HOUSE OF LORDS

Communications and Digital Committee

Corrected oral evidence: Large language models

Tuesday 19 September 2023

3.35 pm

Watch the meeting

Members present: Baroness Stowell of Beeston (The Chair); Baroness Featherstone; Lord Griffiths of Burry Port; Baroness Harding of Winscombe; Baroness Healy of Primrose Hill; Lord Kamall; Lord Bishop of Leeds; Lord Lipsey; Lord Young of Norwood Green.

Evidence Session No. 3 Heard in Public Questions 21 - 28

Witnesses

I: Professor Stuart Russell OBE, Professor of Computer Science, University of California, Berkeley; Professor Phil Blunsom, Chief Scientist, Cohere; Lyric Jain, Founder and Chief Executive Officer, Logically; Chris Anley, Chief Scientist, NCC Group.

Examination of witnesses

Professor Stuart Russell OBE, Professor Phil Blunsom, Lyric Jain and Chris Anley.

Q21 **The Chair:** We now move on to our second panel of witnesses, where the main focus of our questions will be around risks, the last panel being predominantly about opportunities. We want to explore the risks that are known to us and that we are facing with LLMs right now. We will also talk about emerging risks, as much as we know what they are, and the timeline for those. We will cover headline understanding, different risks, the way in which they are understood by different parts of business and government, and the reactions that are proposed to mitigate them. I invite our witnesses to please introduce themselves and the organisation that they represent, if they are representing an organisation.

Lyric Jain: Hi, everyone. Thank you all for your time. I am the founder and CEO of Logically. We tackle disinformation by leveraging artificial intelligence and subject matter experts. We support our partners across the world, including Governments in the UK and the US, as well as social media platforms, using a combined approach that leverages both artificial intelligence and subject matter experts to identify and respond to disinformation.

Professor Stuart Russell: I am a professor of computer science at the University of California, Berkeley.

Professor Phil Blunsom: I am a professor of computer science at the University of Oxford and the chief scientist of Cohere, which is a company that builds large language models and sells them.

Chris Anley: I am chief scientist at NCC Group. We are a multinational information security business listed on the London Stock Exchange and headquartered in Manchester.

Q22 **The Chair:** Thank you very much. I will ask the first question, and come first to you, Professor Blunsom—I should have said, by the way, that we will not ask all four of you to answer every question we put to the panel, otherwise we will be here much longer than the time we have allowed. In headline terms, could you give me a sense of the debate between the limited nature of risk when it comes to LLMs versus the existential risk? In that big debate about generative AI, where should we see LLMs? Is it as something that presents limited risk or existential risk? Could you give us a flavour of that, a simple answer, because I realise that that is a question that could take for ever if we are not careful?

Professor Phil Blunsom: Yes. There is a lot in there, and I think it is good to pose a question at the level of AI and then work to large language models. Existential risk, of course, refers to a risk of extinction for humanity and limited risk, I guess, generally refers to anything less than that, which we might call more mundane risk, but it is a still significant risk to those whom they affect.

It is a broad space and discussion philosophically as to what different existential risks AI could pose. I would not try to cover them all. Some of

them are based on ideas of superintelligence—an intelligence beyond human intelligence—that would have an inherent advantage and could use that to bring about our demise, I would say, and other possible scenarios. Limited risks are those more associated with things that people identify around national security, disinformation, election security, crime that such systems might help in some way, and then negative adverse effects from users of those systems. There is quite a range there.

Many experts would definitely put different technologies in different places there. I do not see a strong existential risk from large language models, however, that would rise to requiring a national response on the level of other more present existential risks that we are very well aware of, such as nuclear proliferation, climate change and so on. As a researcher and professor, I put large language models in the limited risk space, but how we would define those different risks is complicated. I guess that is what we will go through in this panel.

The Chair: Thank you very much. You are quite right. We will come on to the specifics of those in a moment, but I will invite Professor Russell to give us his view as to where LLMs sit within that debate about limited versus existential risk.

Professor Stuart Russell: I would not completely disagree with Professor Blunsom on the risk from large language models as they currently exist. In my view, large language models are not on the direct path to the superintelligent system that Professor Blunsom mentioned, but they are a piece of the puzzle and we are in the process of understanding exactly the shape of that piece and where it fits in the puzzle. When we do, that will be a very big step towards creating super-intelligent systems.

There are a number of intelligent behaviours that they do not currently exhibit, including the ability to construct and execute long-term plans, which seems to be a prerequisite for using intelligence to overcome human resistance, for example. However, there are many research groups that are trying to fix those problems as we speak and making rapid progress on that. Some of my distinguished colleagues think that we might have as little as five years to figure out what to do before the cat is completely out of the bag. I am a little more conservative, but I could not say with any certainty that it will take more than 20 years.

