HOUSE OF LORDS

Communications and Digital Committee

Uncorrected oral evidence: Large language models

Tuesday 7 November 2023

2.25 pm

Watch the meeting

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

Evidence Session No. 7 Heard in Public Questions 51 - 63

Witnesses

I: Dan Conway, Chief Executive Officer, Publishers Association; Arnav Joshi, Senior Associate, Clifford Chance; Richard Mollet, Head of European Government Affairs, RELX; Dr Hayleigh Bosher, Associate Dean and Reader in Intellectual Property Law, Brunel Law School.

USE OF THE TRANSCRIPT

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

Examination of witnesses

Dan Conway, Arnav Joshi, Richard Mollet and Dr Hayleigh Bosher.

Q51 **The Chair:** This is the Communications and Digital Committee. We are continuing our inquiry into large language models. We have one panel of witnesses this afternoon. Our focus will be on copyright. I will start by asking our four witnesses briefly to introduce themselves and the organisations that they are here to represent.

Dan Conway: Good afternoon. I am chief executive of the Publishers Association.

Dr Hayleigh Bosher: Hello. I am a reader in intellectual property law at Brunel University London.

Arnav Joshi: Hi. I am a senior lawyer at the law firm Clifford Chance. I have had similar positions at other firms here and in Asia over the last decade or so.

Richard Mollet: Good afternoon. I am head of European government affairs with RELX plc.

The Chair: Thank you. You are all very welcome. Thank you all for being here today. I would be grateful if you could move closer to the microphones. The audio did not sound too brilliant. I am sure that everybody tuning in at home can hear you loud and clear, but not all of us here have as good hearing as we need.

As I say, we are going to examine copyright. I am sure you will bring us a range of perspectives. I hope that in the course of this discussion we will get to hear some practical options and ways forward for dealing with this rather difficult topic. Without further ado, I hand over to Lord Foster to get us going.

Q52 **Lord Foster of Bath:** Thank you very much, Chair. Thank you all for coming. You are welcome. I think we all know the vital importance of intellectual property to creative industries, creators and so on. Many of them firmly believe that, if companies developing AI models are going to use the stuff that they have created as part of developing a model, the intellectual property law should apply, and people should get licences and should pay. We know that some people argue differently.

We talked to some of the big companies in an earlier session. Interestingly, Google said: "We always seek to be compliant with IP laws when training our models". Amazon said that it agreed with that statement. Meta said, "We obviously agree on compliance". The question is: what did they mean by compliance? Given that we will have the opportunity with my colleagues to go into details about what we can do moving forward, could you begin by outlining your understanding of the current situation and whether people should or should not be paying? Can we start with you, Dan, since you have a lot of creators in membership?

Dan Conway: Thank you very much. I would be delighted. The first thing to say is that large language models and technological advancements in this area are a force for good and are hugely exciting. The creative industries will be innovating alongside that technology, so I do not want to position us as being antithetical to that sort of innovation, but the truth is that current market conditions mean that AI is not being developed in a safe, responsible, reliable and ethical way. That is because large language models are infringing copyrighted content on an absolutely massive scale.

We know this in the publishing industry because of the existence of something called the Books3 database, a database of 120,000 pirated book titles that we know have been ingested by large language models. We also know, because of the outputs of the models—what is coming out of the other end of the processes—that the ingested content has to be published book content. So we know that the content is being ingested on an absolutely massive scale by large language models, and they are not currently licensing them. You quoted them saying that they were being compliant with IP law. Respectfully, they are not currently compliant with IP law.

We have had conversations with technical experts about the processes undergone by these large language models. It is our contention to the committee that these large language models infringe copyright at multiple parts of the process: when they collect the information, how they store the information and how they handle it. It is our contention that copyright law is being broken on a massive scale.

Dr Hayleigh Bosher: From the copyright perspective, on principle, copyright is about granting creators, who you mentioned, rights in their work. Generally speaking, you ask for their permission and seek a licence when you want to use it. Copyright is very much a technological and cultural tool that needs to be applied in different circumstances. We try to write copyright law in such a way that it is technologically neutral, to the degree that it lasts a long time, even when technologies evolve.

The principle of when you need a licence and when you do not is clear. To make a reproduction of a copyright protected work without permission would require a licence or would otherwise be infringement. I think that is what AI does at different steps in the process: the ingestion, the running of the program, and potentially even the output.

As you mentioned, we are in the position where some AI tech developers are arguing a different interpretation of the law. I do not represent either of those sides. I am a copyright expert. From my position, understanding what copyright is supposed to achieve and how it achieves it, you would require a licence for that activity.

Lord Foster of Bath: So you are entirely agreeing with Mr Conway that what is currently happening is in breach of your understanding of copyright law, but you acknowledge that other people have a different interpretation of it.

Dr Hayleigh Bosher: That is right. There are some legal AI technologies, of course, where they have sought a licence. I am talking about those where they have not.

Arnav Joshi: Given my area of expertise, which is more on digital regulation and data protection matters, I would rather defer to Richard on that one. I am happy to cover privacy aspects, if you would like me to.

Lord Foster of Bath: Richard, it is good to see you again.

Richard Mollet: Lord Foster, it is good to see you. First, I should say that RELX approaches the issue of copyright and AI from both perspectives. RELX is, in fact, four different businesses. We have Elsevier, which is a world-leading publisher of academic research. LexisNexis is a global publisher of legal information. We have a risk-solutions business that applies data analytics machine learning to data across a range of industries, including financial services, and we have an RX and exhibitions business. In all those areas we are taking data and content, some of it proprietary—we are a rights holder—aggregating it with other data and then applying analytic tools, including AI and lately generative AI, to help our business customers to improve their decision-making.

