Baroness Beeban Kidron
Submission to the Data (Use and Access) Bill: Call for Evidence

The focus of this evidence are the amendments to the Data (Use and Access) Bill, which I tabled at Report Stage in the House of Lords and were voted into the Bill, and now make up clauses 95(1), 135, 136, 137, 138 and 139. This evidence will provide brief descriptions of the effect of these provisions and focus on dispelling common myths about their impact from tech lobbyists and others.

Introduction

These provisions would retain the current copyright framework for text and data mining—where the onus is on AI firms to seek a license – and ensure the law is complied with by all web scrapers and general-purpose AI models linked to the UK. *This is the only way to guarantee that UK creators have control and will be compensated fairly for the use of their work.*

Meaningful transparency, at a level of granularity that allows a rights holder to understand when their work has been used – as well as transparency over the crawlers that scrape content – will incentivise AI firms to comply with the law and allow creators to identify illegal conduct and seek redress. These measures are underpinned by strong enforcement powers.

In short, instead of trading away copyright protections for a vague promise of transparency which the tech lobby is already decrying, these provisions promise meaningful transparency within a time frame that is necessary to respond to widespread illegal scraping. It is clear this illegal scraping is happening, but these provisions would make clear which firms are doing it. This would make the UK's existing 'gold standard' copyright regime enforceable.

As parliament and public has begun to understand the government's proposed 'opt out' they are increasingly confused as to why the government has chosen a position that overwhelmingly benefits companies based in US and China, at the expense of an industry that meets all the desired outcomes of growth, productivity, international excellence and employs a highly skilled workforce. Meanwhile lobbying from tech firms and aligned organisations has become increasingly desperate, leading to some fundamentally misleading claims being made.

In a recent briefing sent to parliamentarians that argues against the copyright and AI provisions in this Bill, industry lobby group techUK gave the examples of AI that "*detects diseases like Alzheimer's and cancer*" and models that "*forecast hurricanes, wildfires, and floods*" when making the case for their removal of the copyright and AI clauses from the Bill.

It is highly cynical – verging on dishonest – to use these examples when arguing against the robust protection of creative content through copyright law: UK AI firms innovating in science, medicine, and many other fields do not have the slightest need for creative content. Simply put, AI firms do not need Dua Lipa's lyrics to cure cancer, or the works of Stephen Fry to prevent a flood. And the small language models that do use creative content - which are developing to serve specific business needs in various sectors - are trained on very niche, focused datasets of high-quality copyright content. Indeed, the gold standard data set for medical breakthrough is utilised by AlphaFold, which uses highly focused, curated and legally acquired data including the Protein Data Bank**.**

The techUK briefing also rejects the framing of the AI and copyright debate as a *"zero-sum contest between technology and human creativity"* and sets out examples of the uptake of AI by creatives. That is not contested by the creative industries: the creative sector are early adopters of technology, and the UK creative industries are eager to embrace AI. The eagerness to embrace AI does not extend to giving our work away for free so that we can rent it back from AI models. We will not be able to sustain ourselves if we are not fairly renumerated for our work. That is why clauses 95(1), 135, 136, 137, 138 and 139 have been crafted with the intention of spurring a dynamic licensing market for creative content in the UK.

It is clear that the government's 'opt out' proposal was written by Big Tech, for Big Tech, and will not spur growth in the UK economy. It will inevitably stymie growth in the creative sector, and the claims it would bolster the tech sector are entirely unevidenced. Ministers have been unable to say how a fall in the creation of a key ingredient in generative AI models – high-quality creative content – will do anything to bolster the UK AI sector.

Preferring to back vague claims from large tech companies, that the UK will 'be left behind'. It is crucial for the Committee to ask: *"What money will be lost? What money will be made. Who will it accrue to?"*. It is worth considering the potential contribution to the public purse from UK companies, rather than Big Tech who's negotiated position is not to pay at the same rate, or maybe at all in the near future.

These clauses simply seek to make a pre-existing and long-established UK property right enforceable in the age of AI. It is absurd to frame the creative industries as blockers on AI innovation when generative AI would simply not exist without our work. AI firms do not seek free computer chips from Nvidia, free electricity from the National Grid, or free labour from tech experts: they should not seek to avoid renumerating the creators without which their models would not function.

This is not just a question of fairness for creators: robust copyright law with transparency and strong enforcement mechanisms is the only route to a future where creators and AI firms can develop a mutually beneficial relationship.

Clause 135

This clause would:

- Ensure web crawlers and general-purpose AI models comply with UK copyright law when they are linked to (e.g. marketed in) the UK; and
- make clear that the law applies to the entire life cycle of an AI model.

Retaining UK copyright law

This clause retains UK copyright law which – despite claims by Ministers and AI firms – is completely clear that the unlicensed commercial use of copyright material is theft.