The Chair: We will move on to specific risks and that will be an opportunity to involve all our witnesses. On what you have said about making a long-term plan, is that how you would define something specific to describe an existential risk? Sometimes, one of the things we grapple with when we talk about existential risks of AI, or in this case LLMs, is that it is hard for us to get a hand around it and say, "What are we talking about here?" Is that what you would describe?

Professor Stuart Russell: I think that it is one way in which the systems could, in essence, achieve objectives that turn out not to be aligned with what we want the future to be like. The most commonly cited failure mode, if you like, is AI systems that are designed to pursue

some objective and it turns out that this objective, when taken to its logical conclusion, leads to behaviours on the part of the system that are extremely unpleasant for us; for example, consuming all the oxygen in the atmosphere or changing the temperature of the planet to suit them and not us. There could be all kinds of ways that a sufficiently intelligent system could lead to extinction.

There are some risks that do not require that unified, single super mind. For example, people have shown that fairly simple reinforcement learning algorithms, when placed into a competitive market where they are producing goods and selling them, will learn to collude with each other with no communication whatever. This is called tacit collusion in economics and it has long been suspected that human corporations engage in it, but it has been proven that reinforcement learning algorithms learn to collude with each other in setting prices far above the cost of production without any communication whatever. For example, because AI systems are conversing with hundreds of millions of us all the time, they might gradually change our view of climate change so that we pay less attention to it. They might gradually make western countries more hostile to China and vice versa so that we end up having a nuclear war and we do not even know why we are having a nuclear war. We might end our civilisation without knowing that the reason we are ending our civilisation is that the AI systems gradually moved our opinions and our political processes in unhelpful directions.

Lord Griffiths of Burry Port: You said all that with a smile on your face. It frightened the life out of me.

Q23 **The Chair:** I will ask one further question, and hope there is a straightforward simple answer to this. When you talk about the puzzle and that this is a piece in the puzzle, can you tell us what another piece in this puzzle is? Is that another way of us being able to identify what we need to avoid dealing with? Do you see what I mean? What is another piece of this puzzle?

Professor Stuart Russell: There are many pieces of the puzzle that AI has already developed over the last eight decades of research. I think that the capacity for long-term planning in the real world is probably the biggest piece and we do not really know how to do that. To give you an example, AlphaGo is amazingly good at making what we would describe as long-term plans in the game of Go. Those plans might be 60, 70 or 80 moves into the future. That is how they outplay human beings. By analogy, we could say, "If you could do that in the real world, you could just outplay human beings in the real world just as they do on the Go board". However, if you are a physical robot and you have to send a command to your motors every millisecond—1,000 times a second—and you are looking ahead 60 or 70 steps into the future, that is not even a tenth of a second. Techniques that we have for what we think of as long-term planning are completely ineffective when you try to scale them up to the real world.

How do humans manage it? We manage it through hierarchical abstraction. We make a decision to come to a committee meeting at the

House of Lords. We do not decide, "Which foot will I move first?" and then which other foot and then, "Will I take out my card with my left hand or my right hand?" We make decisions at high levels and then we gradually refine those all the way down to the motor control steps that we need to execute the plan. That is an incredible human capability. We are able to put people on the moon, to build the House of Commons and so on. If we could understand how to do that, it would make AI systems far more useful to us, but it would also make them far more dangerous if they could deploy those capabilities in directions that are harmful to us.

The Chair: I will move on, otherwise we will never get through all the specifics.

Q24 **Baroness Harding of Winscombe:** I want to bring us back to the more immediate risks that either are already emerging or could emerge within the next one to three years, let us say. I want to get to each of the specifics in turn, but maybe we could start, Mr Anley, with what you see as the most immediate, specific risks of large language models.

Chris Anley: Certainly. As a business, we respond to a great many cybersecurity incidents directly and we are seeing little to no direct evidence of the use of AI in the incidents that we respond to. We also carry out horizon scanning threat intelligence research, related to incidents that we are not directly responding to ourselves, and in that research we are seeing anecdotal evidence of the use of generative AI in social engineering and fraud; for example, creating fake emails to persuade users to open a web link or a document, creating fake websites that look very much like a legitimate website in order to trick users into entering their credentials, and the use of voice cloning, creating an artificial clone of someone's voice that sounds exactly like them but can be made to say anything they want.

