HOUSE OF LORDS

Communications and Digital Committee

Corrected oral evidence: Large language models

Tuesday 12 September 2023

2 pm

Watch the meeting

Members present: Baroness Stowell of Beeston (The Chair); Baroness Featherstone; 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. 1 Heard in Public Questions 1 - 11

Witnesses

I: Ian Hogarth, Chair, Frontier AI Taskforce; Dr Jean Innes, Chief Executive Officer, Alan Turing Institute; Professor Neil Lawrence, DeepMind Professor of Machine Learning, University of Cambridge; Ben Brooks, Head of Public Policy, Stability AI.

USE OF THE TRANSCRIPT

This is a corrected transcript of evidence taken in public and webcast on www.parliamentlive.tv.

Examination of witnesses

Ian Hogarth, Dr Jean Innes, Professor Neil Lawrence and Ben Brooks.

Q1 **The Chair:** This is the Communications and Digital Committee, and we are very pleased to be starting a new inquiry into large language models. This is our first public hearing, and I will come to our witnesses in a moment.

Large language models, as we know, are a form of generative AI, also sometimes known as foundation models, and the experts before us, I am sure, will be able to say more about them in a moment. AI is, of course, a massive topic. This inquiry is seeking to be specific and targeted, and we are focusing on large language models, not least because they very much entered the public consciousness some months ago with the arrival of ChatGPT, although our experts, I am sure, will testify that that is not where this journey started.

We will concentrate over the next few months on how large language models are likely to develop in the next three years. Rather than necessarily focusing on the future, we want to look at the here and now, and at the risks and opportunities in that timeframe for the UK. We will be assessing the Government's White Paper and their proposed approach to regulation. Is it right, and can our regulators meet what is expected of them? If we get this inquiry and our resulting report right, I hope we will be able to offer a case study for other parts of AI in the way in which the Government wish to approach this in the future.

Can I start by inviting you, our witnesses, to state your organisation, if you are representing one? Then we will move to the first question.

Ben Brooks: I am the head of public policy for Stability AI. We are a global AI company based here in the UK.

Dr Jean Innes: I am the incoming CEO of the Alan Turing Institute in the UK.

Ian Hogarth: I am the chair of the UK's Frontier AI Taskforce.

Professor Neil Lawrence: I am a professor of machine learning at the University of Cambridge, but I am not representing that organisation. I am speaking independently.

Q2 **The Chair:** With our first session of an inquiry, we try to stick to strategic, big picture questions. We are hoping to get different perspectives from the four of you on some of these big questions in the context of large language models, and to get your differing views on the Government's response to these big questions. We will also be talking specifically about the task force and the other institutions that are expected to play a role in the context of what the Government have proposed.

Professor Lawrence, would you like to describe for us what LLMs are, how they differ from other types of AI and how they have developed over the recent past?

Professor Neil Lawrence: I should start by saying that large language models are a form of machine learning. I am a professor of machine learning and am quite uncomfortable with the term "artificial intelligence", because it harks back to an era when many people believed we could solve all problems in the world from first principles, through things like logic. You see this in philosophical traditions as well, so traditions of rationalism.

Machine learning is really a tradition of empiricism: "Let's see how things play out in practice". This has led to the confusion, when people deploy machine learning, that because both are occurring on a digital computer that fundamentally works through logic, a digital computer that is using machine learning will always be correct, but it will not; it will make mistakes, which are derived from whatever source of data the machine learning is using. In the case of large language models, the source of that data is human data.

This is very interesting, because for many years the so-called classical AI community was interested in how we can improve communication between humans and computers. They tried to do that through logical techniques, language parsing and looking at grammar structures—work going back to Wittgenstein.

In practice, this proved impossible to do, because there is no form of universal intelligence that can ever dominate. Our form of human intelligence is just a form of intelligence that we use. What machine-learning techniques have been able to show over the last decade—and the interest in so-called artificial intelligence is actually an interest in machine learning—is that progress is best made on these activities so far by using data, often from humans, to reconstruct what human thinking is.

When we have large language models, the form of that data is at an unimaginable scale. It is the whole of the internet. It is everything that has ever been written, all being compressed into a computer. It can do all that, because it has a bandwidth that is many millions of times greater than the typical bandwidth of a human being. Because it has such a large bandwidth, it is reconstructing human processes of thought through attempting to reconstruct our language.

This has led to a very large confusion, understandably, among the wider population that the thing they are listening to, talking to and finally communicating with—that dream from "Star Trek"—is combining this notion of rationality and logic simultaneously with something that a human can understand. In practice, all we have created is something that can reconstruct human language and, therefore, human conversation.

I prefer the term "human analogue machine". What do I mean by that? There was a period, before digital computers became popular, of analogue machines. One of my favourites is the MONIAC, which was an analogue to the UK economy. You can see it in the Science Museum. It has water flowing around the machine. The idea was that you could change tax rates and see tanks fill up, because money flows like water

does. What we have built is a human analogue machine. We have built a simulation of the human mind that has analogous quantities operating inside it that are analogous to the way we think, and we built that through human data.

The Chair: Rather than ask you all to give me your own essay answer to the question, "What is an LLM?", do any of you disagree with what Professor Lawrence has just said and want to offer an alternative, or do you want to add to what he has said before we move on?

Ben Brooks: I might just add that, from a functional perspective, these models are, essentially, analysing vast datasets. We are talking about hundreds of billions or trillions of words. To Professor Lawrence's point, they are understanding the relationship between words, ideas and fundamental textural structures that make up our system of communication. They are not storing the data that they have reviewed. Instead, they build up a body of knowledge, essentially how to read and write, which they can apply to a broad range of new tasks.

It is also worth pointing out at this juncture that models are just one component in the AI tools that we interact with online, such as chatbots. They are a very important part of those tools, but, to Professor Lawrence's point, they have certain limitations that we have known about for some time, and it is very important that we understand those before we integrate them in progressively more sensitive contexts.

Dr Jean Innes: What you said at the outset, Chair, is particularly important. These models have been current in the research community for some time, but your point about them entering the public domain and the public consciousness is very material.

The technical additional point that I would like to make is that these models are out in the wild, so it is just worth recognising that people with the right capabilities and technical understanding can use these machines or these capabilities for a variety of purposes, and we are seeing that proliferate now. That really is why the Alan Turing Institute so welcomes the inquiry, because we need to move with pace. We cannot put the toothpaste back in the tube. I am not unoptimistic about our ability to set out a positive way forward, but we have to be thoughtful about it and we have to move with pace.

The Chair: Mr Hogarth, you are the only one who has not spoken in answer to question one. That is not to suggest that I want everybody to answer every question that is asked, because otherwise we will be here for ever, but is there anything on this introductory question that you would like to add?

Ian Hogarth: There are two things that I would add. There is a famous post by Richard Sutton, the father of a particular area of machine learning, called "The Bitter Lesson", where he says that the only thing that matters in the long run is the leveraging of computation. One of the reasons why we are talking about language models is that we have a general-purpose architecture—the transformer architecture. We have a huge amount of textual data, as Professor Lawrence said, basically most

of the internet, so it is a very effective way to leverage computation. You can see this as a broader trend towards figuring out how to make the most of all the computing resource that machine learning researchers have available to them.

It can be a bit of a trap to think of language models as just being useful at language. These very highly trained models are able to do things beyond language. A couple of examples are working with tools. There was a famous research paper called *Toolformer*, which looked at taking a language model and, effectively, letting it drive tools using the same underlying training. The other is the application of language models to modelling biology, using the ways in which proteins are analogous to language.

Q3 **Lord Hall of Birkenhead:** This is a question for all of you to give us your views on. It would be helpful if you could begin to tell us what feels like hype and what feels like reality looking forward over the next three years. What risks and opportunities can you see? How do you think large language models will evolve over the next three years? As you four know what you are talking about, we are trying to get to what seems like a truthful position of what we can expect, even though I completely accept that you can never predict the future.

Professor Neil Lawrence: This is a difficult question to answer, because we are in a transformational time. For historical equivalents, you have to look back at things like the invention of writing. I mention the invention of writing, because for the first time we could start storing our ideas in ways that we could share and that could be held over time, with a transformational effect on our society and our cities.

One of the really interesting components of the invention of writing is how we were very lucky in Europe to receive a system of writing via the Phoenicians that was already reduced in complexity. That meant that, when it came to inventing the printing press and literacy, we could very quickly share that capability, and we have literacy today.

One challenge, and perhaps the most difficult, that we have in this space is that computers are being controlled by the modern equivalent of scribes. The software engineering profession exists as the modern equivalent of guilds and has an incredible amount of power over Governments. A lot of the things that we are looking at are about how to deal with those power asymmetries. During my time at the AI Council, when it existed, this was the type of question that we were concerned about.

To your specific question about hype versus reality, these are, first of all, very complex problems. I have learned to be very wary of simple answers and simple solutions, because we know that they tend not to work. Unfortunately, when problems are very complex—so-called wicked problems—it is government that has the responsibility. If the problem was easy to solve and you could make money out of it, market economics would deal with it.