From both those perspectives, however, we agree with what has been said. I should say, for clarity, that Elsevier is a member of the Publishers Association. From a copyright point of view, we think it is vital that there is transparency about what goes into the models, not only so that creators can be rewarded and credited, give their consent and get compensation, but to incentivise the creation of high-quality data. Unless we can trade off intellectual property rights, there is no incentive in the long run for companies to ensure that data is of the highest possible quality and that it is peer reviewed and authenticated.

In the two areas where we work—scientific research and legal research—it is vital that we have that quality. Copyright is important for that reason. For any AI developer, the cliché of "garbage in, garbage out" is never more apposite than in the world of generative AI. Unless we can see what is going in, including protected works, we cannot have trust in the outputs. For both those reasons and from both sides of our business, as it were, we think that copyright should be upheld. I certainly agree with Hayleigh and Dan that there is strong evidence that hitherto it has not always been.

Lord Foster of Bath: I read out the comment from Google and the others saying, "We always seek to be compliant with IP laws when training our models". How can they be saying that? Do they have a totally different interpretation of IP?

Richard Mollet: They are operating in a different jurisdiction, of course. I am in no way an expert on US copyright law and fair use, but some people maintain, and our council says erroneously, that US law allows this. The UK law, on the other hand—and, indeed, EU law—is pretty clear: if you are reproducing works for the purposes of text and data mining, and you are a commercial entity, you have to have the permission of the rights holder.

If you do not have that permission in the UK, it is an infringement. As I say, they might be saying that from a different jurisdictional perspective, but I do not know.

Lord Foster of Bath: Thank you. I know that my colleagues want to look at how we can move forward in more detail, so we will leave it there.

The Chair: Can I check something with you, Dr Bosher? You said that some AI developers have complied with the law, but some of them have not sought a licence and are therefore, in your view, non-compliant. Are you able to give us an example of the latter?

Dr Hayleigh Bosher: The only one I know is Musiio, because I have spoken to the person who founded it. There is no transparency, as has been picked up already, so there are a lot where we just do not know what they are ingesting. If there are licences, they would be done privately, so we would not know that either. I just know that one, because I have spoken to the founder.

My point is that it is possible. What is interesting about the Google example is that, for instance, an AI bot can give you lyrics to a song. Google has a lyric system that it currently licenses from the music industry. There, it is licensing with the AI chatbot that is currently unlicensed to do that. There is a parallel. I think that is an interesting comparison.

My main point is just that it can be done. I do not want to say that all AI is infringing, because we do not know. Where there are licences, maybe they will not be infringing because they have sought permission.

Q53 **Lord Kamall:** It is interesting that there is clearly consensus between the three witnesses who have spoken. I am trying to think of this from first principles. If you think about machine learning, clearly you want as much data in the dataset for the model to be accurate and give accurate output. In this case, I can see, although once again it might be too theoretical, where copyright might be a barrier to having access to some of that data or information. Therefore, in your view, what is the purpose of copyright? Is it there purely to protect rights holders? Is it to encourage innovation, or can it do both? I will ask you first, Dr Bosher.

Dr Hayleigh Bosher: I am really keen to answer. There are a couple of things. There is no evidence to suggest that copyright is a barrier to innovation in that context. We can see that, for example, with the UK IPO consultation on the text and data mining exception. Lots of tech firms did not opt for the broadest option there, which demonstrates that it is not a barrier to them, and that they understand that the tech industry and the creative industry are not separate entities. Creative industries are also developing AI. AI tech companies are also creative. Everyone can benefit from the copyright framework, so I do not believe that it is a barrier.

The purpose of copyright is to encourage creativity and innovation, and the dissemination of that creativity and innovation for culture and knowledge. It does that by balancing the protection of the creator's output with limitations, such as exceptions or the length of copyright.

Lord Kamall: Does anybody else want to add anything?

Richard Mollet: I agree with Hayleigh. Not only is copyright not a barrier to innovation; it is indeed a driver. Certainly, for companies such as ours, we can only invest, as I said earlier, if we can trade off intellectual property rights. The innovation that we have done, working in AI for the last 20 years, is largely because copyright is one of the drivers of our business.

The Chair: We will come on to innovation and the trade between innovation and respecting rights holders later. Perhaps we can pick some of that up when we get a bit further along. Before we move to a completely different topic, Baroness Harding wants to ask a supplementary.

Q54 **Baroness Harding of Winscombe:** I have quite a nerdy question. In previous evidence—I will paraphrase as a layman on copyright—we heard that, thinking of copyright as establishing a distinction between reading something versus copying it, the challenge with large language models is that it is not clear whether they are reading or copying. That is the nub of the argument that you have just been rehearsing.

The nerdy question, starting with Mr Conway, is: could you elaborate in more detail on what it is about the collecting, storing and handling of copyrighted materials that constitutes copying rather than reading, in your view?

Dan Conway: Thank you for the question. I am aware that I am sitting next to Dr Bosher, who might be able to unpack this in a more expert way than me. Copyright is a bundle of exclusive rights. If you are a creator or an author and have written your book, or a publisher who has taken on those exclusive rights, you can control the way in which that is used, including reproduction and communication to the public.