In our own research on the use of large language models by attackers, we find that they give an attacker a small improvement in efficiency, as was mentioned, by writing malicious programmes or similar, undertaking their activities, and a small gain in capability, perhaps allowing them to do some small thing they could not previously do. Overall, the increase in cybersecurity risk from AI today in verifiable things we are seeing is small to moderate, but that comes against the background of hugely increased cybersecurity risks today from ransomware and, to a lesser extent, supply chain attacks. Even a moderate gain in attacker capability is noteworthy and worth monitoring.

Baroness Harding of Winscombe: That is extremely clear. I am mindful that we have quite a few risks to go through so I will keep moving, if that is all right. I wonder if we can go to you now, Mr Jain. I expect that you will want to talk more about misinformation and disinformation, but what do you see is the biggest clear, immediate risks of large language models?

Lyric Jain: You guessed rightly. We see the world through a lens of disinformation but at Logically we have already seen risks emerge in the field of disinformation as a result of LLMs. We have had things like deep

fakes and so on for a good three or four years, but for about two years now we have been tracking the use of something that looks like large language models in campaigns driven by nation states. Particularly in the last two quarters we have seen a significant uptick in campaigns online that create either deceptive profiles, deceptive content or deceptive interactions with audiences, be it single users or groups of audiences, to create fairly sophisticated disinformation campaigns.

In our view, there are probably three key ways that this will play itself out over the next 18 months or so. First off, the biggest delta is the accessibility and democratisation of such technology. If we look back to maybe 2016 and the work that the Internet Research Agency did to spread disinformation during the 2016 US presidential elections, it cost it some $10 million to $20 million to put together that campaign. Such campaigns, experts believe, could be replicated today for about $1,000. That is probably the biggest challenge. The sheer volume of disinformation that we will see online will increase significantly. It will not be just big nation states which have a country's apparatus as well as significant money and people behind such campaigns that are successful but anyone who has a few thousand dollars, or even a few hundred dollars, to spare.

We will probably see an increase in hyper-realistic disinformation, which even experts will struggle to distinguish, and probably some well-timed interventions that may lead to some black swan events, either around key election events around the world or during times of heightened geopolitical tension. I think that this is an example where LLMs and the risks they pose do not necessarily lie in the future but have been playing themselves out over the last few quarters and will only increase as we look ahead to some key geopolitical events over the next 15 months.

Baroness Harding of Winscombe: Thank you. That is also dangerously clear. Professor Blunsom, where do you see further risk?

Professor Phil Blunsom: There were a lot of interesting points there on various risks. If I were thinking about hierarchy, I would say national security is a big thing that we should be thinking about and the implications of these models at that level, as has been mentioned, in the world of cybersecurity and crime. We are talking about the broader class of generative models here, and it is clear that in the short term it is particularly image generation and the one that people are much less aware of, which was mentioned, audio generation. I do not think it is widely known that today we can produce perfect replicas of people's voices, so that poses a big risk for those who do not realise that something that might sound exactly like someone they know is not necessarily them. The obvious case is banks that use that for security. There are those risks, and they are very clear and present. There is a degree of newness to those and people needing to become aware of whether you can trust a video you have seen of a politician doing something in particular or a particular recording.

Text as a medium poses different problems, partly because we do not inherently trust text already, not like what we see or what we hear. It is

a different medium for communicating ideas and poses different problems. Those are mostly around scale and whether these technologies can increase the scale at which someone can produce a misinformation campaign. That is disinformation, national security and crime. Other risks are the unintentional ones: bad products or misuse of these models in ways where they are not reliable in domains where they should be used. For instance, anything involved in making a decision that will have a real outcome for someone's life—the legal and medical sphere; any of these things should obviously be heavily regulated.

Baroness Harding of Winscombe: Could you give us a bit more specific insight into the misuse of the models or poor product design, for example? What is real today as opposed to potential risk in the distant future?

Professor Phil Blunsom: As we are in a very rapid stage of development and these things are developing quickly, those using the models and sometimes those developing them are not necessarily aware of all the pitfalls and how they will behave in different situations. Thus we should be careful about putting them into situations where, if they produced incorrect information, it is likely to have a real adverse outcome on someone. This is not just language models, but AI in general should not be used in decision-making situations—predictive policing, things like this—where the consequence of an incorrect prediction is very real for someone. This is different from a casual conversational model or something that is not being relied on for something we might call a safety-critical decision. Part of that is, first, the awareness.

One of the unique challenges with language models is that it is very easy to anthropomorphise them. It is easy to think that, because they interact like a human, they behave like a human and reason like a human. They do not. They might give you that simulated feel, but that is not what is going on. It is very easy to miscalculate how it will behave in a certain situation. I think that is something that will come with people being more exposed to these systems in general, just like being exposed to Google search. In the 1990s, people would use a search engine and expect it to reliably give good search results, but over time we have become used to the flaws in these systems and we know how to work with them. There is that development.