We have a real, critical moment of steerage here, and it is difficult to know what the right decisions are, but I see the two principal components of that steerage as an ongoing argument about whether these models are open or closed. That argument is about asking, "Who is building and controlling these models?" If we look at the previous wave of disruption that occurred around the internet, one of the most important aspects was the system of open-source software that enabled companies such as Google to compete with Microsoft. You can see that there is a lot of interest among the big tech companies in maintaining closed ecosystems, because they do not want to be disrupted. They are currently dominating and there is a very serious danger of regulatory capture, but there are serious and important points around security and the availability of these models for bad actors.

My own feeling is that it is a little like the security debate. There was a large debate under the Labour Government at the dawn of the internet about whether we should be making security capabilities available. Of course, we have the Online Safety Bill being read again, which looks at the consequences of those capabilities in society right now.

At the time, the debate was about security and defence, but the counterargument was, "You're not going to be able to do commerce on the internet if people can't be secure about their communications". There is a wicked problem that they are having to debate in Parliament right now.

The other aspect is the business model. One of the most pernicious effects of the previous wave of machine learning, which was much simpler, is that the simplicity tended to drive emergent societal effects that you could not necessarily predict when you built the individual models, which are often quite explainable and understandable—things like echo chambers and groups being created online that make it easy to share misinformation, because these models were simple and easy to form.

Part of that is coming about because the alignment of the business interests—in particular, those of commercial companies—to the business model is to sell advertising. It is not to think, "What do the people interfacing with these systems really want?" It is to ask, "How do I get them to engage with my advertising platform?" It was perfectly innocent and existed for a long time, but, when done at that scale, it creates these incentives that are not in the interests of the wider population.

Subscription business models do not have that kind of effect. You can imagine that, on the other spectrum, if people were subscribing to these models, there would be more of an interest in the companies to ensure that people were getting the services they wanted, not just in the short term but in the longer term. It will be interesting to see how that pans out.

I do not have clear answers on either of these, but I am just trying to give you the spectrum of quadrants. You can imagine subscription but open, subscription but closed, et cetera.

Lord Hall of Birkenhead: That is really clear. What you said earlier about the invention of writing was a very vivid thought. You mentioned scribes. Who are the scribes?

Professor Neil Lawrence: The scribes are the software engineers. I have brought a 3D printout of a cuneiform tablet, which, when you look at it, is a little Excel spreadsheet. We have exactly the same thing going on. This is a document record from the city of Nippur of the yields of corn from a number of owners, which is being held, presumably, for tax purposes or whatever. The computer version of that is Excel, and that is what we have developed and all been working with.

The scribe is doing that and being an accountant as well, so writing was invented by accountants but used later by poets. This is the transition moment from accountants to poets, and there are possibilities around that moment of significantly affecting the ability of many people across the country to access these technologies, including the left-behind companies that we understand have such a challenge in the productivity dichotomy that we have.

I spent a lot of my life in Sheffield. I always think of a car parts manufacturer in Rotherham. How is it accessing these capabilities? How is the NHS accessing these capabilities? There is a possibility for democratisation of these. That is the utopia, but the dystopia is that those people somehow do not get access and we have a wave of bad actors who have interests, such as the Internet Research Agency, which presented fake information for the US election, and are getting access to and deploying these technologies. This is where the steerage is so important, because these are all imaginable worlds, with a spectrum all the way between.

Lord Hall of Birkenhead: I want to come back later to another point with you, if I might. Mr Hogarth, do you want to comment on the question of how you see things moving in the next three years?

Ian Hogarth: Your original question was whether it is overhyped or underhyped.

Lord Hall of Birkenhead: Yes. Therefore, can you help us to steer to where you think things might be over the next two to three years? The point about steerage that Professor Lawrence has just been making is really interesting.

Ian Hogarth: Again, it goes back to the point about more computing resources, and computing resources driving so much of the outcome here. We have increased the amount of computing resource that a given model has access to by a factor of 100 million in the last decade. There are not very many things that we have increased by a factor of 100 million in a decade, so the question is whether that will continue. I believe that it may not continue quite as quickly, but that we will stay on an exponential where you see a 10 orders of magnitude increase in compute in a given period.

The reason for that is that there is a huge amount of money to be made. These tools will be used for lots of commercial purposes, so the amount

of money being invested in making these systems more powerful will increase, as it has already. A decade ago, $20 million was invested in companies trying to build super-intelligent AI, and now it is $20 billion. There is a race occurring between companies and countries to build these very powerful systems.

You should assume that there may be as much as a six orders of magnitude increase in the amount of compute that you see given to the next generation of models within the next decade, and we should assume that that will yield some pretty breathtaking capabilities, if the last decade is anything to go by.

If I had to choose between underhyped or overhyped, I would pick underhyped, because we are always taken by surprise by exponentials in general, since we do not intuitively think in exponentials.

Lord Hall of Birkenhead: That is interesting. With growth continuing at that exponential rate, will we see fewer and fewer nations or companies with the resources to be able to grow at that speed?

Ian Hogarth: That is a really hard question to answer. The question of open versus closed is very nuanced, and maybe one that we should return to. The other side of the compute story is that compute is getting cheaper, so more people are able to build these models. If we use history as a guide, when OpenAI trained GPT-3 there was one model at the scale of GPT-3 in the world, and now there must be 100-plus.

When OpenAI trained GPT-4, there was one GPT-4 scale model, and there will probably be 10 by the middle of next year. We are seeing quite a significant proliferation of these capabilities already, and the question is how safe that proliferation is.

Dr Jean Innes: I will talk briefly about opportunities and risks, and then make some comments about how we steer our way through. I am fundamentally an optimist, but there is a real challenge here.

On the opportunities, I would highlight education, for example. We do not have enough qualified maths teachers in this country, never mind in emerging economies. There is a wonderful possible outcome of this, which is really well-personalised learning, where the software matches its pace to the learning of the student. Even from the patterns that it is observing, it understands why a particular student is struggling with a particular concept. That is really exciting.

A second opportunity that I would point to is in productivity. There is great literature on the pace at which technological change delivers productivity increases into the economy. I do not plan to summarise that today, but I would just observe very specifically that large language models can write code. One of the pinch points in the labour market at the moment is people who can write code. Those are two very specific opportunities, which are not without risk—they need to be managed properly.

I would also point to something a bit more left-field. I have heard discussion of whether large language models could be applied to weather

forecasting—to Neil's point that these are not necessarily specific to language. There is a foundation capability.

On the risks, in the short term, the ones that loom large to me are mass disinformation, which we are already seeing. I would look for short-term labour market impacts, because they are likely to come quickly. The last would be the proliferation of large-scale sophisticated cyberattacks. That was just a top-level discussion of opportunities and risks.

In terms of how we steer it, which is really the seminal point, I would like to offer two thoughts. One is the pace point. My background is in bioscience. I think about the Human Fertilisation and Embryology Act 1990. The first in vitro fertilisation happened in 1978, so we had the luxury of 12 years and the Warnock report to think about a really thoughtful regime for how to manage both the potential for wonderful genetic cures and the quite concerning societal risks that could come from cloning. We do not have 12 years, which, again, is why I am so pleased that this inquiry is taking place. We need to come together with pace to think really thoughtfully about the way ahead.

Ben Brooks: From our perspective, we expect these language models to become more capable over time. They will be able to comprehend more complex information. They will be able to reason in ways that are progressively more accurate.

But this focus on an arms race for larger and larger models and more and more compute overlooks some other trends that, frankly, will be just as, or perhaps more, consequential for AI adoption across the economy. In particular, we are now seeing this wave of grass-roots innovation among independent developers, researchers, creators and small firms.

Previously, to put this into context, something like GPT-4, which is one of the state-of-the-art closed-source models, cost something like $100 million to train and develop. It is an insurmountable barrier to entry. We are now seeing three related trends. One is that you now have these open base models, raw language models that have been trained at great expense and have been put out there into the world, essentially as a public resource. That means that another developer or small business can come along and develop its own AI tools, build its own AI models and create its own AI ventures, without having to spend millions of pounds on compute and research.

To give you an example, the most powerful open language model was released just last week. It is a large language model. It rivals GPT-3.5—one of the models behind ChatGPT—for performance. And it was funded by the Government of Abu Dhabi. That is now out there in the world as a resource. Others can build on it. We have seen the same from Meta, Stability AI and others.

The second very important trend is this Cambrian explosion in specialised models that have been adapted for very particular tasks. We have given an example of some of those in our written evidence, but the bottom line is that, by building on these open base models, a developer

can come along and customise a new model that demonstrates better safety or performance for a particular task, such as a chatbot.

To put that into context, when Meta released Llama 2, which was a big, powerful open language model, our team fine-tuned that model within a matter of days. For a period of time last month, that fine-tuned model was the highest performing open language model available anywhere in the world. It was from a British company made available as a resource, and it was downloaded about 300,000 times over the past month.

The third very important trend that I want to touch on is the development of small language models: language models that you can train using desktop hardware, or that you can run as a user on a smartphone.

Together, all these trends are helping to make these language models safer, more useful and more accessible, and that will be essential if we want to see meaningful adoption of this technology across the economy. People tend to forget that adoption of language models in real-world settings is very nascent. No one has figured out the killer use case.

