# PUBLIC SUBMISSION

**Received:** May 28, 2025
**Tracking No.** mb8-cshp-cz2u
**Comments Due:** May 28, 2025 **Submission Type:** API

**Docket:** NSF-2025-OGC-0001
NITRD_FRDOC_0001

**Comment On:** NSF-2025-OGC-0001-0001
Request for Information: Development of a 2025 National Artificial Intelligence Research and Development Strategic Plan

**Document:** NSF-2025-OGC-0001-DRAFT-0148
Comment on FR Doc # 2025-07332

## Submitter Information

**Organization:** Association of American Publishers

## General Comment

Please find attached a comment for NSF-2025-OGC-0001 from the Association of American Publishers.

J. Carl Maxwell
Senior Vice President
Association of American Publishers

## Attachments

NSF-2025-OGC-001_AssociationAmericanPublishers

AAP | ASSOCIATION OF AMERICAN PUBLISHERS

Electronic submission via www.regulations.gov
Docket ID No. NSF-2025-OGC-0001

May 29, 2025

**NCO/NITRD**
2415 Eisenhower Ave.
Alexandria, VA 22314

***Subject:*** ***Request for Informationon the Development of a 2025 National Artificial Intelligence (AI) Research and Development (R&D) Strategic Plan (NSF-2025-OGC-001)***

To whom it may concern:

The Association of American Publishers (AAP) is the national trade association of the U.S. book and journal publishing industry. AAP represents the leading book, journal, and education publishers in the United States on matters of law and policy, advocating for outcomes that incentivize the publication of creative expression, professional content, and learning solutions.

AAP appreciates the opportunity to provide its views in response to this Request for Information (RFI) regarding a strategic research and development plan for artificial intelligence. We note at the outset that, in addition to identifying national priorities, the government should put in place a cohesive framework defining accountability measures and policies to foster responsible and ethical development of trustworthy, human-centric AI technologies. AI technologies will likely improve our lives and positively impact society — in health care, finance, the environment, education, and in accelerating scientific discovery. However, for this transformational technology to affect beneficial outcomes for all of society, the public must trust this technology, and trust presupposes accountability. Trust exists where there is knowledge and understanding, and in the AI context, this means knowing:

1) how the AI technology was developed, including whether the technology conforms with applicable laws or regulations (such as data privacy) or best practices,

2) what content was used to develop the datasets on which the AI technology was trained, and whether this content was legitimately sourced, and

3) the vetting or testing processes the technology has undergone to assess its effectiveness and reliability, and therefore, its readiness for deployment to the public.

This requires that AI technology developers adequately communicate to business end users and to society how and on what content the models were trained, as well as the results of assessments and audits performed and by whom. These obligations of transparency and disclosure should likewise apply to downstream entities that build on and incorporate AI technologies into the products and services they

deploy to customers or other business end users.[1] Transparency as to data provenance, training processes, and testing is essential to building the public's confidence that AI technology generates fair and reliable outputs.

AAP strongly believes voluntary self-regulation by the companies developing and promoting the deployment of artificial intelligence technologies will not suffice, nor ensure such companies comply consistently with applicable existing laws. As we have seen over the last three decades, companies engaged in developing "transformative" information technology have acted to within the letter of the law but refused to otherwise consider common sense regulation that would appropriately address societal concerns. It is essential agencies set a framework, guardrails, and oversight, with strong corrective tools, to ensure safe and ethical development of artificial intelligence technologies.

There are many issues robust transparency, safety, and accountability policies should address, including protecting privacy, preventing the use of generative AI models to promote misinformation or perpetrate fraud, and how to safeguard intellectual property.[2] In these comments, AAP outlines its priorities that should be included among those OSTP defines as essential to defining the U.S. government's AI development agenda.

**Protecting rights, safety, and national security:**

*Accurate Record Keeping; Assessments and Audits* — Developers of AI technology should be required to maintain accurate records as to (a) the types of content or materials used to create the training datasets ingested by AI foundation models; (b) whether the material is copyright protected, and if so, whether the use of this content is licensed and from whom; and (c) the assessment and audit processes conducted, including the benchmarks used to interrogate the AI algorithm or system as to its readiness for deployment.

This information is essential to the internal assessment process the AI developer undertakes as the first stress testing of the AI technology — for instance, to assess existence of bias in the system due to the nature of the data sets on which it was trained, or to identify the risks inherent in the system because of possible gaps in the training data set, or because of how the content used to create the training data sets was sourced. The results of the internal assessment, as well as the information underpinning the internal assessment, are critical to an external audit where independent third-party researchers can validate the earlier assessment, by examining the training datasets and the gaps identified in the training data sets or training processes, the risks arising from identified gaps, and the adequacy of the measures that the AI developer has taken to mitigate the risks identified. Assessments and audits should recur throughout the AI life cycle, from development to deployment, by the AI technology developer and the downstream entities that adopt AI technology.