There have been claims that we must weaken our copyright laws to attract AI investment, and to persuade Big Tech to abide by our laws. But it is completely illogical to argue that weakening copyright law will make AI firms more likely to comply with it. Moreover, there are multiple examples of companies that do pay (Voice-Swap and Adobe among them). As UK AI – the trade body for the UK AI sector - say: *"The opt-out model would significantly harm the creative sectors to achieve a minimal gain for a handful of global tech companies. This creates unnecessary conflict which will erode public trust in the nascent AI sector, limiting its future growth. In the long term, this will undermine economic growth more than any benefits the opt-out could deliver".*

The government's proposed 'opt out' will simply make it easier for AI firms to use content for free whilst continuing to train AI models outside the UK. Indeed, cheaper energy costs and weaker data protection laws in other jurisdictions will encourage them to take the data and run. In contrast – if properly protected - curated, specialist and valuable data sets will entrench the UK position as a data rich nation (in the English language) and with an unparalleled history and future of original content.

That is why extra-territorial provisions are required, and these are discussed below.

Alignment with other jurisdictions

It has been claimed that the UK must weaken its copyright law in order to compete with other jurisdictions such as the US. But the position of US copyright law as regards AI training is far from settled.

If the US courts determine that the use of copyright material for AI training is not 'fair use', then the US will effectively have a copyright framework with the same requirements as existing UK law – that AI developers must seek a license before using copyright material.

It seems increasingly likely that this is the case. Recently, in Thomson Reuters v ROSS Intelligence, a judge determined that the use of Westlaw headnotes – which are created by Thomson Reuters – to train ROSS's Bulk Memos AI tool was not 'fair use'. Whilst not focused specifically on generative AI, this ruling is a key indication that the US copyright

law in the context of AI is far from settled, making claims that the UK must change its law to align with the US entirely spurious.

Extraterritoriality

The application of copyright law to any model linked to (e.g. marketed in) a jurisdiction – as opposed to copyright law only applying where a model is trained in the jurisdiction – is not novel, and a similar provision is contained in the EU AI Act. Indeed, UK creatives argue that UK law already covers the training of models outside the UK when the model is brought here. For example, Getty Images alleges that Stability AI's actions constitute secondary infringement because Stable Diffusion is an infringing 'article' brought to the UK without authorisation.

Tech lobbyists have sought to claim that we cannot impose our laws on models trained outside the UK. Yet legislation such as the Online Safety Act (which these clauses draw upon to define a model "*linked to*" the UK) and the Digital Markets, Competition and Consumers Act demonstrates that we are not powerless to control the products that Big Tech bring to the UK.

Tech lobbyists have also sought to claim that ensuring that AI firms who wish to operate in the UK abide by our laws will discourage AI investment and reduce the availability of AI products in the UK. This is a threat that has been made by tech firms again and again without foundation: the government needs to make a choice about whether it wants our laws to be made here or in Silicon Valley. And again, the UK AI trade body clearly want a trusted and ethical data ecosystem to support an alternative business model to Silicon Valley – many firms, from aerospace and cyber security, to fashion and creative writing desperately need this ecosystem to develop.

Clauses 136 and 137

These clauses would:

- Provide transparency over the crawlers used to scrape creative content;
- ensure crawlers deployed for specific purposes are separated out to prevent content being scraped for search services also being used to train AI; and
- provide granular transparency for creators – to a level that allows individual works to be identified – over the creative works scraped by web crawlers and general-purpose AI models.

Granular transparency

Tech lobbyists have criticised the government's very vague commitment to transparency in the recently concluded consultation. The Computer and Communications Industry Association UK (representing the very largest tech firms) has said that transparency requirements could see AI firms give "*away trade secrets and highly sensitive information that could jeopardise the safety and security of their models.*"

It is clear that the transparency requirements in clause 137 would in no way risk trade secrets being given away or create security risks. It is the curating, prompts and weighting of information that forms valuable IP for an AI model. The only reason for the secrecy is that transparency will shine a light on a business model that is taking its raw material from others without permission and payment. All the provisions would do is ensure that copyright holders can be informed whether *their own work* - and nothing else - has been used, incentivising AI firms to comply with the law.

General-purpose AI models have billions of data points, so it is laughable to suggest that allowing individual businesses or creators transparency over the use of their IP could constitute the giving away of trade secrets.

I would also note that, if AI firms have acted within the law and agreed licenses with copyright holders for the use of their content, it will be extremely easy for them to be transparent about the use of content as the license will already exist. It is only firms acting illegally who would potentially find the transparency requirement burdensome.

Clause 138

This clause would:

- Apply powers given to the Information Commissioners Office (ICO) under the Data Protection Act to the enforcement of the copyright and AI provisions, including the power to fine and issue enforcement orders; and
- allow copyright holders to take private action against AI firms for non-compliance, providing a dual route to enforcement.

The suitability of the Information Commissioners Office as a regulator

The ICO conducted a series of consultations on this issue of AI training and web scraping in 2024, concluding that: "*Web scraping for generative AI training is a high-risk, invisible processing activity. Where insufficient transparency measures contribute to people being unable to exercise their rights, generative AI developers are likely to struggle to pass the balancing test*." Clearly, the ICO has already considered very similar issues to the ones that are the focus of these clauses.

Given that the regulator has already considered the legality of scraping in depth as regards personal data, they are eminently well suited to discharge the powers given to it by these clauses. It will also be more efficient for the same regulator to consider data protection law and copyright in relation to AI training in parallel, rather than AI firms having to deal with two separate regulators. Minsters say this has not been costed but given that their preferred policy is undermining a £126 billion industry, the added capacity to the ICO is a small price to pay.