To Professor Lawrence's earlier points, they have limitations that make them very difficult to deploy in a sensitive environment. But if we have these open base models, and thousands of developers experimenting to build newer, specialised models, and smaller models that do not require super-compute to run or to train, we can start to drive adoption in useful ways across the economy.

There are risks associated with that, and I am happy to go into greater detail on our view of those risks as they apply to open-source development, but those trends are very important to be aware of. The UK's competitive advantage in AI will be nurturing and supporting all that grass-roots innovation, not just the frontier innovation in big models from big brands.

Lord Hall of Birkenhead: That is really interesting from all four of you. Can I go into the particular area of misinformation with Professor Lawrence? Dr Innes also mentioned it. There is also that word that sounds rather nice but is not: "hallucinations"—in other words, you are getting it wrong—and the impact that has. Can you help me to steer between those who think this is a fixable problem, and those who think it is inherent in LLMs and not fixable?

Professor Neil Lawrence: The jury is out on that. I have opinions, I try to give my government advice according to the honest broker. There is a spectrum of opinions, and it is difficult to know which is correct.

One of the dangers with this type of technology is that everyone says, "It's very complicated. There are a lot of things going on. It's very difficult". We have world-leading people in the UK. I think of the University of Edinburgh, for example, which has outstanding groups who do understand these things.

Sometimes when we think about policy, we can step back and ask, "What trends do we expect?" If I go back to the internet of 2006, pre

social media, I tend to think of the internet as this land of information with lakes of misinformation that you could all avoid, the dark web, et cetera. We have already seen this creeping process whereby the internet is shifting. It is becoming a sea of misinformation with islands of information, but there is great hope, because you can look at systems where people take a sociotechnical approach, which is critically important. You cannot deploy technical solutions on their own and expect to get answers.

Take Wikipedia. When it came out, everyone said, "This is the end of encyclopaedias. This is the end of truth", but most of us rely on it. It is not that it is always perfect, but the reason is that there is a dedicated system of editors who spend time, and there are processes around those editors. Even for controversial pages, we can generally rely on Wikipedia. Indeed, all these very large companies are basing a lot of this technology on this highly curated, human-edited dataset.

My expectation on the misinformation front, just trying to think perhaps beyond the three years you wanted, is that we will see that happening increasingly. We will see a return to brands and publishers that are standing up for data. One error that we are seeing in this space at the moment is that a lot of people are saying, "Watermark AI language model-produced output". It is very unlikely that people will watermark stuff that devalues it, but it is very likely that we will find mechanisms and approaches for humans to stand up for the text, whether assistedly produced or in whichever form, and say, "I wrote this and I stand by what it says, even though a computer helped me".

That brings us back to some of the most important lectures on this, Baroness O'Neill's "A Question of Trust" Reith lectures, which are all about how you cannot put process in place of trust. The error that everyone is making is assuming that that is not true for AI. Everything that she said in the three most important lectures on AI that I have ever heard applies today.

I sit here before you wearing my dead brother's jacket—he was a lawyer—because I want him with me today. A large language model can never say that to you. I sit here with a reputation. The evidence I give you is based on the work I have put in and who I am in society. That is what Onora O'Neill is talking about in "A Question of Trust". We are getting very distracted by the technicalities of it, when, a lot of the time, we should be looking at how we ensure that these are empowering people in their decision-making, not replacing people for consequential decision-making.

Q4 **Baroness Fraser of Craigmaddie:** Professor Lawrence, that was very powerful. I am really struck by this question of trust. I want to explore how confident we are about the data that we put in. I need to declare an interest, by the way. I am on the board of the British Library, which houses the Turing Institute.

This question of trusted knowledge is very close to the heart of the British Library. Where do IP and copyright sit in this space, not only for

publishers but writers, artists and authors? How confident are we that what we are putting in is of quality?

Ben Brooks: There are two very big issues there. In general, on the broader point about reliability and quality of data: we do not take a fatalistic view that AI will never be capable of delivering reliable information when and where it is needed. Eventually, in the future, we will have chatbots that support communication between patients and clinicians. We will have personal assistants that provide personalised tutoring. AI will support decision-making in some of our most important public and private institutions. We will get to a point where we are comfortable that the reliability of these systems meets acceptable levels.

My background is regulatory policy across many different sectors, and our regulatory systems are well geared to grapple with this problem. When we certify doctors and lawyers, we are not saying, "You can never be wrong". We are saying, "You need to assure us that you can produce this information and this advice reliably to a minimum level".

The challenge today is that the evaluation frameworks for measuring reliability and thinking about issues such as transparency are very underdeveloped. We do not have good benchmarks for assessing the performance of these systems in those kinds of sensitive environments, and we need a lot more investment and research in that space before anyone would be comfortable seeing language models deployed into these very sensitive environments.

There is a second issue there, which is the question of copyright and intellectual property. It is top of mind for us and other AI developers. From our perspective, large and diverse datasets are essential not just to make this kind of language model technology possible in the first place, but to make it safer and less biased over time.

To put that in context, our flagship image model, Stable Diffusion, accounts for about 80%, by some estimates, of AI-generated images globally. That is trained on 2 billion images that have been curated from publicly available images online. These language models are being trained on hundreds of billions or trillions of words of content—to Professor Lawrence's earlier point, a significant portion of the observable internet.

That is important not just to make it possible for these systems to read, write and vaguely reason, but also because that is how we start to correct for issues such as bias and toxicity. It is how we start to improve reliability. From our perspective, the training of these systems is an acceptable and beneficial use of existing data, and we believe that it is protected by fair use doctrine in jurisdictions such as the United States. That said, we are very sensitive to the emerging concerns among creators and other kinds of rightsholders about the use of that content, and we are working to address those emerging concerns in a number of ways.

To give you some examples from the image space, we have introduced opt-out good practices. We say to creators, "Tell us if you don't want

your content to be included in the training data for these systems". We have received about 165 million opt-out requests to date and are honouring those in upcoming training.

Another example in the image domain is tagging AI-generated content when it is produced through our computing platform. Again, it is not that every piece of content that is ever augmented by AI will need to be tagged, but if we are able to tag at least some of that content, it makes it easier for platforms and users to identify when they are interacting with AI-generated content or AI systems.

To the broader point about trust in the data that sits behind these models, there are a lot of different kinds of good practices and standards, and new kinds of developmental techniques, which can help to address those concerns. But a big part of this will be establishing an oversight framework in which we can measure reliability and transparency to an acceptable level. A lot of that does not exist today.

Baroness Fraser of Craigmaddie: Will a human have to be involved in the measuring?

Ben Brooks: That is an interesting point, because some of how we evaluate these systems today is exactly that: it is specialists in different fields sitting down and just playing around with these systems. Last month, one of our language models was subject to community-led safety testing at an event in Las Vegas that had been announced by the White House. Through the course of hundreds of thousands of messages, researchers attempted to elicit undesirable or unsafe behaviour from these models. That is one example of how we can assess for behaviour.

Another example is through standardised benchmarks or evaluations. There has been a lot made about how language models can pass certain bar exams, medical admission exams and things of that nature. Again, those are not very highly developed, but we will need more of those evaluation processes, particularly when it comes to larger models that may have novel and unforeseen capabilities.

It is in everyone's interests, open source or closed source, to understand those capabilities before we release a model out there into the wild. If we have those evaluation processes and frameworks, we can check whether there are unsafe capabilities there, and decide whether we expose or release that model publicly.

Baroness Fraser of Craigmaddie: Dr Innes, do you want to add or disagree?

Dr Jean Innes: I would like to agree and augment. To your question about trust, evaluation of data privacy and copyright is critical, but complicated and difficult. There are lawsuits going on already in the US over these issues.

Professor Lawrence touched on something that I find very important and interesting. You potentially move into a new regime on the internet with curated safe spaces using various technological techniques. The fundamental point is that we need to come together to work this out. We

do not know how to do this, and there will need to be a meeting of minds and a gathering of expertise.

If I can just make it real for a moment, we did some work with the Information Commissioner on addressing the explainability gap. Everybody is probably quite familiar with the term "black box". A lot of organisations are already using AI to make decisions that have impact on real-world people and individuals. This black-box problem is inherent to the nature of large language models and any neural network. There are many groups doing technical, thoughtful work to think about explainability. This is very high-priority work. We put together and released some guidance on how to explain AI-assisted decisions. It has been very widely downloaded. I make no greater claim than that it is a useful contribution to the debate. It is a pointer of the direction that we need to be going in.

On the point about trust, the other concept that strikes me is that of an online harms observatory, which is a place where you can at least start to curate, recognise and systematise the sorts of harms that we are observing. There is an example of that as part of the AI Standards Hub, which homed in on misinformation, hate and extremism. It unpacked the facts and found that 50% of abusive tweets were directed at 12 Premier League football players. This just gets underneath the concerning headlines and works through in a technical way to make a discussion that you can engage with rather than an amorphous concern.

In terms of trust, I see many cultural fractures in this space. As a scientist, I found CP Snow's "The Two Cultures" essay extremely resonant when I was studying. It is that sense of many well-motivated people maybe just missing each other a bit, because we do not have a shared sense of the issues, hence the importance of coming together.