---

[1] See ChatGPT is now invading websites with plugins — Expedia, Instacart and more | Tom's Guide (tomsguide.com), accessed June 2, 2023.

[2] See https://g7digital-tech-2023.go.jp/topics/pdf/pdf_20230430/ministerial_declaration_dtmm.pdf, accessed June 2, 2023.

There remain questions as to the regulatory framework within which assessments and audits are undertaken, the standards against which AI technologies are measured or evaluated, and the designated government agency that might be charged with oversight. However, as noted above, there should be no question regulatory oversight is necessary to promote responsible and ethical development of AI, and to ensure adequate risk mitigation.

*Information Disclosure* — Record keeping would, however, be of little utility absent a disclosure obligation, i.e., to provide pertinent information to the appropriate parties along the AI lifecycle. As noted above, for external audit purposes to ascertain whether an AI technology is safe and therefore ready to publicly deploy, there should be a requirement to disclose information regarding the nature of the training data set, how sourced (i.e., identification of the authorized licensor(s) and/or other source(s)), how the foundation model was trained, and how the AI technology "works" (for instance, how does the algorithm generalize and what does the algorithm predict on its own) to independent third-party auditors with the requisite technological skills to interrogate the AI system. The results of such an audit would be included in the information accompanying the release of AI technology to the business customers that may build the technology into its own product or service. As to a general disclosure obligation, an AI developer should be obliged to provide a summary of the copyright protected and other materials used to create the training datasets, the licensor(s) from whom the content is sourced, and identify the risks inherent in the AI system given the nature of the data set on which it was trained, and whether and how the known risks have been or are being mitigated.

*The Question of Training Data* — Questions regarding data quality are related to questions regarding intellectual property. AI systems developed or trained on data sets derived or created from authorized sources are more likely to yield reliable outputs than data sets obtained from illegal or pirated sources. It is essential to trustworthy and reliable AI that developers utilize high quality, curated content to create training data sets for their models. Scientific, technical, and medical journal publishers, and educational publishers already license their databases and education content to AI developers on reasonable terms, providing access to the valuable curated material on which to train trustworthy AI systems that yield verifiable and reliable outputs. In the case of AI training based on professional and scholarly communication, it is essential AI developers only use the Version of Record (VoR), with appropriate licensing. The VoR is the final, publisher-maintained article, updated and archived continually in consultation with the author. Accepted manuscripts, pre-prints, or illegally uploaded text versions of the article may be subject to post publication modification or retraction, which could create serious and cascading scientific or medical errors in AI generated outputs.

As standard practice, suppliers of data for training AI systems should be required to affirm that they either own the data or have secured the appropriate license or authorization to permit an AI developer to use such data. Organizations may be in possession of copyrighted material and other intellectual property with the potential to be used as training data, but mere possession does not and should not imply the authority to provide access to this content or intellectual property for text-and-data mining or AI training purposes.

Regulation or policy regarding training data should be careful to define "data" to ensure that proprietary information or content created by or generated by the private sector is inappropriately

categorized as "public" and made widely available in violation of legitimate rights. Specifically, care should be taken to ensure that "data" is not defined to include copyrighted works, such as books, journal articles, and other creative works developed for and created by private sector rights holders. The Open Government Data Act (OGDA), Public Law No: 115-435 (01/14/2019), is instructive with respect to how to establish a framework that facilitates access to "open government data assets," while also ensuring that such assets meet and comply with intellectual property rights protections. While data created for and by the government may properly be the subject of un-permissioned or uncompensated use for purposes of AI training, the use of data embodied in copyrighted works created for or by and owned by private sector actors is protected by copyright law, and its use should be the subject of licensing arrangements.

*The Role of Intellectual Property (terms of service, contractual obligations, or other legal entitlements in fostering or impeding a robust AI accountability ecosystem)* — The AI (or foundation) models that have gained recent notoriety were trained on large corpuses of text and images — much of it copyright protected material largely scraped from the internet.[3] Unfortunately, little if any of this use appears to have been permissioned or compensated. Responsible and ethical AI development must respect the intellectual property rights of the creators and owners of the copyright protected materials (text, images, and other material) used to create the data sets on which AI foundation models or AI systems are trained. Policy makers should ensure rightsholders have appropriate tools to hold AI and its sponsors accountable.

Securing the appropriate licenses from the copyright owner/rights holder to use their copyrighted material to create training datasets, as opposed to indiscriminate scraping of the entire internet including repositories of pirated materials, offers the greatest assurance that the AI technology is trained on materials that do not include incorrect, misleading, or retracted information, or inaccurate research conclusions, and therefore, is most likely to generate reliable and trustworthy outputs. The adage "garbage in, garbage out" is an apt predictor of what happens when the legitimacy of training datasets cannot be assured.

Given that AI technologies will be (and are being) integrated into applications that will impact the lives and well-being of individuals, whether financially, physically, mentally, or professionally, the importance of using high quality, peer reviewed, vetted material to create the training datasets cannot be overstated. As we have noted, AI technologies should be audited as to whether the material used to create the training data sets was legitimately sourced, and whether appropriately licensed from or its use authorized by the copyright owner or rights holder. Accountability policies that provide assurances that high quality materials are used in training an AI system build confidence and trust in the technology.

