in AI? Yes, I think we can, but it is a very ill-explored area and it will be interesting to see the comments my colleagues have on that.

Another thing that is really important, which is partly cultural and partly technical, is learning from experience. Something goes wrong and we

analyse that. In aerospace, they have moved towards a just, no-blame culture whereby, rather than blaming people for what went wrong, they seek to learn what the underlying causes were, so that those can be rectified. For example, we now have different rules about how communication is managed in cockpits because of accidents that occurred that would not have arisen otherwise. They also encourage open reporting to the databases, whereby problems can be reported with particular aircraft or flights. Maybe we need that for some of these large language models. These all can be applied, and there is some research that shows how you can do some of this but by no means all of it. We need to set some frameworks in place and do further research.

We also talked about legal positions and liability. I am not a lawyer, but I have worked as an expert witness a few times. My straight view about liability is that it lies with the manufacturer. These things are far too complex to transfer liability to the user. That gets more complicated, though, if the user retrains the models. Maybe the Law Commission should be prompted to look at such issues. From my experience of working with them, they really get to the crux of such issues, and that might be a good thing to come out of it.

From my experience as an expert witness, it is often very hard to determine cause and effect. That will be one of the major challenges. In aerospace, there was an incident in 2014 when the UK air traffic control system failed. It shut down and could not manage flights. I was lucky enough to be asked to look at this. Actually, the designers found the bug in the code in 45 minutes, because it was designed to be analysable, so they could do fault detection. If they had just turned it back on again, it would have failed in exactly the same way. Because they had found that, they could bring it up in a different way so that it did not fail.

For AI and LLMs, this is a huge research issue, but we need to get to the position where we can do that sort of diagnosis so that we can make them resilient and can learn how to avoid those problems in future. That will also help in unravelling this tangled web of cause and effect and where liability should be properly placed. That is difficult. I do not know how to do that—I do not think anybody does—but it must be on our research agenda as a community if we are going to use these things in really complex environments.

That was a long answer, but it is quite an important point.

Lord Foster of Bath: That was a very helpful answer. You have not said anything about whether it is worth the UK doing this or whether we should just leave it to international agreement. I would value your comments on that.

I will pick up on one other point you made, which is the issue of knowing what we need to test, but, at the same time, ensuring that we have the means to carry out that testing. There is a bit of a chicken and egg issue here. Otherwise, we will end up with a simplistic set of safety regulations, because we will only have very simple tests that we are capable of carrying out.

Professor John McDermid: On that latter point, again that requires a level of design visibility. In traditional safety engineering, you start with the top level and flow down to the components of the system as appropriate. We have some experience of doing that with systems that use AI, but it is still quite difficult. If you do not try to do that, you do not know what to test.

Typically, ML models are trained on average performance. If, for example, I am working on an autonomous vehicle, average performance for detecting obstacles is not very interesting. I want to know much more about detecting things close to me than things that are far away. You want to shape the requirements to say, "Actually, I know the performance of the model that I want". If you have done that, you can test against that. I am not saying it is easy. It must be done in a context-dependent and context-specific way. The requirements for a train with obstacles on a track versus other objects on a road are very different. I think that can be done.

In my view, it is worth the UK doing it rather than leaving it to be dealt with internationally, partly because of our safety culture and the way we regulate things. The balance that we get between innovation and regulation is in a better place than other countries. If we do that, we can lead the way. Some of the questions that I have mentioned are quite hard, but I think we will ask some of the hard questions that need answering, not just for us but more widely. If we leave it to international consensus, I am not convinced that we will get good answers.

One other aspect of that—part of it relates to your fourth question—is that there are international bodies for dealing with different domains, such as civil aviation and maritime. They move very slowly. If we wait for those bodies to put international regulations and standards in place, we will be far behind the curve. We need to be able to move with a level of agility that would not be possible through those international bodies. I would very much encourage the UK to take the lead in this area.

Lord Foster of Bath: I would love to raise a further question about legal liability for AI models that are not developed in the UK, but we will leave that for another day. Do any of the other three have anything they want to add to what has been said, or any different perspective?