Baroness Fraser of Craigmaddie: I was going to say that the examples that you gave us are quite stark. Mr Brooks, you mentioned biases. I find that concerning in this: the gathering of unintended consequences and unknown unknowns. Mr Hogarth or Professor Lawrence, do you have anything to add?

Professor Neil Lawrence: On sociotechnical solutions in the curated space, it is always important to have "socio" in front of "technical", but I would otherwise agree.

I would direct members to my colleague Sylvie Delacroix's paper, which looks at the changes to intellectual property that may be necessary and how we have to rethink. Copyright comes in only with the invention of the printing press. There is a fundamental distinction in copyright law between the notion of copying and reading. I am quite entitled to read anything; I am not entitled to copy it. That distinction is gone with these models. It is absolutely in tatters.

It is really important that we have institutions setting regulatory environments for companies like Stability AI, so that they can work towards the rules, but it is clear that there has been a radical shift, just as there was on the invention of the printing press.

The other point she makes in that paper, which is vital, is that there is a human right to access to culture. What is this model doing if it is absorbing everything that was ever written? It is true that it is not just about language, because we write about everything. It is a distilled version of human culture. What does that mean?

My colleague Sylvie Delacroix, when we convened the UK AI fellows, gave this talk in the Natural History Museum, a free museum, where children were running around looking at culture. We have a new question. What does it mean if there is restricted access to these models that absorb the whole of human culture but do not pay anything back into that ecosystem in any way? What does intellectual property look like for such a large community? We can then break that down to subcommunities such as indigenous groups.

There is a lot of interesting work on data intermediaries. I have been a driver of that. I know that the House of Lords has asked for the national data strategy to be pushed forward. I am worried that it has stalled. I have not seen anything come up on it in the last two years. It is in these pieces of legislation, albeit perhaps not AI pieces of legislation, where people have thought these issues through and are starting to push forward the mechanisms that will be necessary if we are going to look at what the new intellectual property landscape might be, again, probably going beyond the next three years.

Q5 **Baroness Harding of Winscombe:** After such great big thoughts, I bring us back to a bit more practicality and focus, if we can, on your task force, Mr Hogarth. Interestingly, the briefing that I have from last week is that it was called the Foundation Model Taskforce. I note from the update published at the end of last week that it is now the Frontier AI Taskforce. I wonder whether you could start by explaining what the objectives are and what you hope to achieve in the short and medium term.

Ian Hogarth: The core objective is state capacity. Thinking about Dr Innes's comment on the Warnock report, one of the reasons why we were able to navigate such a significant technological revolution there is that there was fundamental public sector capacity around that technology. There was a huge amount of capability within academia that allowed us to get our hands around it, put guardrails around the technology and decide how to steer it.

What has happened with AI is slightly different. If you think about the last decade, most of the frontier has been driven by private companies, so you have had a slightly strange dynamic whereby some very brilliant academics—such as Shane Legg and Demis Hassabis at DeepMind, Ilya Sutskever at OpenAI or Dario Amodei at Anthropic—have moved out of academia into the private sector and have used start-ups as a mechanism for accumulating very significant resources to bring to bear on this problem of scaling up these systems, getting more compute, getting more data and producing more capable systems.

We find ourselves in 2023 in a situation where the frontier is very much defined by private companies. That presents a couple of quite structural problems. The first is that, when it comes to assessing the safety of these systems, we do not want to be in a position where we are relying on companies marking their own homework.

As an example, when GPT-4 was released, the team behind it made a really earnest effort to assess the safety of their system and released something called the GPT-4 system card. Essentially, this was a document that summarised the safety testing that they had done and why they felt it was appropriate to release it to the public. When DeepMind released AlphaFold, its protein-folding model, it did a similar piece of work, where it tried to assess the potential dual use applications of this technology and where the risk was.

You have had this slightly strange dynamic where the frontier has been driven by private sector organisations, and the leaders of these organisations are making an earnest attempt to mark their own homework, but that is not a tenable situation moving forward, given the power of this technology and how consequential it could be.

At the Frontier AI Taskforce, we are trying to deliver on the Prime Minister's mandate. I took the role 11 weeks ago. The core thing I was told was this: "You report to me and the Secretary of State for DSIT. Rapidly build out a research team that can help us get to grips with the appropriate guardrails for this technology". When I came into the Department for Science, Innovation and Technology, there was one lone researcher with a background in machine learning who had dropped out of his PhD to help the Secretary of State as her special adviser, which is a testament to her vision for the department.

In 11 weeks, we have recruited a team where there is now 50 years of collective experience in Frontier AI. The idea there is that we can start to mark the homework in the sense of being able to run evaluations to look at the risks in these models. Dr Innes referenced cyberattacks. What risks do these models pose to rogue actors conducting cyberattacks?

The other thing that we are doing with the Frontier AI Taskforce on safety is recognising that some of these challenges are fundamentally matters of national security. An AI that is very capable of writing software, as many people have referenced, can also be used to conduct cybercrime or cyberattacks. An AI that is very capable of manipulating biology can be used to lower the barriers to entry to perpetrating some sort of biological attack.

There are experts in the British state who have devoted their lives to protecting us from cyber and biological attacks. We are trying to synthesise the expertise we are bringing to the task force in frontier AI research with the existing expertise inside the British state on national security frontier challenges. In some ways, that is expressed by the work that we are doing on the ground as well as on our advisory board. We have Yoshua Bengio, who is one of the pioneers in the field. He basically invented deep learning, the foundational intellectual idea behind these models that we see today. He has joined our advisory board. Paul

Christiano is one of the world's leading AI alignment researchers, formerly of OpenAI. He has joined our advisory board, but so have Anne Keast-Butler, the leader of GCHQ, and Matt Collins, the deputy National Security Adviser for the UK.

I think this is the first time that those four people have sat on an advisory board together, but in many ways it captures this fundamental challenge of frontier AI generating novel national security challenges, and that requiring us to rapidly build out state capacity so that we do not leave companies marking their own homework in such a critical area.

Baroness Harding of Winscombe: Where you have gone is very interesting compared to the conversation that we had earlier with Professor Lawrence and Dr Innes talking about the sociotechnical issues. I am struck that there is no ethicist on your advisory board, and no mention of ethicists. Dame Mary Warnock was a philosopher, not a biologist, and there is no one on your advisory board from civic society either.

A number of us here have spent a lot of time in the last six months on online safety, which is not just about national security but very much about how we protect the most vulnerable in society and everyday life. How will your task force address those issues, or is it more narrowly focused on the national security guardrails of large language models?

Ian Hogarth: First, this is 11 weeks in, so I would beg your patience. In 11 weeks, we have been pretty rapid in building out a team and an advisory board, but it is definitely work in progress. We are very clear that we will announce other members of our advisory board in due course.

To be clear, though, there is a certain urgency to the national security challenge, in that people are giving briefings on potentially quite consequential increases in risks in certain areas of national security. It is incumbent on us to try to tackle those. Ultimately, I think back to the WannaCry attack on the NHS system, where cryptocurrencies enabled a new attack in the form of ransomware. As a result, you had £90 million of damages, and processes shut down at over 80 NHS trusts, with a huge amount of patient disruption. Although these are national security challenges, they have a concrete bearing on everyday life if we do not get them right.

Helen Stokes-Lampard would probably qualify as a representative beyond national security and AI expertise. She is a practising general practitioner, as well as being a very experienced chair of the Academy of Medical Royal Colleges and having a broader capacity in the UK around general practitioners. I do not know whether that seems like the sort of person who would represent a different point of view, but we brought Helen on to the board because we felt that it was really important early on to get someone who could represent the front-line experience of a patient turning up, having diagnosed themselves with ChatGPT. Does that answer the question?

Baroness Harding of Winscombe: It does. I have been in your shoes

and I recognise that you have not had very long, but perhaps it is worth reflecting on the broader sociocultural challenges beyond pure national security. Is it fair to say, then, that the task force is focused on developing the risk mitigations rather than catalysing opportunities for the UK?

Ian Hogarth: No, not at all. It is a sequencing challenge. Right now, people inside the frontier community have started to point at some of these challenges around cybersecurity and biosecurity. We are trying to get ahead of that, because it feels like the cost of not doing so could be quite significant in the large-scale harms that that could lead to.

Once you have got to a place where you feel like you have got to grips with the challenges, you can really start to think about the opportunities. I am fundamentally someone who is very excited about the potential for machine learning to improve lots of aspects of our day-to-day lives. It is just a sequencing challenge: what do you tackle first? We have chosen to prioritise the national security risks up front and start to build that capacity around the other areas in parallel.

Baroness Harding of Winscombe: Could you give us a sense of how you intend to spend your money?

Ian Hogarth: It is primarily on recruiting expert AI researchers into government, by bringing people at the forefront of the field into government. That will give us a state capacity to develop the safety infrastructure required to assess all kinds of risks from these systems, to apply AI to public services and to make sure the UK stays at the front of this technology in terms of both risks and opportunities.