Part of what I was saying earlier about the actual technical process by which these tech companies produce these large language models—by the way, a consideration would be whether they should be transparent about those processes, so that we can be absolutely sure about this—is that copies have to be made from a technical perspective in order to process that data. Therefore, from an input perspective, that would trigger copyright. It is a copyrighted act.

One of the arguments that might be employed on the other side is that the temporary copying exception to copyright comes into play, and developers might rely on that exception as a way of not needing to seek a licence. Again, it is our contention that that temporary copying exception, even if the copying is temporary in this context and they get rid of the content afterwards, does not fully come into play because it is a stipulation for that exception that the original work should not have any economic value. I am sitting here representing a £7 billion industry that would argue otherwise.

Baroness Harding of Winscombe: Dr Bosher, could you elaborate?

Dr Hayleigh Bosher: First, I find metaphors really unhelpful in copyright, because they give us a whole lot of images in our mind that are not accurate to the thing that we are trying to decipher. When you say "reading", you are thinking of when you read something. You look at it and you might remember it. When the AI reads it, it is reproducing it, typically for commercial purposes.

The way to consider that in the copyright context is to think about the point in copyright. If the point in copyright is to remunerate the creator in order to encourage creativity for the benefit of society, you have to take a purposeful approach. Is the purpose of your reading a book to benefit commercially from the story within it? No, you are just enjoying and consuming the story. The purpose of AI reading a huge dataset of information of value that is owned by copyright, however, is in order to create a new business and be remunerated themselves from something created by someone else that is typically a licensed model. That is the distinction in the reading.

At times, copyright mentions technology. As I mentioned earlier, we try not to be too technologically specific for longevity, but also because it is not the point. It does not matter how you do it; it is why you are doing it.

Baroness Harding of Winscombe: That is hugely helpful. Thank you.

The Chair: It is really helpful. Thank you very much.

Q55 **Baroness Healy of Primrose Hill:** My question is directed to Mr Joshi. Please could you set out your view on the arguments about privacy and personal data in relation to large language models? I am particularly interested in your view as to whether it is likely that LLMs breach UK privacy law as it stands.

Arnav Joshi: Thank you for the question. Having spent most of my time in private practice, but a little bit in academia, I like to think I have a reasonably objective view on the question.

In my experience, companies are genuinely looking to comply with data protection law in this space and not to jump the gun or look for loopholes. An overwhelming majority of my clients are quite careful with how they adopt generative AI. That is particularly true for industries like financial services, which, given where we are sitting, in London, is an incredibly important industry.

What they need is a lot more guidance. They are looking for specificity. That is something this inquiry is looking to do, which is very helpful. There has been a huge amount of hype surrounding AI, particularly in recent weeks. Some of that hype, particularly on questions like existential risk, is setting back the debate on issues such as data protection and copyright a little bit.

I want to share some findings from a recent report that I think came out just last week. It is an Ipsos study done by a team that I know quite well and respect. It said that 9% of people in this country are currently using generative AI, whether at work or for personal use, and 19% think that it will lead to our extinction. Issues like that about existential risk and a lack of understanding by rights holders is a significant gap that needs to be addressed.

Coming back to data protection, is there wide-ranging non-compliance with data protection law as things stand? No. Do I think there is a risk that that might happen if sufficient guardrails are not kept in place? Yes. I think the risk exists, but I have not seen instances of wide-ranging non-compliance, given where we are.

In order to understand what non-compliance might look like, it may help to recap very quickly what good compliance looks like under GDPR. As you probably know from other experts and from briefings, GDPR is a technology-neutral law, which means that it applies to AI and has applied to many other forms of data processing before it and will continue to apply to others that come after it. It presents a sliding scale of risk-adjusted compliance requirements and decisions, so if you are a utility company trying to understand what usage projections might look like based on user patterns—some of that might be personal data—you still have to comply with the same core principles under GDPR as you would have to do while building a large language model for customer services. These things become more complex as you try to start to process high-dimensional data and high-volume data, and in slightly trickier use cases such as healthcare.

What are the tests that need to be met under GDPR to make sure that you are in compliance? The first big one is known as the purpose and means test. You need to understand, as you would under IP law for instance, what purpose you would like to process the data for and how you will go about doing it. That is a test that needs to be done at every significant step of processing. Sometimes it can be the same at the development stage and the deployment stage, but in many instances, particularly with large language models, it can defer, because these are general purpose applications, as we know; they can be built for one purpose but then used for another.

Data protection assessments need to be done, as I have just mentioned, at every step. That is an important piece. The other thing that has to be kept in mind is applying a lawful basis. GDPR offers data controllers—the entity doing the purpose and means decision-making—six lawful bases. I will focus on the one called legitimate interest because, first, it is the most widely used and, secondly, it is arguably the most controversial. It requires you to identify why it is in your legitimate interest as a company to have access to, and to process, personal data, ostensibly on billions of different individuals, and how you want to do the processing. That is training your model, pre-processing and then putting it on the market. How is it necessary to achieve that legitimate interest? You need to balance it against those individuals' rights and the risk to their rights and freedoms.

The test has to be done in a systemic way. There is GDPR guidance. As far as I am aware, there is no extensive guidance on how you would apply the test to large language models, which is where some of the risks start to come in.