Dr Moez Draief: I would love to add something about whether it is worth this happening in the UK. Sometimes we think that building and testing models are two separate things. Actually, they require the same skill set, so if the UK is not involved in building or testing models, it will not have the capability to take advantage of these technologies when they are customised and fine-tuned for certain applications or be able to test the safety of those applications.

It is important for the UK to work with the international community to invest in the ability to test models that require an understanding of how neural networks work, the relationship between the data and the models, and what happens if we do this guard rail or that guard rail. There is an existing community in the UK—the academic community—that is looking into this. We also need practitioners who can do this to then do

secondments with regulators and to support the Government to create that thriving eco-system of people who know these technologies and who can contribute to them.

Dr Adriano Koshiyama: Setting the minimum standards for safety is probably the million-dollar question of the decade, not only for language models but for all the other AI systems out there. It is really where most of the resources and money are going for the next few years. We know that is happening in the EU, the US, and here in the UK, because if we talk about standards, we need to think about which metrics we are going to assess and what the bounds of acceptability are.

It tends to be quite easy to decide the metrics. The problem is that the bounds of acceptability tend to be the ones that require quite a lot of agreements and disagreements with people as it goes along, so it is extremely difficult.

Just to use the audit term, on liability I think about the UK positioning on the global stage. There is a huge market opportunity for insurance for AI systems. The UK has a huge insurance industry. We are seeing Microsoft, Google and now OpenAI saying that they will take the liability in case there is corporate infringement. Thinking about this aspect, providing a form of insurance would be quite important. I am not saying that that is the whole solution—buy insurance and live your life—but it would be a form of dealing with residual risk and a market-based solution for that.

Lord Foster of Bath: I will look at that as a future career. Thank you for the suggestion.

The Chair: Thank you. Just before we move on to auditing and assurance, coming back to liability, Professor McDermid, you acknowledged that this is a complex area. To be simple for a moment, one of the simple debates in the business world in looking at AI is whether liability lies with developers or deployers, and the liability question is perhaps a deterrent to some businesses wanting to deploy AI if they are not clear where liability for this technology sits. If possible, can you say very simply whether liability in the aerospace airline industry is with the manufacturers or with the airlines?

Professor John McDermid: It will depend a little on whether you are looking at technical failures that occurred, such as with the Boeing 737 MAX. Those are clearly demonstrable as design limitations or flaws, so liability clearly lies with the manufacturer. However, often the airlines will do maintenance on the aircraft, so they need to have carried out maintenance operations correctly. Another problem is using counterfeit parts, which are cheaper than the real ones. So it will depend on the things that are caused.

When we translate this across into the AI and LLM world, you are saying that concerns about liability might be deterring people. That could be true. As I say, it is quite difficult to pull apart. My understanding is that in the law they look at causality in a slightly different way than I would as an engineer. However, they are concerned with causality, and if these things are very complex and you cannot tell, for example, the extent to

which some local prioritisation or customisation that has been done by a deployer is causally responsible for something as opposed to how the model was built originally by the manufacturer, that will lead to a lot of ambiguity and argument—and will keep some lawyers and expert witnesses quite rich, I suspect.

Work on this is beginning to be done, as well as the EU regulations. I have a colleague who works in the University who recently published a book on how he believes tort laws are affected by the introduction of AI, and there are other things. So some work done in that area is beginning to be done, which might give greater clarity. However, it will take a long time to resolve, and, like a lot of legal things, it will come down to case law.

Q71 **Baroness Harding of Winscombe:** I just wondered whether there was anything we could learn from the regulation of other sectors on the right time to start to impose some audit requirements and safety regulation on an industry. I am struck by the differences between the aviation industry, which I know you know a lot about, and healthcare, which existed for hundreds of years before we had regulation. Can we learn anything from those different approaches to the right time for a government-imposed regulatory regime?

Professor John McDermid: It is a very interesting question. There are also different forms of regulation, as well as different timescales. All I can say here is that it depends on the application area. If we are going to use these things in flying aircraft or driving cars, as is happening, we need to have something rather like what the industry does now. One reason why I say that is because we still have to do all the things we do with aircraft and cars, and so on, whether or not they have AI in them, and you have to be able to integrate the processes that are dealing with these critical, functional components into those existing regulatory processes.