Baroness Harding of Winscombe: I know AI researchers are very expensive, but £100 million does buy you quite a lot of them.

Ian Hogarth: How many should we have?

The Chair: How many do we have now?

Baroness Harding of Winscombe: Yes, how many do you have today?

Ian Hogarth: We have 10 people so far with real frontier expertise, with PhD-level through to professor-level experience in the field. We are 11 weeks in. We are going to hire a lot more. These are some of the hardest to hire people in the world right now. It is a real challenge. You are offering to bring people into the public sector when they are being offered 10 times that amount to stay in the private sector.

Baroness Harding of Winscombe: Just to go back to where I was before, from everything I have seen, the great innovations in modern science happen at the intersection of disciplines. It sounds like you are building a single-discipline task force rather than a multidiscipline one. I am struggling to see how you will really be able to build guardrails if all you have are AI researchers.

Ian Hogarth: That is why I was trying to highlight the expertise we have been bringing in around national security. It may not seem like very much progress but, up until now, you have had private sector

actors trying to work out the national security implications of the technology and software that they are building inside their companies. That has not been connected up with the existing expertise in those areas inside nation states.

We are trying to join those communities together as a first step. Fundamentally, bringing in people such as Anne Keast-Butler and Yoshua Bengio is a first step towards having a more diverse group of people thinking about these challenges. Clearly, there is a long way to go. I am not suggesting we are done or anything like that. It is just the first concrete step.

If I have learned anything about building start-ups, it is that focus matters. If you try to do everything at once, you tend to do nothing. I am very conscious that delivering something concrete matters. The Prime Minister has arranged the first global summit on AI safety in November, which is seven weeks away. We are trying to make sure that, in these core questions around safety as it applies to national security, we are in a position to deliver on his mandate.

Q6 **Baroness Healy of Primrose Hill:** I just wonder what the other members of the panel think about the new Frontier AI Taskforce. I would like to start with Mr Brooks.

Ben Brooks: In the time that the task force has been around, the French Government have directed €40 million into open-source development, alongside €500 million to support grass-roots innovation across language models in France. "On croit dans l'open-source", President Macron said—"We believe in open source".

There has been a very concerted effort to support development and research not just at the frontier, but behind the frontier as well.

Frankly, the long tail of use-cases for this technology across the economy will not require a DeepMind-style model. They will require small models that you can build in-house. You can customise them without giving your data away to a company on the other side of the planet and you can inspect them. You can understand their performance and their risks. That is the contribution of open models to the AI ecosystem.

What we would like to see from the Government, if it is serious about becoming an AI superpower, is a sustained commitment to investment right across that ecosystem and not just in these frontier models, which frankly are developed by two, three or maybe four companies all based in the San Francisco Bay Area.

When it comes to safety, we absolutely welcome the task force. We particularly welcome Ian's appointment to lead the task force. But from our perspective, there also needs to be a more concerted effort right across government to support that kind of open grass-roots innovation.

Baroness Healy of Primrose Hill: Dr Innes, is it up to the task force to use taxpayers' money to secure AI for the well-being of the public?

Dr Jean Innes: It is a statement of fact that the majority of the cutting-edge work in this area is happening in industrial labs and, therefore, we cannot do what we have always done to tackle it. When you have budgets like these industrial labs have not just for talent but for compute—even getting hold of the chips you need is increasingly challenging—you fundamentally have to make a step change in what you are doing.

I fundamentally agree with that. I would also make the observation, as somebody who has worked within a government department, that this is a challenge for government. I welcome the creation of DSIT. Having a focus and gathering expertise is very helpful, but I know from having worked inside government that it is hard to find the people with the understanding to tackle these issues, particularly if you are trying to do it at pace.

I fundamentally welcome and really support the level of ambition. It also places the UK very well internationally. I risk sounding glib, but there are no borders with AI.

The Chair: We might pick up some of this in a moment when it comes to the specifics on government and risks.

Baroness Healy of Primrose Hill: I just wanted to ask Professor Lawrence about the AI Council. What is the difference, apart from the amount of money that has been given?

Professor Neil Lawrence: I am very supportive of these approaches. In fact, I agitated actively in the AI road map for the recommendation to have pathfinder projects with executive sponsorship at the most senior levels of government.

I lecture regularly in the Judge Business School to companies about this approach. That was born from experience in integrating machine learning technology, not large language model technology, within Amazon's supply chain, where I found that you needed that executive understanding in order to get the buy-in.

I have to be a bit careful because a number of us went into industry and had these very senior positions. We had this task, and we shared a lot of experience. This was not my idea. My favourite thing that one of my colleagues did in their company was to work closely with the CFO because, once the CFO understands the possibilities of these technologies, they start signing cheques across the rest of the board. You really need that spike to drive forward and alert people to the possibilities.

On the specifics of what has to happen next, it is worth talking and thinking about the concept of absorptive capacity. This is a concept about how an organisation absorbs external ideas. The classic work is on the different ways that Dupont, the defence company, integrated technology and chemistry into its work. I am not an expert in this, but, since everyone else gets to talk about AI, I will try to talk about management.

Broadly speaking, you have two possibilities. First, you can try to deploy close to the coalface. That is very much the Amazon approach. The machine learning we were doing was about trying to reach people who were integrated with the teams, who felt ownership of the problem and who understood the problem deeply. Otherwise, they just deployed things that did not work. This is why Amazon is very good at the assimilation and deployment of recent technology.

The alternative approach is one where you have a siloed central group that is focused on particular challenges. It is much better for the rapid assimilation of deep technological expertise into the organisation, but it really struggles to distribute that expertise within the wider organisation.

The fundamental challenge becomes about bridging and how you do that. Your project selection then becomes really important. How are you choosing projects? What are you choosing? You have to spend a lot of time discussing it, thinking about it and deciding which ones will make a difference to the things people care about.

The security angle is clearly a really live issue, but I am curious as to why we do not do this in the Ministry of Defence. If you look at the United States with its recent task force, it has a pure defence task force that is led by the Deputy Secretary of Defense, who already has this type of experience.

Going way back to September, when I first attempted to give advice on this area, my main argument was that we should be working with Five Eyes. We should be talking to Canada very rapidly around this because a combination of Canada and the United Kingdom is extremely powerful in this space.

The United States can ignore one of us on our own, and it always will. It is Canada's best friend, whether Canada likes it or not. We can work with Canada, which has really deep expertise that it has invested in, to leverage our expertise and start delivering on these defence issues. Of course, we also have shared common interests with the Five Eyes.

My suspicion is that a lot of that work is going on already. I know the Alan Turing Institute has been very active in convening that type of work across machine learning in the base case. I totally accept that security is an interesting direction. I can think about DSIT and the type of problems we are having around online safety, et cetera; in fact, I convened and chaired a group on behalf of DSIT to feed ideas into the task force, but we did not have a defence idea. They are very difficult to get hold of because of that community. Again, it is really important to work closely with them.

Some of the ideas were really interesting and instructive. We had to work closely with teams that are deploying things, such as call centres dealing with government interaction, so that those people from the task force could get a deep understanding of the issues that people are facing at the coalface.

Fundamentally, very large start-ups and companies are able to develop these technologies without caring about their social consequences

because their main business is to invent problems that we did not know we had and solve them by selling their product to us. If you watch the 1980s series "Yes Minister", you can see that the challenges government faces today have changed very little in 43 years. The question is how we adapt these new capabilities to service those problems.

Q7 **The Chair:** I have a couple of questions on the task force. Can I just press Mr Hogarth a little more on the £100 million? I assume you are not suggesting that is just going on salaries.

Ian Hogarth: I am just telling you what we have spent it on so far. In the last 11 weeks, we have primarily been bringing researchers into government. We have been trying to get people to leave their jobs at some of these great academic institutions or frontier AI companies and to bring them into government. If you want to assess the risk a frontier model poses to cybersecurity, for example, you ultimately need to have that expertise in fine-tuning these models and understanding how one can extend their capabilities in a particular area.

The Chair: Are you able to say simply at this juncture that the British public, for their money, are going to get something out of it or is it just a policy framework that will provide the steerage and the guardrails, or whatever you want to call them, as far as this new technology is concerned?

Ian Hogarth: If I think about the end of this project, what would make me feel like I have done a good job is if the average British person felt safer—people are quite worried about this technology in lots of ways—and were seeing a tangible benefit in day-to-day life. That is the fundamental opportunity that the task force has: to make people feel safer and to do things that make life in Britain better in some concrete way.

If we think about what that takes, feeling safer and making day-to-day life better via government requires brilliant civil servants. I have worked alongside some of them and have seen how great the people inside the Civil Service can be. The bit that is missing is the technical expertise to complement that. Fundamentally, I hope to bring a lot of technical talent into government to deliver on those two goals.

The Chair: I just have one final thing before we move on. First of all, thank you for taking on a public role and coming out of industry to do it. As you are discovering, it comes with an element of public exposure that you do not normally get when you are beavering away in your own business. There have been questions about conflicts of interest. If we are to bring somebody into this kind of role, it is inevitable that there will be conflict. That is a perfectly reasonable issue that just needs to be managed.