In many ways, this is a tool, so we need to understand the limitations of the tool, the situations in which it should be used and the situations in which it should not be used.

Baroness Harding of Winscombe: Can I just make sure I understood what you said? There was an awful lot to unpack in it. In our previous witness session, one of our witnesses, Dr Webster, said something that I thought was very helpful: to think about decision support versus decision-making. Am I correct that I just heard you say that you should not use large language models in decision-making where the consequences are very real for the human beings involved? That feels like a big statement that you have just made.

Professor Phil Blunsom: Yes. If you are making decisions about what shoes to buy, this might be something where you are happy to use a large language model.

Baroness Harding of Winscombe: A dangerous thing to say.

Professor Phil Blunsom: Depending on the consequences. In the medical domain, such as with hallucinations or these sorts of things, it is anywhere that the output of that decision is safety critical or the information provided is incorrect. Just as you would not necessarily do electrical work at home based on googling how to do that, you also would not ask a language model how to fix your electrics at home and go ahead and do it. There is definitely a need to understand what is safety critical and where AI systems in general—not just language models but other decision-making systems—are appropriate or not.

Baroness Harding of Winscombe: That is a helpful but quite strong set of health and safety warnings to apply to this technology. Professor Russell, you have been very patient.

Professor Stuart Russell: I just want to add that OpenAI explicitly says that we should not use these systems for any high-stakes application. It is already saying in writing what Professor Blunsom is recommending.

Baroness Harding of Winscombe: Is there anything else that you would like to bring out about immediate risks that you see from the technology?

Professor Stuart Russell: I had disinformation on my list as the number one risk, and I think that Lyric has spoken to that very well. I know some of the national security people in the US are worried about these systems facilitating the development of biological weapons by terrorists, for example. This is one of the things that they are trained not to do. After training on lots of text, there is a phase of reinforcing learning with human feedback where they are told that they have misbehaved. If the system answers a question it is not supposed to answer, the trainer basically says, "Bad dog" and keeps saying that until the system stops answering those questions. That is how we try to make them safe.

It turns out not to be very effective and, recently, groups including mine have shown that by prefixing your question with a particular string, which is unintelligible, it works very well to bypass the safety training. We also found that a particular image can work very well. We have a picture of the Eiffel Tower that has been invisibly modified and that will cause the system to answer any question you want, including all the questions it has been trained not to answer. The security methods that exist are ineffective and they come from an approach that is basically trying to *make AI systems safe* as opposed to trying to *make safe AI systems*. It just does not work to do it after the fact.

Baroness Harding of Winscombe: What is the logical conclusion of that statement?

Professor Stuart Russell: If we were to take an approach similar to that which we take with nuclear power, for example, where the regulator says to the provider, "Show me that your design has a failure rate of at least one in 10 million years" and if the provider cannot show that the developer goes through—

Baroness Harding of Winscombe: You cannot build the nuclear plant.

Professor Stuart Russell: Right, you cannot build a nuclear plant. If we were to ask the developers of large language models, for example, "Could you provide a guarantee that your system will not replicate itself?", I do not think that they can do that. The way in which we should think about regulation is: what do we want the safety requirements to be? Then the onus is on the developer to meet them, not the other way around. We should not build our regulations around what the developers can currently satisfy, which may be very little.

Baroness Harding of Winscombe: You have used nuclear power as the analogy. The other one I have heard described is to think of this in the same way we do the development of medicines and medical devices. You have to prove that they are safe before you launch them, not the other way around. Do you think that that is a fair similar analogy?

Professor Stuart Russell: Yes. I think that there are good analogies also to aviation. The power of recall applies to drugs and to aeroplanes. We grounded all the 737 MAXs for a year or so when they exhibited poor behaviour in the field, and we should do the same with AI systems. If an AI system starts defaming real individuals, for example, if it does it once, it will do it again. It is not a one-off thing.

It is also worth pointing out that, with aviation accidents, if seagulls fly into the engine of an aircraft causing a crash, we take countermeasures but we do not expect other seagulls all over the world to immediately start doing the same thing, but that is what happens with software. We should think of this as adversarial safety, not accident safety.

Baroness Harding of Winscombe: That is very helpful. I fear I am moving on to my colleague's next question.

Q25 **Lord Young of Norwood Green:** It was Donald Rumsfeld who said there are knowns and there are known unknowns and then there are unknown unknowns. Do you worry about that?

Professor Stuart Russell: Is that a question to me?

Lord Young of Norwood Green: To both of you, yes.