You have to do them the same way, with novel applications. I would want to try to do that not just in relation to risk but in relation to our ability to detect the problems and remediate them. Have something that is easily detectable, and I can clearly and quickly fix it by paying compensation or whatever. Then, perhaps, we can afford to be more flexible and look at those things retrospectively and say, "Actually, we didn't really like that. Let's just change the way we work". If you kill passengers on an aircraft, you might have to pay compensation to the family, but there is no real remediation of the problem. You have to take a much stricter approach and have more pre-deployment approval.

Again, I refer you back to the Law Commission report, which I am delighted will be taken to the House for consideration following the King's Speech. The Law Commission has looked carefully at those balances, and that is quite a good model for looking at how you might do that for something that could harm people physically, up to and including loss of life. You have to have different balances. As I say, you can afford to let some things be deployed, monitor them and then judge, but with the

more critical things you need to be clear about them before they are deployed.

We also need to work out, exactly as my colleague here was saying, what risks are acceptable in those different domains. That is the starting point for me with many of these discussions.

Q72 **Lord Kamall:** I want to pick up some of those points. I was very interested in what Dr Koshiyama said about a market opportunity, but I was not quite sure whether it was for insurance or assurance as a product.

Dr Adriano Koshiyama: It was insurance.

Lord Kamall: Very good. We will take that away and try to make some money out of it.

I want to take a step back and ask about one of these issues. Do you really think that we need some form of auditing system to provide assurance, and what do we need to check? Is it training data, methods, and so on? There is also compliance. We had some copyright experts here yesterday. Obviously, there are data protection issues, and safety issues, which we have spoken about.

I have some questions for you to consider. You do not have to answer them all in a systematic way but perhaps you can help answer them. Can this be done meaningfully with open-source models? Are there already good cross-regulator standards for this sort of auditing, or would we need to look at third parties? What about models based outside the UK, in the US?

Another example came up yesterday, and I am interested in how you test for this. More than one witness said that sometimes large language models make things up, incorrectly or inaccurately deduce something, or hallucinate. How would you deal with that in this sort of auditing system? Could you do so, or do you have to wait for deployment?

Dr Adriano Koshiyama: Just to say as a disclaimer, I am a big believer in auditing. My initial engagement in general was that when I was an academic at UCL we were commissioned by the CDEI to conduct AI audits on its behalf in industry—this was in 2019-20, so a few years ago. The main idea at the time was to try to learn how to conduct those audits in a real-world setting.

We got to write a report for the CDEI, which is published on the GOV.UK website, on the need for AI assurance. The word "assurance" came about, because sometimes when we engaged with companies to try to audit their systems, they would get scared—nobody wants to be audited—so we would say, "Actually, we are here to provide assurance to your systems", which was somehow more like a benefit.

Usually, there is a systematic process. The nice thing about the term "audit" is that we are really drawing parallels with financial and IT performance audits, in the sense of coming in to conduct a risk assessment. With that risk assessment, you would estimate something

that we call inherent risk, which is basically the risk of conducting the activity without any controls or meaningful mitigation. Then, depending on the risk, you typically look at the red and the green—low, medium or high—and decide what the next form of action is.

In the context of AI, the risks are a bit more complicated. We are talking about robustness, privacy, security and bias. There are all those risks. Then you try to conduct a form of verification on the system. If the risk is too high after verification, you try to help with mitigation by putting some kind of controls in place—ways for them to take the risk down to some residual level that is enough for the business to take that risk or to feel confident that they can deploy it. That was the initial pre-deployment form of audit, and it was quite interesting, because we could repeat the process so many times, regardless of which AI system we were working on. Whether it was housing associations, police services or HR, we were using the same processes over and over.

Lord Kamall: Was that also regardless of whether they were open or closed systems?

Dr Adriano Koshiyama: Absolutely. It was the same. The open-source models were actually much easier in general, because the problem with closed-source ones is that you have to deal with the IT from the company and the developers. It is a much trickier kind of relationship. With the open source, there were all the assets for you to play with, and as long as you knew what they were going to use the open source for—it is very important that you know the user case they will go for—you could go about providing that testing.