Could you say how your conflicts are being managed, particularly the business that is still investing in new technologies that are being developed by start-ups?

Ian Hogarth: Yes. Speaking very candidly, when I was asked whether I would do this, it was a bit of a daunting challenge because I have to do this on top of my day job. Fundamentally, this is incredibly important. I am doing it because it is extremely consequential and government needs to get it right. There are some wicked problems, to use Professor Lawrence's phrase.

The thing I was most worried about is that I have been an active investor in some of these companies. At the start, I sat down with the Permanent Secretary for DSIT, Sarah Munby, and we went through every investment that I have made. We worked out what was in the scope of the task force and then we worked out a series of mitigations, such as divestments, that would protect the task force from any financial conflict of interest.

Just to be explicit, that was a reasonably expensive thing to do. You are taking shares in very valuable companies and divesting them. There is some definite lunacy, in some ways, in taking an unpaid job and then divesting a load of valuable positions, but it speaks to how critical the moment we are in is.

In practice, that is what we have done. I have done it in partnership with the Permanent Secretary for the department.

The Chair: As I say, I thank you for that level of commitment and public service. I just think it is in your interest that there is as much exposure as possible of how those conflicts have been managed in order to protect you. I do not know whether DSIT has published anything that outlines the steps that have been taken but, if it has not, I might write to the Secretary of State and suggest that she puts that in the public domain to protect you as much as anybody else.

Ian Hogarth: Yes, I welcome that. These are commercial organisations. How they are handling divestiture and so on is a reason for waiting until it is all done. I have a ton of respect for Sarah Munby. She and I sat down and both of us were trying to be whiter than white about it. I very much welcome any scrutiny.

Q8 **The Lord Bishop of Leeds:** Welcome to public service. One thread running through the whole of this afternoon has been what I might call the tension between the technocratic or the technical and Professor Lawrence's point about preceding it with "socio" and the need for ethics and trust. We have even heard Wittgenstein quoted. I would throw in Immanuel Kant as well for good measure.

I want to focus on another tension, which is that held by government in balancing managing risk along with creating opportunity. What is the appropriate role of government in responding to the sort of advances and pace of advance that we have been talking about? What priority actions would you recommend for government over, say, the next one to three years? I wonder whether we might start with Ben Brooks.

Ben Brooks: We completely support the approach laid out in the White Paper, principally because the risk of these language models will vary

depending on the context. What safety and fairness mean for an NHS chatbot will be very different from what they mean for a social media widget or for anything else in between. We welcome the approach taken there.

As I said, if the Government is serious about the UK becoming an AI superpower, we need more than just sector-specific regulation. We need to sustain this culture of open access to data, open sharing of research and open innovation in models. That has helped to bring us to this point in the first place, but is also going to help us make AI safer and more effective over the next one to three years.

There are three areas where government can perhaps do a little more than what has been announced to date. One area, just from a pure policy perspective, is making open innovation and competition in AI an explicit policy objective.

For 20 years, we have lived through a digital economy that has one search engine, two social media platforms and three cloud computing providers. We are at serious risk of repeating these mistakes again with AI, perhaps in more serious ways than in the past. Competition has to play higher in the pantheon of policy priorities. To that extent, we are grateful for the engagement we have had with CMA and others around this particular issue.

The second area, which I have mentioned before, is sustained investment right across the AI ecosystem, not just at the tip of the spear, the cutting edge. We need deployers of AI, developers who are experimenting with models, and those who are building and training those models, to have access to standards, research and computing infrastructure to support that ecosystem.

There has been some work to date. The exascale computing commitments from the Chancellor are very promising, but that is a multi-year problem. There needs to be sustained investment right here, right now, today. I have given some examples today. The French and UAE Governments are supporting that activity and 'putting their money where their mouths are'.

Thirdly, not to put too fine a point on it, is the question of intellectual property. This is not an ancillary AI issue. This goes to the very heart of AI and, frankly, the creative industry would say the same as well. There has been a lack of clarity from the Government in its response to this committee's recommendations and what it intends to do going forward.

The committee understandably urged a lot of caution and further consultation in some of the text and data-mining reforms that were announced last year. We certainly believe there are other alternatives that would help to balance these competing interests. However, since then, the Government has announced that if the AI industry and creative industry will not agree on a code for licensing data, trillions of words of content and such, the Government will legislate that code.

That really sets the UK apart from the European Union, the United States, Japan and Singapore on this issue. If you do not create a

supportive environment for training these models, where you have clear, fair and practical rules governing that training, AI does not cease to exist; AI investment and development goes elsewhere. It goes across the Channel or across the pond.

We would welcome, in the very near term, more clarity from the Government about how they intend to resolve those tensions and competing interests. As I say, I am happy to go deeper on specific alternatives, but there is a lot of inspiration from Brussels and the United States about how different jurisdictions have tackled this issue. Again, this is at the heart of the problem. It is not just a nice-to-have. We cannot 'have our cake and eat it too'. We cannot become an AI superpower and duck the question of intellectual property. It needs to be dealt with now.

Dr Jean Innes: Your question was about the role of government and the priority actions. May I quote from Carlota Perez's work? She is very thoughtful about the diffusion of technology into the economy: "When the economy is shaken again by a powerful set of new opportunities with the emergence of the next technological revolution, society is still strongly wedded to the old paradigm and its institutional framework. The world of computers, flexible production and the Internet has a different logic and different requirements from those that facilitated the spread of the automobile, synthetic materials, mass production and the highway network".

I say that really just to take a step back and recognise how major a point we are at. That is probably a theme you are hearing from all your witnesses. Picking up on something which Mr Hogarth said, this is a really significant issue for government. This is not an issue for government that we cannot find the right way through. Convening the different voices and talents with pace is absolutely the right thing to be doing.

I remarked earlier that AI does not have borders. For us as an economy, there are pragmatic advantages to being on the front foot about AI, but this is important work that needs to be done. We are on the front foot, and in November we will have an opportunity to start to look at that international conversation. I can be more specific if you wish, but I wanted to express that this is a moment. It is one for us to capitalise on and, yes, bring together all the voices.

The Lord Bishop of Leeds: The question was about how to manage the tension between the two. Do you have anything to add, Mr Hogarth and Professor Lawrence?

Ian Hogarth: We have talked a little bit, in a different context, about open and closed systems. That is maybe a nice lens through which to look at this. It is one of the hardest problems. I thought it might be worth giving you my thoughts on that.

The Chair: We are going to come on to open and closed systems in the next question.

Ian Hogarth: I will save it for then.

Professor Neil Lawrence: I agree with Dr Innes. It was considered a courageous decision to create the Department for Science, Innovation and Technology. My understanding is that it was part of a package of reforms that Sir Patrick Vallance was urging around technological advice in government. We need to think deeper and harder. It is not all about spending money.

Mr Hogarth is here representing a £100 million investment over five years. The UKRI budget is something on the order of £7 billion a year. When we think about how we are spending that money, we have new opportunities with the merging of the councils under the UKRI banner and the strategic direction that the Department for Science, Innovation and Technology can set about how that is spent.

When Mr Hogarth talks about the lack of available expertise in government, I feel a little annoyed. I have been working with government for seven years under the employment of DSIT,[1] and I am at the frontier of these technologies. A large part of what we have been doing is trying to link those ecosystems up with the Alan Turing Institute and UKRI-funded grants. The announcement of the £30 million Trustworthy Autonomous Systems Hub a few years ago, which is managed mainly out of Southampton, and the recent announcement of £30 million for a responsible AI hub are the result of the work that the AI Council, in collaboration with UKRI, has been laying down during a period of a lot of churn in government to try to ensure the ecosystem is ready to go.

I am a little nervous about all this attention on quite a small investment. Let us be very frank: I appreciate that it is public money but, given the scale of investment we are talking about, and the scale of the challenge, which is to revolutionise the way we think about many of our institutions, this is not a lot of money.

I hope that you will get evidence from Dame Ottoline Leyser. I am sure that she is actively thinking about this. One thing I have noticed across this whole story is, when these things emerged eight months ago, how little attention was given to the existing investments and what they are doing, and how little that came up in the media.

I was utterly confused by it. We are simultaneously announcing that we will have an AI safety summit and that we are investing £30 million in a responsible AI hub. I am sure it is entirely due to a new department coming together and people finding their feet, which has happened across this period, but we absolutely have to leverage those capabilities.

I appreciate the Haldane principle. There is an interesting question about the extent to which the Haldane principle is really being deployed in practice. If you compare us with the United States, for example, there is

---

1 For clarification: the Department for Science, Innovation and Technology was established in February 2023. Previous employment has involved appointment to the Advisory Board of the Centre for Data Ethics and Innovation (via the then Department for Digital, Culture, Media and Sport (DCMS)). Work on the AI Council was conducted jointly across the then Department for Business, Energy and Industrial Strategy and DCMS; that board was not financially renumerated.

an interesting question about the extent to which we are bringing academics into the management of our research budget and our ability to respond actively to new directions. I appreciate that ARIA is looking at that, but there is more that we can do. This is a big amount of money. AI will affect all those existing sciences, and we need to think a little more creatively.

The Chair: That seems like a nice segue for Lord Lipsey.