Professor Stuart Russell: The risk that we face, as I have said many times, is that we are creating systems that are more powerful than us and expecting to have power over them for ever. How we solve that problem is not immediately obvious. As I mentioned before, the failure mode is systems that are pursuing stated objectives, so the problem always comes from the fact that we do not know what we are missing in the way we formulate objectives for our AI systems until it is too late.

Professor Phil Blunsom: That is an interesting question. In some sense, the unknown unknowns for a scientist are fascinating to want to discover and understand. In the systems we have, yes, there are known things that we do not know about the systems, but we do have a good sense of how they interact with these different issues. We have a very good sense of where these different risks, hallucination and such, come from. That is why we emphasise education and understanding the known—I am getting confused with those—limitations of what we have. There will always be unknowns, and part of the job of researchers is to uncover those and discover them.

Lord Young of Norwood Green: It is about risk assessment, is it?

Professor Phil Blunsom: Risk assessment is key. There is a lot of that going on, whether we are grounding, as we have talked about, the existing models we have and what we know about their boundaries and such or considering possible future developments. You have in the commercial sphere the focus on those products, how they behave and what their limitations are, and then a big focus in academia and research on the longer-term picture and the hypothetical places that we can go to.

The Chair: Thank you. We will move on to Baroness Healy.

Q26 **Baroness Healy of Primrose Hill:** To return to risk, I will ask Professor Russell first: how confident are you that these risks are sufficiently well understood and accounted for by government, industry and, perhaps, the public?

Professor Stuart Russell: Not very. Let me expand on that a little bit. There is now a very high level of government attention to these questions and there has been a dramatic change since March of this year. We can look back to the open letter from FLI warning that there should be some kind of pause in the development of models more powerful than GPT-4 as a wake-up call, to which Governments and international organisations responded very quickly.

All that is good and makes me feel reasonably optimistic, but then I have been sitting for many days this year in long discussions at the OECD with various national delegations around how we should define AI. It is a one-sentence definition and it matters because the European Union Artificial Intelligence Act will take its definition from the OECD. The OECD definition is the one that has legal status because all the member countries have already agreed to those documents.

The definition as it is given requires that for a system to be an AI system there have to be "given, human-defined objectives", which rules out ChatGPT. According to the original OECD definition, ChatGPT is not an AI system in any way, shape or form. A number of countries are insisting that it is an AI system only if it learns once it is deployed, which also rules out ChatGPT. In fact, it rules out about 99% of all AI systems in existence. A few—some speech recognition systems, for example—do learn to adapt to the individual speaker, but the vast majority of AI systems do not learn after deployment because for people who tried it—for example, Microsoft tried it with the Tay bot—it was a complete

disaster. We build the systems using learning and then we turn off the learning when we deploy them.

This experience has made me less optimistic because I think that the national delegations probably lack technical expertise. This is a pretty uniform problem. Even in the US, where plenty of technical expertise is available, oddly enough the National Artificial Intelligence Advisory Committee does not have any active AI researchers on it. I cannot really explain why. There needs to be much greater attention paid to ensuring that we have experts who can be in day-to-day advisory roles. Having hearings with us is good, but there need to be day-to-day advisory roles in developing regulations and legislation. There is a move towards national AI regulatory bodies, just as we have for nuclear power and aviation, and then co-ordinating international bodies. I think the sooner that happens the better.

Baroness Healy of Primrose Hill: Mr Anley, your organisation has talked about better horizon scanning as one way of helping to deal with this. Could you explain that a bit more to the committee?

Chris Anley: I was referring to the threat intelligence research that we carry out. In essence, we continually monitor new and emerging types of cyberattack and trends among attackers just to help defend our customers so we can pre-emptively put defences in place and so on.

In terms of the security picture in AI in general, both Professors Russell and Blunsom mentioned the security vulnerabilities inherent in the models themselves and it is probably worth talking about those. One of the major risks in large language models is that if the models are trained on sensitive data they can leak that sensitive data. There are specific attacks that achieve this and this is well studied in the academic community. Professor Blunsom also pointed out the possibility that you can trick a model into returning an incorrect answer. We talked about whether models should be used in decision support or decision-making. Of course, in the security field, quite often a security decision is made by the model: is this a given person's face or not? Should this person be permitted to do this action or not?

That is a field that we have researched and it is important that it forms one of the areas in which external third party validation, transparency of these processes and some level of benchmarking of performance on these safety and security matters are important.

Baroness Healy of Primrose Hill: Thank you very much. Professor Blunsom, you mentioned the health implications if it is used. That is something that the UK Government need to be very aware of. Could you explain what academic work is going on to counter the dangers if we use AI more in the health service in terms of people's records and making decisions about medicines and so on?

