frontiers

RE: Frontiers Response to Request for Information on Artificial Intelligence and Copyright (Federal Register Number 2023-18624, Document ID COLC-2023-0006-0001)

December 5, 2023

Summary response

We welcome the chance to respond to this important request for information from the United States Copyright Office (USCO). Frontiers is a leading research publisher and open science platform, the third most cited and sixth largest in the world. The science we publish is peer-reviewed, globally shared, and free to read.

Our mission is to make all science open – so that we can collaborate better and innovate faster, for fairer and more equitable outcomes in all parts of society. That is our social purpose as a business.

We are above all a knowledge, information, and technology company. We made the founding decision to build our own open science platform and we continually develop, improve, and customize it to meet the evolving needs of the scientific community.

And it is worth nothing that the articles we publish come under the CC-BY licence, allowing others to distribute, remix, adapt, and build on them, provided that attribution is given to their original creator. While it is standard practice for subscription paywall publishers to require that authors surrender their work's copyright, all authors in all our journals retain all their rights, and copyright is not transferred to Frontiers.

This approach delivers truly open science that is freely and permanently available for anyone to view, download, and disseminate in interoperable, machine-readable formats, allowing all authors to commercially manage and exploit their intellectual property as they wish.

Our position is three-fold:

First, as a business whose driving purpose is to support the scientists who wish to publish their research with us, we back the principle of author attribution.

- We think applying attribution to AI and LLM outputs – and doing so in a way that is easily accessible to both end user and rights holder – is both ethically and commercially compelling, notwithstanding that the cost and regulatory implications have yet to be fully understood.

- The creative and intellectual contribution of our authors derives in large part from extensive community collaboration – backed in publication by a robust approach to editorial oversight, rigorous peer review, and immediate testing in the court of public opinion, all of which sharpen the quality and impact of new published knowledge.

- The effectiveness of that scientific collaboration and publication is in turn founded on attribution – and the ability to trace, test, cite, and develop the science. A growing corpus of AI and LLM outputs that attributes authors will be key to scientific collaboration at scale. If we are to overcome existential threats, from health emergencies to climate change, global scientific collaboration will be essential to healthy lives on a healthy planet.

Second, it can easily be argued that the better the AI and LLM inputs – defined here as scientifically accurate, validated, verified, peer reviewed, and tested by public opinion – the better will be the outputs. We would encourage the USCO to consider how scientific publishers might provide more inputs to train AI and LLMs.

- Frontiers provides a publishing platform for such research that is immediately tested by public opinion. We do this with industry-leading editorial rigor, technology, and expertise, and a robust peer review underpinned by industry-wide ethical standards.

- The Washington Post reported this year on the websites contributing most to Google's C4 data sets, which in turn "have been used to instruct some high-profile English-language AIs... including Google's T5 and Facebook's LLaMA." Frontiers is listed in the top 15 contributing websites but is joined by only one other open access publisher.

And **third**, in that context, we think the opportunity for better quality AI and LLM inputs is being squandered by a legacy scientific publishing system that locks scientific knowledge behind paywalls.

- The global investment in research, which today approaches $3 trillion a year, generates, among other things, around four million peer reviewed papers each year that capture new knowledge. Most of that knowledge[1] remains locked behind paywalls, slowing down the societal benefits of leveraging AI and LLM with higher quality inputs.

Within this framework, we fully support the August 2022 Office of Science and Technology Policy (OSTP) guidelines on immediate public access to federally funded research. We strongly welcomed them at the time, and the White House amplified that support.

We think it is possible to achieve the fullest possible access to our collective knowledge – for fairer outcomes in all parts of society – in a business model where AI and related technologies are cost-effective, commercially sustainable, and underpinned by private sector innovation. We stand ready to support the USCO and its partners in the federal government. It is vital we back responsible AI efforts for the good of open science and to meet the public appetite for accountability, transparency, and trust.