The ways in which legitimate interest tests have been done, as far as I am aware, have been scrutinised by a number of regulators, not in this country, but we still largely apply the UK GDPR. They have been stress-tested in Italy, France and Germany. With a few rare exceptions, it looks as though most large organisations putting large language models on the market have been able to justify their legitimate interest to regulators when scrutinised. As an example, even though there is some controversy about how a large corporation can argue that hoovering up billions of data points is in their legitimate interest and does not override the users' interest, most regulators at the moment seem to be satisfied with their justifications.

Lastly, one of the things that need to be kept in mind is that companies need to conduct data protection impact assessments. This is not that different from what we see in the AI Act in the EU, which will require conformity assessments of exactly how you will comply with the AI Act. Once again, it is a step-by-step assessment of how you look at each of the measures that I have just mentioned, and a number of others, and then put things on a sliding scale of risk. One of the things that I mentioned before, which Dan may have circulated to you, is the really helpful risk matrix I have here. It is from the Information Commissioner's Office and its guidance on how data protection impact assessments should be done.

There are two axes on the matrix: severity of impact, and the likelihood of harm. As you will see, it is only as you start to get to the top-right quadrant where serious harm is likely. The likelihood of harm is more likely than not that you start to get into a high-risk category. When you are high risk, at least under GDPR, that it triggers a mandatory consultation with the Information Commissioner's Office. You need to sit down with them, show them all your assessments and say, "Here's how I think I can mitigate that risk". If you cannot mitigate it, your product cannot be on the market. Unless you are at that top-right quadrant, most of GDPR compliance essentially allows you to mark your own homework, unless the regulator steps in and tests it.

What we see happening in the market at the moment is that most people, although they are marking their own homework, when stress-tested seem to be able to come out on the right side and prove that they are compliant. That is not to say that there have not been instances where it has been problematic. The last example was just two weeks ago. The ICO has issued preliminary enforcement proceedings for a major technology company. In that instance, we do not have a lot of public information available about it just yet, but I think the ICO asked for the data protection impact assessment, scrutinised it, and found that it was not up to scratch.

Q56 **Baroness Healy of Primrose Hill:** That is very thorough, thank you. I am also concerned about the inadvertent leaking of private information because of the way the large language models take up all this information. Does the individual have any right to complain? Once it is out there, it is impossible to get it back. What is your view on that? I might then ask another member of the panel.

Arnav Joshi: I can start and then I am sure the others will have something to say.

The Chair: If you were able to be a little briefer in your answer, that would be very helpful, not least because there is a lot of information.

Arnav Joshi: It is an occupational hazard of being a lawyer, I am afraid.

The Chair: We will be reading the transcript of this very carefully.

Arnav Joshi: Sure. Coming back to specific rights, GDPR, including over AI, gives you a number of very helpful rights. There is the right of access, so you can get a copy of the data that someone might have on you. There is the right of erasure and the right of rectification. All of these are actively in use for large language models as well.

Some technical solutions already exist to the question: at which point in the large language model processing activities can you exercise those rights? Arguably the easiest—I caveat that by saying that it can end up being a very manual process—is to exercise that right over training data. At the point at which it is being hoovered up, whether that is billions of Wikipedia pages or a licensed dataset, you can go to the data controller of the company building the model and say, "I would like to exercise my right of erasure and here's why. I have the right under GDPR". They will then have to undertake a mix of a very manual and somewhat automated process to help fulfil that right.

The guidance from regulators for the actual trained models is that there is so much pre-processing of the data that there is very little meaningful personal data in the actual model. If the controller can justify that by saying that it is highly unlikely that any personal data still subsists in the model itself, you can get away with not having to fulfil that right for the model itself. There are some ongoing debates about whether models constitute personal data. There is no definitive guidance on it just yet.

Lastly, coming back to how you may end up seeing the information yourself, one of the things about large language models is that they are a predictive technology. It is not a search engine that is giving you factual information. It is coming up with and predicting what the next word might be. The easiest solution to that—for rectification and for your personal data not to be misused—is inputs and output filters. These are being used by many of the major technology companies putting these services on the market already. That will mean that you complain through the slightly tedious and slightly manual process, and they will introduce a rules-based check into the system that will say, "Arnav Joshi has exercised his right. His responses in relation to his personal data should not be given out by the system".

Baroness Healy of Primrose Hill: Mr Mollet, do you have anything to add to that, please?

Richard Mollet: Only to say that RELX is a company driven by the use of data, and much of it is personal information. Of course, everything that

Arnav says is what we fastidiously comply with. As I have said, we have been developing AI systems for some time and we published our responsible AI principles last year, one of which, of the five, is on privacy and data governance. It is baked into every stage of our AI development process, for all the reasons that have been stated.

Baroness Healy of Primrose Hill: Thank you very much.

Q57 **The Lord Bishop of Leeds:** I am not a technologist, so I do not quite get all the detail, but if the models are basically being trained on data, some of which will be wrong, what form of redress is there in the system if that is propagated by these large language models? Presumably, you cannot correct it. If you were to make a complaint about it, what redress is there? It would need to be granular.

The Chair: Who are you directing that question at?

The Lord Bishop of Leeds: Arnav, first.

Arnav Joshi: I can go first.

The Chair: Keep it tight, though. We do not need a lesson in the law. Just tell us your answer.

Arnav Joshi: Yes, it can be exercised, and the data can be corrected. It is very hard to understand what is going on in the model itself; we have heard the references to black boxes. With the training data itself, you can identify people, because those are largely searchable databases. Not a lot of it is what is known as unstructured data. They are searchable systems. You can identify whether someone's personal data is in there. If they have an output that says, "This was clearly incorrect, so please can you do something about it, big tech company?", they will be able to search through that dataset and say, "All right, we're taking you out of it and the next time this model is trained your data won't be in it any more".