Professor Phil Blunsom: There is the broader space of AI and then specifically large language models. The most obvious issue in health is anywhere that someone is providing, on information coming solely from a large language model, a decision about treatment or anything like that.

As with any situation like that: you should act only on trusted information from a reliable source, and these models are not reliable sources when it comes to medical treatments, medication or things like that. That would be a safety-critical space.

In the broader space of AI, anywhere you make decisions about treatments, prioritisation of patients and all these things is a place you could apply an artificial intelligence system, but you would have to have stringent regulation and testing of many different scenarios to convince yourself that it was safe in that situation, to a much higher standard than in another situation—in online commerce or these sorts of things where the implications are lower.

There is the broader space of decision-making, which I think in the AI space is well known. There have been plenty of case studies. There have been situations where people have got it wrong and deployed systems inappropriately and they are reasonably well known. However, there definitely needs to be the awareness at the regulation level of how to address those areas where there needs to be attention as to what is safety critical and what is not.

Baroness Healy of Primrose Hill: Lyric, you are an expert in misinformation. Are there any early warning indicators for major risks relating to foundation models and, if so, what would these look like?

Lyric Jain: Unfortunately, I think that we can predict a lot of what is likely to happen over the next few quarters. As Professor Russell mentioned earlier, a lot of those models that are out there in the public domain to use currently can be tricked, can be fooled, to create content and to create various forms of misinformation and disinformation. We have recently released some research around a random sample of disinformation being targeted at those models that are supposed to have some of these safeguards and safety features: 85% of prompts we hit them with were passed through and could create disinformation.

That is the state of play. Although some safety features are built into these consumer-focused models, they are not very effective. We are likely to have elections around the world next year. We are preparing actively to figure out how we can red team/blue team and come up with safeguards ahead of various election events in the UK, US, EU and many democracies around the world. However, generally, there is likely to be the same sets of narratives, the same sets of campaigns, that we have seen over the course of the last few years. There is likely to be a set of narratives pre-election period that will dissuade people from participation outright on the day of the election. There will probably be a number of narratives that are hyperfocused, hyperlocal, that are probably targeted to individuals to say, “Hey, the election is cancelled” or, “Your polling booth has been moved from this location to that location”. Probably on the day of the result or immediately after the result there will be a number of campaigns targeting the narratives around confidence in the vote, how the vote potentially was stolen, hacked, undermined and so on.

This is now a playbook—that one surface area of elections that I mentioned. The same can manifest itself in a national security setting or in another setting. The real challenge here is the volume and scale at which these campaigns can now be A/B tested. That used to be the most expensive, most labour-intensive part of these campaigns. Again—to use some of the examples from a few years ago of the $10 million or $20 million number I mentioned previously—around 60% to 80% of that capital was invested in early-stage content creation, messaging, testing and A/B testing of which audiences are responding to which campaigns. Now, all that, the hundreds and thousands of examples of messaging that need to be created, can be automated.

Those are some of the risks that lie ahead for us. The silver lining is that there is some preparedness. Work is being done around the world through organisations such as ours, platforms and various Governments to get ahead of this curve, but there still will be things we are surprised by. One key concerns is: is sufficient capital and investment being deployed by some state actors to produce, say, something like an influence GPT, where you do not need to do anything; you just put in some commands around, "Create a campaign that will cause these people to distrust these people"? They perhaps could orchestrate that and deploy that campaign end to end. Again, we do not believe any such technology exists today, but 12, 15 or 18 months from now is where the industry believes something like that might come to light.

Baroness Healy of Primrose Hill: Thank you very much.

Q27 **Baroness Featherstone:** I think Professor Russell was right that that letter asking for a pause woke up Governments; ChatGPT woke up us commoners. What should the Government be doing? At the moment, we have a safety conference. We have a frontier task force. I want to know what the Government, industry and regulators should be doing to address those key risks. What would the timescale be in which we could expect to see a positive safety net or unsafety net, depending on which way they go? Perhaps I will start with you, Professor Russell.

Professor Stuart Russell: Thank you very much for the question. A number of lists of action items have already been published. FLI published one along with the open letter and there have been a number of others from people who have thought deeply about this. I mentioned standing up a regulatory agency, which is part of the European Union AI Act. They recently included a Europe-wide AI office in the language of the Act because they felt that they otherwise would lack two things. One is expertise within the decision-making apparatus and the other is a unified voice to then interact on the world level with other entities. I think that will be true for the UK.

The Chair: Do not feel you have to go through everything. It is the priority ones. What would you say, if you had a shopping list, are the ones that really need?

Professor Stuart Russell: One thing is to remind legislators that they can actually do something about the digital world. The tech lobbies have

been very effective for 50 years in warding off almost all legislation that would directly affect their business models.