We tackle below the questions we feel best equipped to answer.

Full Response

General Questions

1. As described above, generative AI systems have the ability to produce material that would be copyrightable if it were created by a human author. What are your views on the potential benefits and risks of this technology? How is the use of this technology currently affecting or likely to affect creators, copyright owners, technology developers, researchers, and the public?

We think the USCO has posed critical questions in this request for information about the opportunities, challenges, risks, and benefits of AI. General, human-compatible AI could empower us all, but public trust in good science will be key.

[1] By published articles. Dimensions data for published articles 2022.

Our starting point is that applying attribution to AI and LLM outputs – and doing so in a way that is easily accessible to both the end user and rights holder – is both ethically and commercially compelling, notwithstanding that the cost and regulatory implications have yet to be fully understood.

The creative and intellectual contribution of our authors derives in large part from extensive community collaboration – backed in publication by a robust approach to editorial oversight, rigorous peer review, and immediate testing in the court of public opinion, all of which sharpen the quality and impact of new published knowledge.

The effectiveness of that scientific collaboration and publication is in turn founded on attribution – and the ability to trace, test, cite, and develop the science. A growing corpus of AI and LLM outputs that attributes authors will be key to scientific collaboration at scale. If we are to overcome existential threats, from health emergencies to climate change, global scientific collaboration will be essential to healthy lives on a healthy planet.

So, we think the benefits of this technology are substantial in that context, both in accelerating that collaboration, and in gathering better AI and LLM inputs – defined here as scientifically accurate, validated, verified, peer reviewed, and tested by public opinion. Frontiers provides a publishing platform for such research that is immediately tested by public opinion. We do this with industry-leading editorial rigor, technology, and expertise, and a robust peer review underpinned by industry-wide ethical standards. We would encourage the USCO to consider how scientific publishers might provide more inputs to train AI and LLMs.

Moreover, at Frontiers, we apply AI tools to help establish public trust in the scientific record. Our Artificial Intelligence Review Assistant (AIRA) verifies that scientific knowledge is accurately and honestly presented even before our people decide whether to review, endorse, or publish the research paper that contains it.

AIRA reads every research manuscript we receive and makes up to 20 checks a second. These cover, among other things, language quality, the integrity of figures and images, plagiarism, and conflicts of interest. The results give editors and reviewers another perspective as they decide whether to put a research paper through our rigorous and transparent peer review.

Our platform ensures manuscripts sent to peer review reflect the standards that are essential for high-quality scientific research. Peer review is widely recognized in the scholarly community as a vital process for ensuring quality in scientific literature. However, the demand for peer review is growing at an immense rate. According to data from Dimensions, in 2019 more than 4.2 million articles were published, compared to just 2.2 million articles only 10 years prior. The increasing amount of scientific literature published and the growing demand for high-quality peer review now makes the use of novel decision support technologies essential.

2. Does the increasing use or distribution of AI-generated material raise any unique issues for your sector or industry as compared to other copyright stakeholders?

It is clear that the better the AI and LLM inputs – defined here as scientifically accurate, validated, verified, peer reviewed, and tested by public opinion – the better will be the outputs. Frontiers provides a publishing platform for such research that is immediately tested by public opinion. We do this with industry-leading editorial rigor, technology, and expertise, and a robust peer review underpinned by industry-wide ethical standards. We would

encourage the USCO to consider how scientific publishers might provide more inputs to train AI and LLMs.

By way of example, the Washington Post reported this year on the websites contributing most to Google's C4 data sets, which in turn "have been used to instruct some high-profile English-language AIs... including Google's T5 and Facebook's LLaMA." Frontiers is listed in the top 15 contributing websites but is joined by only one other open access publisher.