The problem is that these models are normally only trained on a six-month or a one-year cycle, because it takes hundreds of millions of dollars to train them. The short-term solution is upward filters, which I mentioned. Before the model checks out an output for someone to view publicly, it will apply almost a blunt force weapon that says, "I'm going to check this output before it is available for people to view". Although the model is thinking it, no one will really see the output. Those solutions are already in use.

The Chair: Are you any the wiser, Lord Bishop?

The Lord Bishop of Leeds: I have a lot of follow-on questions, but not for now.

The Chair: All right. It looks as though Mr Mollet might want to offer a simple answer.

Richard Mollet: It would be also very difficult to know what falsehood looked like. If I went on to a large language model now and said, "Tell me about Richard Mollet", and it came back and said, "He's six feet four", I would say, “Well, that’s wrong”. It is not because anyone ever told a large language model that I was six feet four; it is because that is the sort of thing that large language models do; they make stuff up. It would be really hard to know when the wrong data about you is being put in or whether it is just something that has come out, which is one of the difficulties.

The Lord Bishop of Leeds: Part of the reason for asking is that I got into big trouble when my Wikipedia entry was manipulated by my youngest son’s mates, and it would not allow me to correct it. I still get people accusing me of stuff that is inaccurate. With that sort of stuff, if you multiply that—I am just an individual; I am not a company—you can see the damage it can cause. Your chance of redress is very limited. I hear what Mr Joshi is saying, but I am not convinced.

Arnav Joshi: I do not think we have—. Sorry, Hayleigh, would you like to come in?

Dr Hayleigh Bosher: I just want to add that I agree. First, what you are saying is that it can be removed but only from the next model, not from the current model. There are things that you can block. If you put in a prompt for something, they can block certain outputs, which is a good start, but you cannot actually remove your data from the current model. Even if you did that, it does not prevent misinformation. I did an interview about this for the *Guardian*. We put in, “Tell us something about Hayleigh Bosher”, and it said that I went to Harvard University, which I did not, and it does not say that anywhere on the internet. It has deduced that from the fact that I am a legal scholar. It does not prevent misinformation even if you take the data out in the first place.

In terms of redress, the only thing to add is that I know of a case happening in Australia where someone is suing for defamation because false information has been put out, but I would not necessarily say that that is an effective mechanism. It is just an individual trying to protect themselves in a similar situation to yours.

The Lord Bishop of Leeds: And has the resource to do it.

Arnav Joshi: Speaking of halfway-house helpful solutions, one other that many of the companies are certainly coming up with and working through with regulators is greater transparency. With a lot of these models, the services now say on the tin, “This is an experiment. This is not a factual statement”. When you sign up, it will give you a little text box that says, “Please don’t treat this as fact. This is just predicting a bunch of text. It is not a search engine”. Things like that are coming online. Whether people are believing them is a different question, because everyone looks at it and is quite excited by it.

Lord Kamall: Richard, you talked about large language models making stuff up. You might want to say, “inaccurately deduced”. Are large language models really doing that? Are they making stuff up or inaccurately deducing, as opposed to saying, “We don’t know this”, or avoiding an inaccurate statement?

Richard Mollet: Absolutely. May I give you one example from my world of scientific research? A year ago, when ChatGPT launched, a few weeks and months afterwards we at Elsevier were getting requests from researchers asking to see papers that they had read about on ChatGPT and were frustrated because they could not access them on our platform. It turns out that they could not access them because the papers did not exist; ChatGPT had hallucinated scientific citations.

Lord Kamall: That is a really important point. Thank you.

The Chair: Mr Joshi, from all that you have told us, am I to take it that it is your view that the large language models currently are not breaching privacy law?

Arnav Joshi: There is a sliding scale of what GDPR breach would look like. In the same way that there is a sliding scale of risks, there is a scale of what non-compliance may look like. At a very technical level, the fact that someone had a response that spit out the fact that Richard is six feet four when he is not is arguably a breach in some form. There is, in a sense, a de minimis that a regulator will think about. It will look at what level of harm needs to materialise before something can be considered directly actionable at various levels.

At the lowest level, if you have some incorrect information and you think that someone is not processing your personal data in the right way, you go to that company and say, "Please can you correct this?" It will then have 30 days, three months or whatever its timeline is to correct it and bring itself back into compliance. One level up, you will need to consult with regulators. A regulator will say, "You've done this. I'm probably not going to fine you if you fix it in the next three months", because there is a potential breach, but it is not serious enough that it cannot be remedied. At the highest level, the regulator will step in, as it did in Italy, and block your service from operating altogether.

There is a sliding scale of compliance. Where I do not think there is a lot happening is at the very top-right quadrant: very high non-compliance, very high risk, and very severe non-compliance.

The Chair: Do you think that is a design issue of the technology or an issue with the legislation not being fit for the technology as it develops?

Arnav Joshi: I agree with the ICO view on this. GDPR is an incredibly powerful tool to address things like large language models. To the extent that they process personal data, they can process lots of other non-personal data as well. I do not think there is a problem with the law. Given that the technology is new, everyone is trying to innovate, and it is highly scalable, there are teething issues associated with how the technology is being built up. The law allows some degree of iteration and mitigation in consultation with regulators, which is happening, and that is essentially where we are at the moment. It will take us some time to make sure that compliance happens at a greater degree. It will never be perfect. It never is with any law, but the highest-risk incidents can be avoided.