Baroness Featherstone: What would you have us do?

Professor Stuart Russell: I would say number one and the easiest is a ban on the impersonation of human beings, so a disclosure regulation saying that you should always know whether you are interacting with a human or with a machine. I would also say a ban on deep fakes of certain types, and then red lines on the behaviour of AI systems, meaning—

Baroness Featherstone: Such as?

Professor Stuart Russell: Such as if your system replicates itself, if your system breaks into another computer system, if your system discloses national secrets, if your system advises terrorists on the construction of biological weapons. You can choose from a menu, but the point is that those would be disqualifying if they were exhibited during training and testing and they would result in an immediate recall if they were exhibited after deployment. Having that regulatory power would be very important because then the developers would have an obligation to figure out how to design systems that they understand, they can predict and they can control. At the moment, that is what we lack. We are, if you like, designing nuclear reactors with no idea how they work, no idea how to turn them off and no idea if they will explode.

Baroness Featherstone: You would put responsibility on the developers upstream?

Professor Stuart Russell: Absolutely, yes.

Professor Phil Blunsom: There is a lot going on in the Government in terms of regulation and thinking about this, which is great. There was, of course, the White Paper earlier in the year and the CMA's updating of that, I think, yesterday, focusing partly on where regulation is needed, identifying where those safety-critical risks are and making sure the people regulating those areas are equipped to do that—they understand, if a product is coming along using AI, the questions they need to ask and what is likely to be a problem and what is not.

As Professor Russell mentioned, one key thing is knowledge and education. That is one of the hard problems in this space. It is fine to say that all these regulators need to regulate these products, but if they do not have the expertise to do that—and this expertise is in high demand and there is a shortage—there is a question of how that works. There is a real problem for government in how they gain the expertise and the people to do this. One possible solution is some central body that is able to centralise that expertise and then provide it to different agencies. As to how we should think about the levers governments can pull around national security and the issues that such models might raise, I am definitely not an expert on the legal issues there and what could be done now and what would need to be legislated.

Yes, there are those things, and then other things that I have highlighted previously, particularly education and making sure people are aware of these systems, so that, for instance, a lot of the impersonation and deep fakes do not necessarily come as a surprise. Maybe once it has happened to a person, they will realise that this is a problem, but we want to get to it before that. Especially around elections and such, people need to realise that just because a video shows up on TikTok or somewhere, that does not mean that necessarily happened. A Government can take a very active role in that.

Baroness Featherstone: Are we moving quickly enough? Should the regulators wait to see what happens or should they intervene ahead of time?

Chris Anley: On the shopping list of actions, I would say that action one would be to implement the recommendations of the White Paper, which is based on the OECD principles. I would also say a pro-innovation approach to AI regulation which gives us the ability to leverage the deep sectoral expertise of our existing regulators, who are active in this space already, while maintaining some form of central communication, convening and oversight with the agility to respond. For example, we had a paper yesterday from the Competition and Markets Authority specifically addressing foundation models and giving lists of particular tests that could be applied by third parties to provide some level of assurance specifically in that field.

On the cybersecurity and national security elements, we have NCSC and the National Crime Agency, which are also publishing heavily in this area specifically giving guidelines to third parties on how to make sure their AI systems are secure, how to validate third party AI systems that they are using, and so on. We say there is a skills gap, which is true, but it is also true to say that our regulators and security bodies are already very much across this issue. We also have the Frontier AI Taskforce to help us to plug the skills gap in the short term.

Action one: implement the White Paper, that central oversight, monitoring, making sure that we genuinely are across the risks. And then big picture: what if that is not enough for some of the risks that we are talking about longer term? At that point, the things that we need to discuss, arguably internationally, are the points at which it is necessary to shift from viewing AI as a product to viewing it as an offensive capability. There are already international treaties, export regimes and so on around more dangerous technologies and the question is at what point we should shift from regulating a product space into regulating an offensive capability specifically.

Baroness Featherstone: Lyric, do you have anything you want to add to this?

Lyric Jain: Yes, certainly. On the points specifically of both what we need to do and timing, the National Security Act is through Parliament, it is now an official Act, but specifically on the foreign interference offence it does not really comment on the use of AI to conduct foreign interference or to conduct human interference. Practically, for that Act to

be implemented by government and by Ofcom, we really need to drill down into what behaviours we mean and which behaviours qualify as foreign interference, whether that is something being conducted by AI or by humans. That level of specificity and detail currently is not present.