In that context, we think the opportunity for our sector to generate better quality AI and LLM inputs is being squandered by a legacy scientific publishing system that locks scientific knowledge behind paywalls. The global investment in research, which today approaches $3 trillion a year, generates, among other things, around four million peer reviewed papers each year that capture new knowledge. Most of that knowledge[2] remains locked behind paywalls, slowing down the societal benefits of leveraging AI and LLM with higher quality inputs.

Leveraging AI to that end is already underway, but more can be done. At the same time, it is vital these AI solutions do not create or perpetuate inequity. The governance mechanisms and safeguards being proposed vary widely, and substantial new thinking and public funding will be required to bring that variance down and lift standards for compliance. The public funding of AI infrastructure and oversight must be as efficient, scalable, and as good a value for money as possible.

Open access to the results of publicly funded research is of huge value, offering significant social and economic benefits. The open access publishing model, defined by the Berlin Declaration in 2003, improves the pace, efficiency, and value of research to society. By its very nature, it improves the visibility of authors' work and enables better scholarly exchange, and therefore, the potential impact of that work.

3. Please identify any papers or studies that you believe are relevant to this Notice. These may address, for example, the economic effects of generative AI on the creative industries or how different licensing regimes do or could operate to remunerate copyright owners and/or creators for the use of their works in training AI models. The Office requests that commenters provide a hyperlink to the identified papers.

Of particular note are the following proposals put forward in this white paper, 'The AI Governing Challenge,' from June 2023, namely that:

"Congress should create special select committees on AI in the Senate and House to take on the task of rising to these challenges with the Chairs of key committees represented. The Administration should create an AI Policy Coordinator and permanent office in the White House with direct access to the President.

"Only a handful of firms will have the capital, computing capabilities, and talent necessary to create and deploy powerful AI models. A licensing regime should govern their operations and they should be subject to regular public audits. Before a licensed entity releases any AI tool to the public, it should undergo rigorous safety testing and auditing for protection against specific potential harms. Every company that evaluates and distributes such a tool should have usage policies for its users and partners and effectively enforce those policies.

[2] By published articles. Dimensions data for published articles 2022.

"Users of powerful AI systems for the delivery of content or services should ensure that they make transparent to end users or consumers that the content consumers are viewing or the service they are receiving was generated using AI in part or in whole. They must make available an appeal process for any AI enabled decision of significant consequence for the individual. No AI tool should enable the violation of people's privacy or civil rights.

"Agencies should use existing authorities to enforce existing law on creators and users of AI systems. Congress should provide the appropriations necessary for the agencies to acquire the technical expertise necessary to evaluate AI models and tools and their use."

4. Are there any statutory or regulatory approaches that have been adopted or are under consideration in other countries that relate to copyright and AI that should be considered or avoided in the United States? How important a factor is international consistency in this area across borders?

We think the results of publicly funded research should be open to all of society. There is exponential growth in the knowledge produced by scientific, medical, and technical (STM) research, and new tools are being developed that can exploit this data in powerful ways. One of the most promising of these tools is text and data mining (TDM), i.e., the automated computational analysis of digital content.

Many STM publishers license their databases for TDM; these license structures can provide a template for licensing to generative AI models and systems.

As with TDM before it, the European Commission recognizes the potential of AI and LLMs and is considering updating and clarifying the legal provisions for their use. In the case of TDM, Frontiers urged European legislators to support a copyright exception that clearly includes all research bodies (i.e., businesses and SMEs, as well as universities, institutions, and citizen scientists) that have lawful access to the digital content.

5. Is new legislation warranted to address copyright or related issues with generative AI? If so, what should it entail? Specific proposals and legislative text are not necessary, but the Office welcomes any proposals or text for review.

Frontiers supports the recommendation in the Oct 30 Executive Order (EO) issued by the White House setting new standards on the development and deployment of AI systems. Many of the actions outlined in the EO were included in a non-binding agreement signed by CEOs of the largest tech AI companies earlier this year.