Going back to my earlier point, we think of big chatbots and things that are out there for use, but if we look at high-risk settings, things that are in defence or in the healthcare sector, even some forms of financial services-related customer services, people are extremely careful about where they get their data from and what they do with it. Most companies are still in trial mode; you cannot access the systems they are building.

The Chair: Okay, thank you.

Q58 **Baroness Harding of Winscombe:** I would like to bring us back to copyright to begin with. A number of you have been very clear that you do think that copyright laws are being broken today, so I would like to ask quite an open question. What are the options for ensuring that large language models do not break copyright law and that they reward rights holders for use?

Dan Conway: We think it is quite clear what should be happening. There should be a process of permission, transparency, remuneration and attribution. If you are operating a large language model, that would mean effectively that you seek permission to ingest content prior to doing that. You are then transparent about what you have ingested, both retrospectively and moving forwards, and from that you go through a process of remuneration of the creator or the rights holder and attribution for the citations that Richard talked about—making sure that there is proper attribution in the output process.

That is really about licensing. We need market-based solutions for licensing that are as seamless as possible and flexible, and can make sure that the access point from AI systems to data is done in the best possible way. That is probably a combination of direct licensing between a rights holder with a large proprietary dataset of clinical information, for example—the *Harry Potter* catalogue; pick a set of rights—and a collective licensing model that might be helpful, for smaller businesses particularly, both on the AI development side and on the rights holder side. If you do not have the capacity to do those commercial deals in real time, you would effectively pursue a voluntary collective licensing option as well. That is not on the market yet. It has not been delivered yet, but it is where we need to get to.

That is one of the options that the IPO looked at initially in the text and data mining conversation: how do we improve the licensing system so that those deals can be done and we can ensure that the right information and the right creative works are going into the machines and we get the right outputs at the other end?

Baroness Harding of Winscombe: Thank you, that was very clear. Dr Bosher, what is your view?

Dr Hayleigh Bosher: I largely agree with Dan. I do not want to take up more time just repeating. We need licensing structures. It would help if the policymakers confirmed that the law says what it says. It seems silly that you would have to do that, but there is a need for it, given the fact that you are hearing arguments to the contrary from people who believe that US law applies here, or even that some exception would apply where I do not believe it does. Transparency is important for the purpose of enforcing those rights when they are breached.

It would not be a bad idea to consider whether we need to extend any of the rights that might need to be put into place considering the new technological developments. We are currently in the process of thinking about implementing the Beijing treaty, for instance, and there are different ways that you can do that, such as extending moral rights, which goes back to attribution, and equitable remuneration, which is about remuneration. Those could be very relevant in this context as well.

Richard Mollet: There are two things that I would add to what has been said. One would be to look at what the European Union AI Act currently says. Of course, it is not quite at the end of its legislative process, so this may change. The European Parliament introduced an amendment calling on the developers of foundation models to provide a sufficiently detailed summary of the protected works that have been used—to give statutory effect to the call for transparency. We could look for something similar in the UK. Indeed, we are talking to the Intellectual Property Office about it, perhaps not in statute but at least at a voluntary level, so that rights holders are able to see what is being used and act accordingly.

In the world of scientific publishing, we have had licensing systems for text and data mining in place for a number of years that are used by researchers and large corporates, and we have good licensing arrangements with those companies. They could use the same licence, with text and data mining being the basis for the development of AI models. So we have some of the blocks in place, but, as has been said, we need more clarity out there so that rights holders can see what is happening.

Baroness Harding of Winscombe: Mr Joshi, do you want to add anything?

Arnav Joshi: No, I am good.

Baroness Harding of Winscombe: This is probably a very dumb question, but how would you get from where we are today to the world that you have all just described, as the UK? How do you even start that negotiation and create the conditions where the parties come to the table?

Richard Mollet: Are we that far apart? There are voluntary code discussions going on. There are good actors. All the large language model developers have signed agreements with the White House about the need for safe and secure AI. Most companies that we work with are committed to safe AI, and that means transparency. I might be rather optimistic, but I do not know that we are so far apart that we cannot get some agreement on a voluntary code, get some transparency into the system and allow the licensing models, as Dan says, to really get off the ground in other areas.

Baroness Harding of Winscombe: Do you agree?

Dan Conway: I agree. The IPO has started a process of voluntary round tables with a view to a voluntary code of conduct. We support those and support the IPO in the work that it is doing. It is doing a great job with what is a very difficult topic and lots of very difficult stakeholders. The reality is that tech companies have still not acknowledged that copyright applies. Without the acknowledgement that copyright applies, that voluntary process will run aground. We still support a voluntary process, and we would like to see a high-level set of principles from the Government pretty urgently saying, "Copyright applies and transparency applies". That could be the EU version of transparency or the White House version.

The G7 Hiroshima AI process, in which the UK played a role, also talks about transparency. There are lots of global models out there already where we can pick and choose and forge our own way as the UK on what we think is best, but for that voluntary approach to apply we need the companies to acknowledge that the law currently applies. I would support the voluntary approach, but very much backed up by a legislative handbrake if the voluntary conversations fall apart.

Baroness Harding of Winscombe: Thank you. That is very clear.

Q59 **Lord Hall of Birkenhead:** This may be for Hayleigh or Dan; I am not quite sure whom. Given the vast amounts of data that are involved—it it is of a scale that we have not come across before—if I were a copyright holder, what does transparency look like, and could I cope with it?