There was a question about what kind of parallels we could draw with other parts of the world. The EU may have a potential commitment to third-party conformity assessment—it uses that term—for high-risk applications. In the US, they have a few applications where they do third-party conformity assessment or AI audits. The most famous one is a law in New York on conducting bias audits for HR technology solutions, so that every employer in New York City, before they used some technology to recruit people, needed to conduct a bias audit. Can you imagine how many companies and technologies got affected by that? The next thing about the US in that case is that it has good standards for which metrics to use, what is and is not acceptable, et cetera, which really helps. If that law were to happen in the UK, we would be discussing what standard of acceptability we should go for. That would be the major issue.

On the LLM side, to finalise on the question of hallucination and that form of risk, at this point in time you can try to test for some forms of hallucination. There are myriad forms of hallucination. One example is when you dump a PDF of text to a language model and ask it to summarise it for you, maybe in one or two paragraphs. It could be that the language model does not summarise but is just predicting the next word and makes things up, or maybe it works as a summary, so you can evaluate whether it is hallucinating the terms of summarisation.

Some tasks are doable, and others are much more complicated because they require access to a knowledge base. The most famous one was also in the US. Some lawyers decided to use ChatGPT to write a court case, and the media used the term "hallucinate". The model hallucinated; basically, it came up with a precedent that did not exist. Someone asked me, "How would you be able to pick that up?". The only way to do so would be if you had access to a knowledge base with all the cases, so you could identify whether it is real or just making things up. In that case, the judge was the one who picked it up and those lawyers ended up getting fired by the court. That is an interesting user case, so it is doable for some cases but not for all forms.

Lord Kamall: It is interesting that AI keeps a judge on his or her toes to make sure that they are up to date. What sort of audits are we talking about here? We hear about governance audits, empirical audits, technical audits, compliance audits et cetera. Is it all of the above plus a few more, to your mind?

Dr Adriano Koshiyama: It depends on the regime in place where you are and the application. Ideally, you would try to have a more systematic process for all of them. If you are dealing with a large enterprise, for example, ideally you would start with a governance audit. It is extremely important to know what their processes are for preventing, detecting or correcting any risks emerging from AI, what teams they have in place and what their accountability mechanism is. When I am talking more about the specific lens of AI, I am thinking about technical audits. Can I go there and investigate the data or the predictions coming out of the model, and do some evaluation of that? It is a top-down and bottom-up approach; we need to work on both at the same time.

Finally, the compliance audits are much more tailored, because you are really testing for a specific form of compliance that you want. Technical or governance audits are much broader in application.

Lord Kamall: Thank you. That was very comprehensive. Do any witnesses want to add anything to that comprehensive overview or disagree with any of it?

Professor John McDermid: I will make just a few points. I agree with what has been said. As well as looking at the organisation, you need to look at the culture. There is a notion of psychological safety: are people prepared to speak up when they feel that there are problems? That is very important too. The Boeing 737 MAX, again, is an example of where there was not that psychological safety.

On hallucinations, as a safety guy they really worry me. If I present something to somebody that is plausible but wrong, they are much more likely to operate on it—and if it is wrong, it could be dangerous. If I am analysing a system, that is one thing that really bothers me. We need to look at that.

A number of industries—aerospace, for example—have some practices which they call stage of involvement reviews. They do those at various points in the development process, basically to check how things are

going and to take or recommend early corrective actions. That is harder to translate across to the AI world. LLMs tend to have a much more iterative development process. We need to learn from that by having much more continual engagement. Again, that is one of the things these models will do. They will change us from saying, "We'll check it at this point", to saying that we have to have much more continual engagement to help to guide things to a good outcome. That is probably consistent with the way some of the audits we have talked about were conducted, but it would be a shift in mindset on how we do auditing.

Lord Kamall: That is very helpful. There is also the fact that as humans we should admit the conceit of knowledge, as Hayek said. We have a limited knowledge and should almost admit that some of these LLMs will have a limited knowledge, based on the data that they are being fed.

The Chair: Thank you. We have covered a lot of ground, and I am very grateful to all four of you for giving up your time this afternoon to join us—and to Ms Solaiman for getting up very early, I would guess, to join us from over on the west coast of the US. Thank you.