The EO does not carry the weight of legislation and includes no mention of how AI companies address concerns surrounding intellectual property. That said, Frontiers agrees with the EO's recommendations on **privacy**, which ask Congress to pass bipartisan data privacy legislation that includes provisions on limitations for training data used in AI systems.

Separately, Frontiers welcomes the bipartisan CREATE AI Act to expand access to artificial intelligence research in the US. It shows foresight, creativity, and the chance to properly weigh the risks and benefits of AI for all.

The bill proposes new national infrastructure for the US that gives researchers, academics, startups, and students from diverse backgrounds much-needed access to the computing power, resources, data, and tools needed to develop safe and trustworthy artificial intelligence.

The principle of open access to these resources – and the ambition to expand research on this cutting-edge technology to the best and brightest minds in the US – chimes loudly with our mission as an Open Science research publisher.

Training

If your comment applies only to a specific subset of AI technologies, please make that clear.

6. What kinds of copyright-protected training materials are used to train AI models, and how are those materials collected and curated?

By publishing Gold Open Access content under a CC-BY license, and in doing so creating publications that are easily searchable and ingestible, Frontiers has long presumed that all of our published content is already being used to train AI and LLMs.

That said, we think there are substantial, untapped opportunities to grow the impact of that training content.

It is clear that the better the AI and LLM inputs – defined here as scientifically accurate, validated, verified, peer reviewed, and tested by public opinion – the better will be the outputs. Frontiers provides a publishing platform for such research that is immediately tested by public opinion. We do this with industry-leading editorial rigor, technology, and expertise, and a robust peer review underpinned by industry-wide ethical standards. We would encourage the USCO to consider how scientific publishers might provide more inputs to train AI and LLMs.

By way of example, the Washington Post reported this year on the websites contributing most to Google's C4 data sets, which in turn "have been used to instruct some high-profile English-language AIs... including Google's T5 and Facebook's LLaMA." Frontiers is listed in the top 15 contributing websites but is joined by only one other open access publisher.

In that context, we think the opportunity for better quality AI and LLM inputs is being squandered by a legacy scientific publishing system that locks scientific knowledge behind paywalls. The global investment in research, which today approaches $3 trillion a year, generates, among other things, around four million peer reviewed papers each year that capture new knowledge. Most of that knowledge[3] remains locked behind paywalls, slowing down the societal benefits of leveraging AI and LLM with higher quality inputs.

6.1. How or where do developers of AI models acquire the materials or datasets that their models are trained on? To what extent is training material first collected by third-party entities (such as academic researchers or private companies)?

The Washington Post did a tremendous service in demystifying ingestion processes by analysing C4, the Google dataset used to train AI and LLMs. The Post's analysis showed that 9% of C4 contains science and health content, and within that subset, Frontiers content is the second-most prominent resource, trailing only another Gold Open Access publisher, the Public Library of Science (PLOS).

[3] By published articles. Dimensions data for published articles 2022.

6.2. To what extent are copyrighted works licensed from copyright owners for use as training materials? To your knowledge, what licensing models are currently being offered and used?

As noted above: By publishing Gold Open Access content under a CC-BY license, and in doing so creating publications that are easily searchable and ingestible, Frontiers has long presumed that all of our published content is already being used to train AI and LLMs.

However, the parties ingesting Frontiers content are not obligated to report doing so; many parties scrape so much data from the Internet that they cannot name all of the content used to train their models. This lack of information stands in opposition to the transparency inherent in the Open Access model and the required attribution of the CC-BY license.

9. Should copyright owners have to affirmatively consent (opt in) to the use of their works for training materials, or should they be provided with the means to object (opt out)?

Frontiers publishes content under the CC-BY license, which requires attribution and allows for commercial and derivative uses of content. Frontiers is not the developer of this license and at present does not take a position on the license's opt-in or opt-out specifications. We do, however, support authors exercising their copyrights to the fullest extent.