Dan Conway: If you are an author of a book or a series of books, at the very least, at a base level, you should be able to contact the large language models and ask whether your creative works have been used to train the model, and you should get a response. That is at base level. It still places the onus on you as the author proactively to reach out. That, as you can see, scaled up to a macro set of processes, will not be the answer. At a very base level, you should be able to find out whether your information is in there. You cannot currently do that.

We encourage the committee and policymakers to think more creatively about establishing a central repository of creative works and datasets that have been ingested. The EU Act might be a way.

Lord Hall of Birkenhead: So all LLMs come together to do an ingestion bible.

Dan Conway: A searchable repository of citations and metadata that would allow a user, a rights holder, to go in and search for whether their content has been ingested without having to make a request about each individual bit of IP.

Lord Hall of Birkenhead: Something like that would be very useful to me as a copyright holder of some sort. Do you want to add to that, Richard?

Richard Mollet: Some people might tell you, "That would be too difficult, because think of the amount of data"; to which we would say, "Yes, but these large language models are already handling large amounts of data".

They might say that it is too difficult to identify which are the protected works. If you are a responsible developer of AI and you do not know what is in your system, I am not sure that you are responsible. A lot of the objections to the ideas that we in the creative and rights holder community are putting forward are weak, frankly.

Lord Hall of Birkenhead: Thank you. Hayleigh, is there anything you want to add?

Dr Hayleigh Bosher: Only that what transparency would give you is the opportunity to enforce your rights. Copyright is a private right. To some degree, the onus is on you to enforce that right, if we are not talking about criminal copyright infringement or anything like that. To a certain degree, creators are on their own in that situation, but there are also collective organisations that represent them and negotiate or even disseminate remuneration for them. It is important as an individual and as a collective to be able to enforce rights. It is not just that I want to find out if my work is being ingested for the fun of it. It is, "Then what?", so that I can enforce my rights.

Lord Hall of Birkenhead: You probably want some reward.

Dr Hayleigh Bosher: Yes, exactly.

Q60 **The Chair:** Have we, in the course of this, covered the opt-in versus opt-out issue? I do not think we have, have we?

Dr Hayleigh Bosher: Not explicitly. I think we mentioned that if you opted out—for example, with the personal data—you would only be opting out of the next model, because they cannot take out your data.

The Chair: I am thinking about the content creators versus the large language model developers. If there was a regime, would it be better for the content creators to have an opt-out, so that everything is in unless you opt out, or it is not in unless we put it in? In the context of a regime, I wondered what views there were on that stark choice.

Richard Mollet: That is the regime we have in the European Union under the digital single market directive, where, when we are dealing with commercial players as a rights holder, we have to expressly reserve our permission against text and data mining, otherwise they can do it. In that sense, there is an opt-in/opt-out regime, and it operates tolerably well. It is still early days, but it requires rights holders to have the ability to have machine-readable permissions on their content so that when a crawler comes along and says, "Can I text and data mine your work?" they encounter a bit of code that says, "No", "Yes", or "Possibly, but come over here and get a licence". Those technologies are being developed.

We can do it at platform level. You come along to a platform and everything on my platform has this permission. It is slightly harder, when you are talking about individual works that might be out in the wild on the internet, to get the technology that expresses the permissions at that level, but that is on the way. We are working on it.

Dr Hayleigh Bosher: From a copyright perspective, the opt-out is only in the context of an exception. The starting position of a copyright owner is that you need to seek permission to copy, so it would start as an opt-in. Does that make sense?

The Chair: Yes, it does. Thank you. We will move on.

Q61 **The Lord Bishop of Leeds:** Part of the reason why we have stuff in statute is that voluntary does not work. There may be a conversation there. Dr Bosher, you said that it would help if policymakers were clear. One thing that is not clear is what the balance should be between government, courts and regulators in this whole field. Could you start by saying something, possibly briefly, about the role of government in choosing between encouraging innovation and respecting rights holders?

Dr Hayleigh Bosher: There are court cases. We see the judges interpret and apply the law as it currently stands.

The Chair: I should add the usual sub judice-type thing here in Parliament on anything active.

Dr Hayleigh Bosher: Yes. Just in general, I mean. That is done in copyright on a case-by-case basis. It might give you an answer to a very specific case with specific parties, but it will not necessarily tell you the interpretation of what the law means in general, so that can be quite confusing.

For policymakers, copyright, as I mentioned at the beginning, evolves with culture and technology. It always has. When we had the internet, we had to update the law, and when we had the photocopier, we had to update the law. The principle of copyright is the same, but the context is slightly different. Sometimes that can happen more smoothly than others, depending on what the technology does.

So, in this situation, we need that confirmation from the policymakers because the technology is moving very fast. I do not think it is necessary to make a specific AI copyright and put artificial intelligent words into the legislation because the principles already apply. It is just the confirmation of it that is required. Does that answer your question?

The Lord Bishop of Leeds: Yes, that is helpful. Thank you. Dan?

Dan Conway: I would be very wary about waiting for this to be decided just by case law. Without policymakers grasping the nettle on this one, there could be litigation on it for the next decade, which would move the goalposts around continuously. As Dr Bosher just mentioned, these cases are often between a single set of rights holders and a single technology operator, and those operators will run different models. It is my view that we need policymakers not to wait for case law, but to make a really strong statement on intellectual property and the ethical use of AI and make those principles binding.

The Lord Bishop of Leeds: What about the specific question of the balance between encouraging innovation and respecting rights holders?