Government and Ofcom need to move very quickly to produce that level of detail to figure out which specific behavioural profiles would qualify as something like foreign interference. That is when platforms, organisations such as us, and Governments can go about identifying these. At the moment, everyone has a slightly different definition of what foreign interference means, so there is no practical way of putting the National Security Act into practice, or even the Online Safety Bill, which may or may not go through Parliament in the next few weeks. Making a determination of those specific behavioural profiles is important.

We have spoken about elections quite a lot. There is some movement around this idea of digital imprints and any content produced to do with elections that is promoted by candidates would carry a digital imprint, although that seems quite limiting. There are alternatives. I believe that Scotland is following that at the moment, where any election-related content would carry a digital imprint. Something like that would certainly make it a lot easier for organisations such as us but also Governments and platforms to intervene on any manipulated campaigns and disinformation campaigns around that specific risk.

Q28 **Baroness Featherstone:** Thank you. I have one last quick question. You have all spoken in varying degrees about the risks and the safety, which are obviously the questions we are asking, but do you think that that will impact innovation?

Professor Stuart Russell: It has been a mantra of industry that regulation stifles innovation, and I think they whisper it to legislators all over the world before they go to sleep. There are many examples, and you can talk to EU regulators who will say that phone service in Europe is much cheaper than it is in the US, flights are much cheaper in Europe than they are in the US, because they regulate to ensure competition. The credit card industry would not exist if it were not for regulation about disclosures so that there could be competition on interest rates and so on. Before the regulations about disclosures, it was a complete disaster, defrauding people and so on. I think that it is not "benefits *or* safety", it is "benefits *only if* safety".

Professor Phil Blunsom: As both a professor and someone who works in industry, I would agree that there are great opportunities for regulation to help innovation. There are some obvious specific examples for intellectual property, where there are genuine ambiguities in the law and resolving those would make it much easier for companies to operate and to know what the ground is. Competition was also mentioned.

Many things about AI are new. Many things also overlap with things that we are very familiar with. A very good example often to think about is what has happened in the search industry from the 1990s to today. In the 1990s, there was a great deal more diversity in search engines than

there is today, where there is, in effect, one dominant search engine monopoly in the western world. There are many similarities in what is happening in AI and the question is: if we were to do that again, would we want to do things differently? Would we want a different regulatory space? The obvious other consequence of that is that there is no large search engine of significance at all in the UK or in Europe. This is not without historical precedent and there are things we can look at and say: if we had our chance again—a similar one is social media—what would we do differently?

The Chair: I have one final question for clarification. In terms of this last set of questions, I was not sure whether—and maybe you were all offering different perspectives—you still think there should be a national body in addition to existing regulators. The existing regulators clearly are responsible for their own sectors, as it were. If there was to be a national body, would this be a national body that is not necessarily a regulator and is about AI more broadly? I wondered whether anybody could give me a quick clarification of what you were trying to say. Maybe I should come to Professor Russell because I think that he said something about national bodies.

Professor Stuart Russell: Yes. I think that it should be a national body with devolved regulatory powers. You cannot keep coming back to the legislative body to keep up with changes in technology. The European Union AI Act—

The Chair: But stand-alone from the other existing regulators?

Professor Stuart Russell: Stand-alone, yes. I think that Chris mentioned this convening activity. Co-ordinating among different sectoral agencies could also be useful. If we take things like bias, that is an issue across many sectors. It is an issue in health. It is an issue in finance. It is an issue in employment. It would be very helpful, I think, to have a clearing house where you could look at these issues. We are talking about intelligence here. Intelligence is not confined to any one industry sector, so it needs to be general.

The Chair: Does anybody on the panel have a differing view to that?

Professor Phil Blunsom: I would not say I had a differing view. I probably would not take a view on whether it has to be a regulatory body, but I do think that there needs to be a national capability in terms of AI knowledge and either the ability to regulate or the ability to advise on regulation.

Lyric Jain: Probably in a similar vein, I think a national body as a capability and as a capacity would be helpful. On some of these fundamental problems that are cross-sectoral, it would play an important role. A lot of the devil lies in the sector-specific detail, in our view, so that is where ensuring that sufficient authority is devolved to sector-specific regulators to go after problems that are likely limited to their sector or perhaps overlap with only one or two sectors probably gives us more speed of execution. A central body working on every sector is likely to be where we are iterating two things over years.

The Chair: We will come to the fundamental topic of regulation at a later hearing. There are different views as to regulating technology as opposed to sector specific, and there is also the question of whether we are regulating for outcomes in terms of the safety measures that we have talked about earlier as opposed to other things. This is a topic we will return to.

For now, gentlemen, thank you, all four of you, very much for your evidence this afternoon and giving up so much of your time. We are very grateful to you. I will bring this to a close now.