News Media Association Briefing Paper
Sustaining Trusted Journalism in the Age of AI

Key Points:

- Generative Artificial Intelligence ("**GAI**") firms have built a business model based on mass copyright infringement. This threatens the sustainability of trusted news publishers, as Large Language Models ("**LLMs**") use content without renumeration and are then used to compete with trusted news sources.
- Without urgent action by policymakers, the resulting fall in publisher revenues will lead to less and less trusted news content being produced. Perhaps even greater than the threat of malicious GAI disinformation and misinformation is the threat that the proliferation of LLMs will lead to a sharp drop in the production of original journalism.
- In this scenario, GAI will have nothing to train on but its own outputs, leading to a phenomenon known as 'model collapse', whereby LLMs collapse under the weight of their own misinformation. This means that, ultimately, the unlicensed use of copyrighted content threatens the sustainability of GAI firms themselves.
- The Government and Intellectual Property Office ("**IPO**") must clarify that publishers have control over their copyrighted content and must be asked for their consent before it is scraped and used to train GAI systems.
- Transparency obligations must be placed on GAI firms in order for publishers to understand when their content has been used by an LLM.
- GAI firms must be required to identify the crawlers used to scrape content, and what purpose the crawler has.
- The Government must support structures and mechanisms which allow publishers and other creators to be compensated when their content is used to train LLMs.

A Business Model Based on Mass Copyright Infringement Threatens the Sustainability of Trusted News Publishers and the GAI Sector Itself

1. GAI firms have created a business model which is based on mass copyright infringement, building GAI tools on the back of the often unauthorised and unremunerated use of copyright works. In the UK, the text and data mining (**TDM**) that trains GAI systems is currently only permitted for the purpose of research for a non-commercial purpose. The UK should not be tempted to adopt a more permissive TDM regime.

2. Trusted news content is incredibly valuable for GAI developers. Google's C4 dataset, which has been used to train the tech giant's flagship Google Gemini (formerly Bard) LLM, has five news publisher websites – including The Guardian - in its top ten sources. Training materials are passed through LLMs again and again, leading 'memorisation' whereby models can reproduce large portions of the copyrighted material that they are trained on. News content is not only valuable in the initial training of LLMs. Once trained, LLMs can

be provided with a reference to specific material by a user – such as a news article – to 'ground' their output, leading to a response which may extensively copy, or paraphrase copyrighted material.

3. If individuals can access publishers' highly valuable news content through GAI firms' products without having to pay for it – bypassing the need to pay for a subscription or go to a publisher's website where advertising is used to monetise content – then many will begin to do so. In this way, GAI firms not only infringe copyright on a massive scale, but then use their LLMs to directly compete with trusted publishers.

4. Microsoft has already integrated its ChatGPT technology into its Bing search engine, whilst Google is experimenting with its Search Generative Experience product. Both of these integrate an LLM at the top of a traditional search engine: if a user is provided with a GAI response to a query, it is unlikely that they will need to go to a publishers website where a publisher can monetise their content.

5. If publishers are (a) not renumerated for the use of their IP when it is used to train LLMs; and (b) those LLMs compete with and supplant news publishers as a source of information, then very soon publishers will lose revenue and investment in original news content will fall. Even greater than the threat of malicious GAI disinformation and misinformation is the threat that the proliferation of LLMs leads to a sharp drop in the production of original journalism.

6. In this scenario, GAI will have nothing to train on but its own outputs, leading to a phenomenon known as 'model collapse', whereby LLMs collapse under the weight of their own misinformation. This will see misinformation proliferate, not as a result of bad actors manipulating LLMs, but simply because it will become impossible to train LLMs on reliable, trusted information.

7. This will leave citizens with less reliable news content and news outlets: this isn't just bad for the public and publishers: it is also an existential risk for GAI firms as publishers and other content creators will stop investing in the copyrighted content that fuels LLMs. Therefore, for the sustainability of trusted news publishers and for the sustainability of the GAI industry itself, news publishers must be compensated for the use of their IP. This is also essential to prevent GAI misinformation flooding the internet.

Policy Solutions to Support a Sustainable Commercial Relationship Between Trusted News Publishers and GAI Firms

8. The Government and IPO must clarify that publishers have control over their content and must be asked for their consent before it is scraped and used to train GAI systems. Whilst mass copyright enforcement has already occurred, new LLMs will be created and existing LLMs will need to be updated frequently over time, meaning these solutions are still urgently needed.

9. At present, publishers are unable to determine precisely how their content has been used to train LLMs, or when it is used to 'ground' responses. Transparency obligations must be

placed on GAI firms in order for publishers to understand when their content has been used by an LLM, both in the inputs and outputs. Absent this transparency, publishers will be unable to assert their rights. Such transparency will also be imperative for users of LLMs to understand whether the outputs have been based on trusted, reliable information.

10. This transparency could be achieved by requiring GAI firms to provide a regulator with the details of what data has been used to train an LLM. A publisher could then be allowed to make a request to the regulator to determine if their content has been used to train the model. If this is the case, the regulator could then allow the publisher with a detailed view of what copyrighted content has been used. Without this transparency, publishers will be unable to assert their rights and negotiate a license for the use of their content.

11. Another key barrier to fair remuneration is the opacity surrounding the 'scraping' of publishers' websites for the data used to train LLMs. GAI firms must be required to identify the 'crawlers' used to scrape content, and what purpose the crawler has. Crawlers must be separated according to purpose, so that rightsholders have a clear choice about which crawler to block and which to license.

12. For example, a publisher may wish to allow access for a crawler so that their content can be indexed on a search engine but may wish to block a crawler from the same firm that is being used to scrape content to build an LLM. Heavy penalties must be imposed on GAI crawlers which scrape content in defiance of a machine-readable signal that the publisher has placed on its website to prevent crawling. This way, publishers will be given the power to consent or deny access to their websites for the purposes of building LLMs.

13. The Government must support structures and mechanisms which allow publishers and other creators to be compensated on fair and reasonable terms when their content is used by LLMs. For Big Tech owned and backed LLMs, this could include the use of the Competition and Markets Authority's ("**CMA**") new powers (under the Digital Markets Competition and Consumers Bill) to compel gatekeepers to "trade on fair and reasonable terms" with third parties such as news publishers. The Government should use its Strategic Steer to suggest the CMA looks at this as a priority, particularly where LLMs have been integrated into existing dominant digital services such as search engines.

14. GAI firms must also be required to abide by UK law – including copyright law – if they wish to trade in the UK. This is necessary to prevent GAI firms claiming that they are not required to license content from UK publishers if the training of their LLM took place in another jurisdiction.

Sebastian Cuttill, Parliamentary and Campaigns Manager
News Media Association, March 2024