Q9 **Lord Lipsey:** I have two quite practical questions, if I may. One is to Jean Innes. For years, the Alan Turing Institute was thought of as the gold standard among the institutions working in this country, but I have noticed an increasing range of criticism of it, not of the individual people who work there but of its sense of direction.

I notice that you joined only in July 2023. I do not know whether that had anything to do with those criticisms. Could you give us an overview of the criticisms and what your answer to them is?

Dr Jean Innes: May I focus on the latter?

Lord Lipsey: Yes, please.

Dr Jean Innes: I do not accept the criticisms. Our track record speaks for itself. It includes keeping our country safer by working with the defence and security community. We are the first public partner of MI5, I am very proud to say.

You could also look at National Air Traffic Services. We have done some very nice work in partnership with NATS and the University of Exeter to look at using AI to improve the resilience of that important critical national infrastructure, not just to improve resilience but to improve its environmental footprint.

I come back to the fact that we all need to evolve. I have talked about the magnitude of the challenges we are facing. Well before my arrival, in February 2023—in fact, I was applying for the job at the time—the Turing Institute published a new strategy with a level of focus on some key challenges, which felt to me absolutely the right direction to be going in. We fundamentally see the case for pace; we see the case for collaboration; and we stand ready to play our part.

Lord Lipsey: My second question is more widely based. We have just rejoined Horizon. How much difference will that make in this whole field? You look as if you are full of an answer already. You start and then I will move around the others.

Dr Jean Innes: I was just going to welcome it. It is part of that global outlook that we need to have. I suspect Professor Lawrence has a more detailed insight to offer.

Professor Neil Lawrence: These things are a little dangerous because they are integrated with challenges that the Government have to deal with. I try to avoid too much public comment on these things.

I will just point out that in the early days of machine learning, because of the funding ecosystem set-up, we found it very difficult to get EPSRC

funding. That was just because of the way EPSRC works. It works very well for established communities. You can see the quantum community has £900 million over 10 years, which I believe is more than AI is getting. I find that extraordinary. AI will likely overtake that.

We really struggled. At that time, European funding came with a lot of additional provisions, such as engagement with SMEs, all sorts of baggage and all the bureaucracy that we do not like, but it was a lifeline for me. I would not be an academic today if I had not had it. It was great to have an alternative route to explore these more interesting areas.

When I listen to the community, what really devastated a lot of people was the loss of the ERC grants. The bureaucratic part of European funding seems to fade away. I have never had one, but they seem like amazing opportunities for deep diving. Of course it[2] was a problem, but in one respect there are larger issues at play. The Government have to make decisions across those larger issues. In line with Dr Innes, it is a welcome return.

The Chair: I am going to move on to Baroness Featherstone, who will give Mr Hogarth a chance to talk about open and closed systems.

Q10 **Baroness Featherstone:** I will. The question is around where the balance of power lies. Will it be with the big tech companies or with open-source AI? I want to know where you stand on these issues. Personally, listening to you, which is fascinating, I am terrified by the motivation of big tech and I worry about the future. I want to know what is best for the UK. What should the Government be doing? What is the motivation? Where is the element of trust going to come from?

These are the issues. They go in such difficult directions. I would call it a million-dollar question, but it is more than that.

The Chair: This is within the framework of the choice between open and closed.

Baroness Featherstone: Yes. One has all these toxic actors and cannot be regulated very easily. The other can be regulated but is probably toxic because its motive is profit. I do not know; I am just listening to you. You are the experts.

Ian Hogarth: The reason I wanted to talk about open and closed systems is that it is a really challenging topic, and it is worth trying to pull out some of the nuance. As I understand it, this session is in many ways a table-setting session, where you are trying to work out some of the core challenges. Open versus closed will be a really big one for the next few years.

To frame it a little bit, I will just talk about some of the problems with closed source AI systems. Take GPT-4 by OpenAI. People say it is trained on the whole internet—Professor Lawrence said that. The reality is that we literally do not know what it is trained on because the training dataset is not public, so they are fundamentally more opaque and less

---

2 Clarification: "It" refers to loss of access to Horizon.

transparent. The companies state that they are doing this for good reason and that it is due to security considerations, but fundamentally you have more opacity and less transparency, which leads to these questions around intellectual property, training data, bias, ethics, et cetera.

Another challenge is competition. Ben Brooks mentioned that competition is good. I agree with that. Again, closed systems can have a "rich get richer" effect, where you get the capital to train the frontier model, which means people use the frontier model, so you get more capital and you can stay further ahead.

Finally, specifically in a government context, there is a sovereignty question as well. If you want to take some sensitive UK taxpayer data and feed it into a cutting-edge large language model, do you want to do that with the Google, OpenAI or Anthropic API or do you want it to be a sovereign system that the UK Government control? Those are some of the set of problems with closed systems.

The problems with open systems are quite different. First, it is quite loaded term. "Open" has a lot of positive connotations, but there are a couple of considerations there. One is that open source is a long-standing business strategy. Just because you say you are doing open source, it does not necessarily mean there is no profit motive underpinning it. You can look at the way Google developed Android as a way of catching up with Apple, iOS and the iPhone.

There is a brilliant blog post called "Commoditize Your Complement", which is about the business strategy of opening up a layer of the stack, so you can compete in a different area and undermine your competitor in that process. When Meta releases Llama or Stability AI releases Stable Diffusion, there are loads of positive things to point to in that, but there is ultimately also a business objective and business strategy behind those systems being open. It is not altruistic. That is what I am trying to get at.

The other point is that you can become a bit ideological about openness in technology. We do not do open gain-of-function research on biological systems. We have a pretty closed system for that, where we have biosafety labs. There is a question about how open you should be when you are dealing with potentially very consequential technologies that could destabilise aspects of society.

A closed system gives a Government more control in certain ways. It gives you something closer to an off switch, for example, which you do not necessarily get with an open system. With open systems you sometimes get this irreversible proliferation. Once it is out, you cannot put it back in the jar. That makes it harder to do precautionary deployment of certain things. A good example of that would be how some of the open-source image generation models have now been fine-tuned by malicious actors to generate child sexual abuse material. These systems are being used for really some of the most heinous things out there.

Finally, maybe the biggest structural challenge that we have to wrestle with regarding openness is that it can make things run more quickly. In a situation where the capabilities of these systems have run ahead of our ability to make them safe, that can also present its own risk because you can speed up the whole system even more.

There are some really exciting things about these open systems. If you look at work that Stability AI has done or that Meta has done with Llama, there are so many great things about it. There is a lot that I welcome. When I was building my start-up, I sued Ticketmaster for anti-competitive behaviour. I have real scar tissue around monopolies. It is troubling that we cannot say what data these systems have been trained on and that they present questions of sovereignty, but it is not helpful to pretend that open source is a panacea either. We need to be laser-focused on the challenges with open source and find a thoughtful and nuanced way forward.

Ben Brooks: I agree with a lot of what Ian just laid out, actually. However, I would make this point: it is not either/or. This is a diverse AI ecosystem. We will continue to have powerful and well-funded closed-source technology pushing the edge of the envelope, doing new and exciting things, including in other domains such as life sciences. We will also have, as I say, this renaissance in open models, helping to make AI safer, more useful and more accessible for ordinary people and the front-line organisations that can best figure out how to deploy it in practice.

At the moment, you have a bunch of companies sitting in an ivory tower saying, "This is how we think AI is going to be used". The reality is that we need those organisations to train and deploy their own models, and to give those tools to their own workers, professionals and analysts to figure out what to do with them.

It is not either/or.

One thing I will say is that we can bet on language models becoming critical infrastructure over the coming years. They will transform access to essential services. They will totally change how we search and interact with information online. They will, as I say, support analysis and decision-making in some very important public and private institutions. In that environment, it is more important than ever before that we can look inside the black box and that we avoid a centralisation of capabilities in one, two or three companies.

Ian has touched on this already. In that environment, open models have a valuable role to play. They are more transparent in many ways. You can look under the hood; you can verify risks; you can verify the performance of these systems before you use them in your NHS chatbot. They lower barriers to entry. A developer or a small business can come along and build on these models without having to spend £10 million or £100 million on training its own model from scratch.

A very important point that Ian mentioned is this issue around what I would call strategic independence. Businesses and public sector agencies

building AI capabilities have a choice. They can either give all their data to a company on the other side of the globe, and give up control of that AI capability. Or they can build on open models, develop those capabilities in-house, keep that data in-house and retain control of those AI models.

These sound like arcane issues in some ways, but they go to the heart of what the digital economy will look like for the next 10 or 20 years. In that environment, open models have a very important role to play. They do have, as Ian mentioned, some very serious risks. One is that the supply chain is much more complex than it is for a closed source model.

If we take OpenAI, it develops the model; it trains the model; it hosts the model on a computing service; and it builds the chatbot that a user will interact with. All of that happens within the same company.

In the open-source or open environment, one company might train the model. It might be the UAE Government or the French Government. Another company or another developer will come along and fine-tune or adjust that model for better performance on a specific task. Someone else will come along and build the tool that uses the model, which the user interacts with. One of the risks is that you get a diffusion of responsibility and it is not quite clear at what point you are meant to be mitigating different risks.