Dan Conway: Copyright law balances innovation and respecting rights holders. It is one of the things that it explicitly does, and that is why we have copyright exceptions. It is why we already have a text and data mining exception in UK law that allows non-commercial mining of huge datasets.

What is the kind of innovation that we want? I would go back to something Richard said earlier about garbage in, garbage out. This is not about stopping innovation. As I said at the beginning, the creative industries are not about stymieing AI innovation; we just want it to be done responsibly and in accordance with UK law.

Richard Mollet: Lord Bishop, to your point about the interaction between government, courts and regulation, on this area the UK Government are getting it right by not hastening to regulation, but setting out principles that they might underpin with statute in the longer run, and waiting to allow the market to develop.

I mentioned earlier that we had published our AI responsible principles some years ago—lots of companies in this area have done the same—because we know we cannot wait for government to tell us what to do. The EU AI Act, which is the first out of the blocks, still might not become law for another two years, even if it gets signed off in December, so we have to be working on this already. That is the tone when you see things like the Bletchley declaration and the G7 principles.

A lot of onus is being put on the private sector to do the right thing in the meantime, because, as Dan said—with apologies to the learned friend on my right—litigation is expensive. We do not want to always be going to court to find out what our rights are. It is better all round if things are clearer through our self-regulation, and then underpinned by statute where necessary.

Lord Lipsey: I want to make sure that I am clear. The Lord Bishop's question, I think, was whether government needs to choose between encouraging innovation and respecting rights holders. The more I heard your evidence, the more I thought it was nothing like as clear a choice as that. It is in both people's interests to reach an agreement. For one thing, the rights holders will all have to go to court at great expense, and the people who are using those rights will also have to go to court at great expense. There is that very simple reason alone. Is there really a contradiction? Is it not rather a question of finding an iterative way forward so that we can find a solution that suits everyone?

Richard Mollet: It is a false dichotomy that you are either innovative or you are a rights holder. Absolutely not, as Dan said. Rights holders are some of the most innovative companies in this country. The trick, as a rights holder, is how to innovate while at the same time preserving all the things we want to preserve about copyright, and allowing the technology to develop. As I said at the outset, as a company that rides both those horses, we definitely do not see the contradiction. It is something we are doing at the same time.

Lord Lipsey: That is helpful. Thank you.

Q62 **Baroness Featherstone:** Don Foster opened by saying that those great American companies are all breaking copyright. Where does that leave creators in this country whose copyright is being done over there?

Dr Hayleigh Bosher: First, I do not necessarily agree that they are compliant. They will court to argue about it in America under what they say is fair use, but it is not that straightforward. If it was so clearly obviously fair use, it would not be going to court. We will have to see how that law is interpreted and applied to see whether that is actually true. If it is true, we contend with the fact that the laws are different in different countries already, with lots of different laws, especially copyright. We have minimum international standards that help us trade with other countries and we have reciprocal remuneration agreements through our collecting societies so that our creators can benefit from the use of their work in different countries. Again, the lack of certainty is unhelpful, and that is something we have to manage in the meantime. It remains to be seen whether that is actually true or not.

Baroness Featherstone: Is there any timing on that? When will the first cases be?

Dr Hayleigh Bosher: Cases can go on for decades.

Baroness Featherstone: Dan was right in saying that you cannot wait for legislation.

Dr Hayleigh Bosher: Yes.

Baroness Featherstone: I mean for court cases.

Q63 **Lord Foster of Bath:** Can I ask a really noddy question? It follows on from the whole business of the legislative frameworks in different parts of the world. If I have an AI model that is generated in the United States but then added to for a particular purpose in the United Kingdom, which laws apply to which stages of production? You said right at the beginning that you had to check at each stage for copyright breaches and so on.

Dr Hayleigh Bosher: It is a great question and obviously something that we already contend with. The internet is cross-jurisdiction. There are court cases about music played on aeroplanes. You can sue where the harm occurs. In general—it would depend on the case, obviously—it is possible to sue where the harm occurs.

Lord Foster of Bath: If one of Lord Hall’s books has been used without permission in generating the base model in America, but that is then used for a particular purpose in the United Kingdom, the hurt is to him here but to his worldwide sales because his books are sold all around the world. Does he sue in every country? Perhaps you could write and explain.

The Chair: That sounds like a question that does not lend itself to a straightforward, simple answer.

Lord Foster of Bath: Yes, but we need to know.

The Chair: Perhaps, Dr Bosher, I could invite you to write to us with an answer to that question.

Dr Hayleigh Bosher: I can do that, yes.

The Chair: That would be very helpful. Thank you, all four of you, for your evidence today. It was very helpful. To reiterate a point that Mr Mollet made earlier, we as a committee very much understand that, when it comes to innovation and contributions to economic growth, both the creative industries and the content that emerges from the creative industries are incredibly valuable, and it is not just all innovation from tech. Certainly, in our creative industries inquiry last year, one of the things that we were very clear about in the context of text and data mining was that all innovation in tech cannot be at the expense of the creative industries. That is a balance that we are very conscious of and very much seek to reflect in the way in which we are examining this very difficult question as part of the inquiry.

Your evidence has been a great help. I am very grateful to you for the time that you have given us this afternoon. For anybody watching us live on the internet, please join us again tomorrow afternoon when we will be meeting to take evidence on the biggest question, which is about open source versus closed source when it comes to large language models. We will pause our sitting for now. Thank you very much indeed.