Transparency & Recordkeeping

15. In order to allow copyright owners to determine whether their works have been used, should developers of AI models be required to collect, retain, and disclose records regarding the materials used to train their models? Should creators of training datasets have a similar obligation?

Put simply, under the Gold Open Access publishing model the input is "nice to know" data and the output is "need to know" data.

Regarding inputs: As a fully OA publisher applying the CC-BY license, Frontiers presumes that any party capable of scraping our content is already doing so. We'd prefer to know all inputs (that is, all parties scraping our content, especially for the purposes of training AI models). However, monitoring every input is not an effective use of time or resources for a Gold OA publisher.

This is a significant distinction between Gold Open Access publishers and legacy subscription publishers that keep content behind paywalls. Companies that overwhelmingly rely on paywalled and/or subscription-based content typically hold the copyright to that content and assign various terms and conditions to its reuse.

Therefore, legacy subscription publishers want to know the input regardless of the potential for output, as input itself will be viewed by these publishers as a licensable event.

Regarding outputs: As stated above, our driving purpose is to support the scientists who wish to publish their research with us. So, we back the principle of author attribution. We think applying attribution to AI and LLM outputs – and doing so in a way that is easily accessible to both the end user and rights holder – is both ethically and commercially compelling, notwithstanding that the cost and regulatory implications have yet to be fully understood.

The creative and intellectual contribution of our authors derives in large part from extensive community collaboration – backed in publication by a robust approach to editorial, rigorous peer review, and immediate testing in the court of public opinion, all of which sharpen the quality and impact of new published knowledge.

The effectiveness of that scientific collaboration and publication is in turn founded on attribution – and the ability to trace, test, cite, and develop the science. A growing corpus of AI and LLM outputs that attributes authors will be key to scientific collaboration at scale. If we are to overcome existential threats, from health emergencies to climate change, global scientific collaboration will be essential to healthy lives on a healthy planet.

15.1. What level of specificity should be required?

If the output includes scientifically meaningful material that in an academic setting would be cited, specific labelling would be vital to understanding how inputs are put to use. The burden should not be on the rights holder to prove that the content helped to generate the output.

15.2. To whom should disclosures be made?

Under the Gold Open Access model and CC-BY license, disclosures (i.e., attribution) should be made to the authors of the content, either directly to those authors or via a notice to their publisher.

15.3. What obligations, if any, should be placed on developers of AI systems that incorporate models from third parties?

Developers of AI systems should be obligated to abide by the terms and conditions of the CC-BY license, which requires attribution.

15.4. What would be the cost or other impact of such a recordkeeping system for developers of AI models or systems, creators, consumers, or other relevant parties?

This is a question best addressed by the developers of the AI models. That said, AI is a licensable event, and significant business norms already exist for the tracking of licensable content.

16. What obligations, if any, should there be to notify copyright owners that their works have been used to train an AI model?

As noted above, developers of AI systems should be obligated to abide by the terms and conditions of the CC-BY license, which requires attribution.

17. Outside of copyright law, are there existing U.S. laws that could require developers of AI models or systems to retain or disclose records about the materials they used for training?

As noted above, AI is a licensable event, and significant business norms already exist for the tracking of licensable content.

Labeling or Identification

28. Should the law require AI-generated material to be labeled or otherwise publicly identified as being generated by AI? If so, in what context should the requirement apply and how should it work?

Frontiers expects that AI systems generating outputs should provide attribution describing how those outputs were derived – a “nutrition label” of sorts indicating how the output was packaged and prepared in response to a prompt.

Admittedly, the “nutrition label” analogy goes only so far; without knowing the prompt, it’s difficult to discern if the output made proper use of the inputs. But if the output includes scientifically meaningful material that in an academic setting would be cited, such labelling would be vital to understanding how inputs are put to use.

Such labelling not only meets the attribution and "Adapted Material" requirements of the CC-BY license but also serves as a citation, the accepted and measurable metric of contribution to the scientific literature.