There is one point I would add to that. This is a complex challenge, but it is not a challenge for which there is no answer. There is no silver bullet to safety in AI. Sometimes we get this sense that with closed-source models you just turn off the tap and all your problems go away. If open-source models and open models are going to be important, you cannot just do that. Once the genie is out of the bottle, it is out of the bottle. However, there are layers of mitigation that we can put in place to make it easier to do the right thing and harder to do the wrong thing with these technologies.

Baroness Featherstone: When you say that we can put it in place, what is the role for government and policy in that?

Ben Brooks: That is a great question. It is important to distinguish between three kinds of risk. These are frequently conflated, which makes policymaking very difficult.

There is a question as to whether an unsuitable AI model has been deployed in a sensitive environment. That is a product safety problem, basically. If you have a chatbot giving you financial advice, is it reliable? That is something where we expect regulators to come in and establish very clear performance requirements, saying, "Here is the minimum reliability, the minimum level of transparency or interpretability, and here is the assurance you need to give us, as the FCA, the ICO or the MHRA, to show that this is safe and fit for purpose".

There is a second class of risk—and this is much more what the task force will be looking at—in these novel emergent capabilities that typically occur in the larger and more powerful models. In other words, can you synthesise a dangerous biological molecule or compound, for

example? In the future, government will have a very important role to play there. There will be very robust evaluation frameworks. If you are building one of those models, you will have to go through that evaluation process before you can perhaps release or expose that model publicly.

There is then a very complex middle ground of risks; I would describe these as intentional misuse. You take these models, you adapt them for purposes for which they were not intended, and you do something nefarious. That could be the generation of unsafe content, as Ian alluded to. It could be the production of disinformation for malicious purposes. It could be identifying software vulnerabilities and trying to exploit them.

Again, there is no silver bullet, but there is a lot that you can do in the model itself before you release it out to the world. There is a lot that the provider of the application, the chatbot, can do to filter out some of those prompts and some of those outputs. And there is a lot that the wider information ecosystem can do as well to detect and identify unsafe content, for example, before it gets amplified through the web.

It was announced elsewhere this morning but, in a few hours' time, the White House will announce that Stability AI has signed up to its voluntary AI commitments. Among those are a number of these mitigations that we are talking about. How can you fine-tune the model to avoid some of that unsafe and undesirable behaviour before you release it into the world? How can you tag AI-generated content so that social media platforms and search engines can identify it as it moves around the internet? Again, there is no silver bullet, but there is a lot that we can do there at different layers in that tech stack to mitigate those emerging risks.

One of the challenges with the framing of the question of "open or closed" is that it implies you cannot have open "and" safe. The reality is that you can have open and safe. You can have open and unsafe, but we are trying to show that you can develop open models; you can release them; people can experiment with them; and people can deploy them in a safe and secure way. That is a lot of our work today.

Dr Jean Innes: We have just heard some very lucid, thoughtful commentary on open versus closed. This is properly difficult. May I home in, then, on what we therefore need to do?

We need to bring the different aspects together at pace. I think the direction of regulation, which is sector-based, is right. Sir Patrick Vallance published a thoughtful report on the sectoral approach to digital regulation. These are cross-discipline technologies, so it is right to take a sectoral approach, but that central risk function will be extremely important in supporting all regulators to make the right interventions in their different spaces.

Again, this loops back to that question of expertise and the sharing of approaches. I see a strong role for government in building that central risk function, which can then support the regulators to approach their different sectors appropriately.

Professor Neil Lawrence: To get an insight into the US point of view, there is Eric Schmidt's report on security. It is a few years old now, but this seems to be the way the US is going. It was about machine learning and AI in general. A lot of Eric Schmidt's public statements since then have said the same things.

He says that the only way of doing this is to back large tech. My concern is that, if large tech is in control, we effectively have autocracy by the back door. It feels like, even if that[3] were true, if you want to maintain your democracy, you have to look for innovative solutions.

The observatory[4] is very important, and the nature of the observatory is key. It is unfortunate that it has been outside the conversation when we have been looking at the AI White Paper. A lot of these things were thought about in the construction of the road map and the national AI strategy, et cetera.

Q11 **Lord Kamall:** I should declare my interest as being on the advisory board of the Startup Coalition, formerly the Coalition for a Digital Economy. I have looked at technology issues for a number of years. This issue is very interesting. As you say quite rightly, it is not black and white between closed and open. There are different dimensions.

One dimension that has not been discussed is this whole argument about open and proprietary data. There will be those who want to challenge proprietary data and say, "If we don't have open data, we've got a very limited dataset and it will be very biased". The other side of the argument says, "Hold on a minute. There are some intellectual property issues here about copyright, et cetera, that say you have no right to that data".

How do we resolve that? It seems incredibly difficult. If you fully respect copyright and intellectual property, you are training your machine learning systems on limited data that could well be biased. Does anybody have an instant answer to that?

Ben Brooks: It is top of mind for us. Perhaps I can venture a quick answer. As I say, from our perspective, training on these large public datasets is not just acceptable under US law, for example, but is absolutely necessary. Otherwise, you will get an AI model that either does not work at all or, to your point, is extremely biased and unsafe.

It is important to understand what is happening in the model. The model is not storing that data. It is not acting as a search engine for existing content. As I say, it is learning to read and write; it is learning a body of knowledge and then applying that knowledge to generate or analyse new content that it has not seen and which perhaps does not appear anywhere in the dataset.

---

3 Clarification: "that" refers to the position attributed to Eric Schmidt referenced in the previous paragraph.

4 Clarification: "observatory" refers to the "coordination layer" referenced in the Department for Science, Innovation and Technology, *A pro-innovation approach to AI regulation,* Cp 815 (March 2023): https://www.gov.uk/government/publications/ai-regulation-a-pro-innovation-approach/white-paper#executive-summary

Sometimes we conflate a language model's apparent ability to recall information with the way it is trained to understand and comprehend language. They are trained to read and write. As an ancillary artefact of that process, they can sometimes recall information that was overrepresented in the dataset, but that is just an ancillary artefact of that process. Coming back to the reliability and trust point, that is not really how these models were intended to be used. It just so happens that they can be used in that way.

All of that is to say that we think it is going to be essential to preserve the culture of open and permissive training that helped to get us to this point. There have been a number of text and data-mining options laid on the table historically. Many of those would do the job. Of course, there is also the fair use doctrine that we see in other jurisdictions.

If we think about what we have at the moment, either retaining the status quo—which pushes this development activity to other jurisdictions—or trying to force the AI industry into some kind of statutory licensing arrangement by another name is going to have very serious consequences for the industry's ability to develop here in the UK and, as I say, will have consequences for safety and bias.

The Chair: I am going to cut in here because the Minister is on his feet in the Chamber. I am slightly worried that we will have to abort before we have properly concluded. I was going to say at the end anyway, as part of thanking everybody, that we will have a session specifically about copyright and some of the challenges that have been raised and exposed in the course of today's session.

One reason why I want to say that loud and clear is that some of those listening to this session will have their own strong perspectives on this topic. This is just to reassure you that we will be coming back to that.

You have all been incredibly generous with your time this afternoon. We are hugely grateful to you. As Mr Hogarth said a moment ago, we try to use the first sessions of a new inquiry to set the scene for what follows. What we have also tried to do today is to be quite challenging and testing with our witnesses and to make sure we understand some of the big issues and big questions.

As I have just said, we will pick up some of those questions, whether on copyright or other things, in the course of the inquiry. On the questions we asked around the current infrastructure, if I can put it like that, and the various government bodies or arm's-length bodies that we discussed earlier in this session, we will clearly want to explore that with Ministers when they are before us later. We have lots to ask the regulators too, by virtue of the evidence you have given us today.

I just have one final quick question for Mr Hogarth. Will there ever be a British sovereign LLM?

Ian Hogarth: How nuanced can I be with the answer?

The Chair: Can you answer it in a couple of sentences?

Ian Hogarth: I may be able to. The core question is this. A cutting-edge

frontier model costs, let us say, $100 million to train. Is that a good use of taxpayers' money versus taking a prior-generation open source model and trying to fine-tune it, extend it and find a more cost-efficient way to achieve a similar outcome? That is an open question.

This is the sort of thing we will be exploring as part of the Frontier AI Taskforce on the opportunities side of the equation. These are very expensive systems. If you want to do all the pre-training yourselves, you are talking about tens of thousands of top-spec GPUs and compute clusters. Next year people will have billion-dollar compute clusters to train these frontier models. It is a tough judgment call, and one that I am just going to be helping the task force give some advice on.

The Chair: Is there a deadline by which that decision has to be made?

Ian Hogarth: I do not think we should be in a rush to spend public money before we have really thought it through. I would not say there is a deadline.

The area where there is a bit more of an urgent question is what to do around compute. There has been some great work already around the exascale cluster up in Edinburgh, but there is an ongoing question around AI-specific compute. That is just another thorny challenge the department has to wrestle with.

The Chair: Thank you all very much again for your time today. I am going to draw this to a close; I am so sorry. I am very grateful to you all.