**Promoting economic growth and good jobs:**

Publishers have been developing and employing AI technologies in their businesses for at least a decade. For instance, intelligent search of scientific and academic publisher databases has long been powered by AI technology, allowing students, academics, and researchers to quickly find the information

[3] See the websites that make AI bots like ChatGPT sound so smart - Washington Post, accessed May 22, 2023; These artists found out their work was used to train AI. Now they're furious | CNN Business, accessed May 22, 2023.

they need. Education publishers, likewise, have been developing AI-enabled educational products and services that allow personalization of instructional materials that enhance not only student success but also the ability of teachers to assess and guide student progress. Publishers — as both users and developers of AI technologies, *and* curators of the high-quality content (such as books, journals, and other reading materials) are essential to developing trustworthy AI technology — recognize the promise of AI technologies but are also cognizant of the risks they pose.

The U.S. publishing industry supports an extensive network of American businesses and thousands of jobs, with revenue of $28.10 billion for 2022.[4] The publishing industry is also an integral part of the broader U.S. copyright industries, which collectively added more than $1.8 trillion in annual value to U.S. gross domestic product in 2021.[5] Beyond these important economic contributions, an independent and thriving publishing industry supports the nation's political, intellectual, and cultural systems. Publishers are incorporating AI responsibly into operations to boost outputs and create jobs and new opportunities in the creative industries.

**Innovation in public services:**

AI, if carefully and appropriately deployed, could create a transformational shift in improving the accuracy and reproducibility of reports about government supported medical and scientific information. Employed in concert with human based peer review, AI presents an opportunity to refine and improve scholarly communications funded by the public. To realize this benefit, it is imperative federal authorities incentivize a robust and competitive scholarly communication marketplace, with a wide array of publishers and business models. Empowering scientific authors and publishers should be a key goal of research agencies, and a core consideration of open science policies to ensure a robust marketplace of ideas and vetted science.

When considering open science and open access policies, it is important authors be given a broad suite of options to disseminate their work, including freedom to choose where to publish their data and manuscripts in a manner consistent with the goals of public access. While broad open licensing and data sharing mandates associated with the products of federal investment can be effective tools for dissemination, they can also exacerbate errors and misinformation by AI. Authors and publishers need tools to ensure proper and appropriate use of materials. Materials published under broad CC0, CC-BY, and similar licenses waive virtually all creator rights and are irrevocable. Moreover, materials distributed under broad open licenses may be manipulated, modified, and monetized in potentially harmful and damaging ways. To ensure generative AI does not engage or digest open licensed materials beyond the author and right holder's consent, all federal open science, data, and access policies must include options for non-derivative, non-commercial licenses.

---

[4] AAP StatShot Annual Report: Publishing Revenues Totaled $28.10 Billion for 2022 - AAP (publishers.org)

[5] *Copyright Industries in the U.S. Economy: The 2022 Report,* by Robert Stoner and Jéssica Dutra of Economists Incorporated, prepared for the International Intellectual Property Alliance (IIPA), (December 2022), https://www.iipa.org/files/uploads/2022/12/IIPA-Report-2022_Interactive_12-12-2022-1.pdf.

As previously mentioned, the VoR should be the only version of scientific literature for training AI. Government repositories should feature the VoR where feasible and apply labels and code to notify readers and AI that any non-VoR displayed accepted manuscripts are not the definitive version, as well as link to the publisher's website. In the case of AI, federal agencies should work with AI developers to direct their AI algorithms to publisher databases, seeking, as appropriate, licensed content to use as training material. This will assure the public that safe and reliable results of taxpayer funded research are used in training AI models.

**Conclusion**

Transparency — providing the public with sufficiently robust information as to how and on what content an AI system was trained — promotes accountability, in turn building the public's trust in AI technologies and AI-enabled products and services. Assessments and audits are likewise critical to a strong AI accountability framework. Interrogation of the AI algorithm, first by the AI developer to self-examine its AI technology for risks or defects, then by qualified independent third parties to verify the findings of an internal assessment process including whether risk mitigation has been sufficient, is necessary to build trustworthy AI.

Another key policy consideration addressed in these comments is the role of intellectual property in responsible and ethical AI development. Data is essential to developing and training AI technologies, and often that data is embodied in copyrighted works created and curated by the copyright owner or rights holder. Responsible and ethical AI development requires that the developer ensure that the material used to create training data sets is legitimately sourced, and appropriate licenses and permissions secured. An AI system trained on data sets created from legitimate and authentic content — whether this content be text, images, or other material — provides greater assurance that AI technology is more likely to generate reliable and trustworthy outputs.

Finally, publishers stand at the nexus point of reliable information to train generative AI, while also taking advantage of AI to continue to act as stewards of audio, visual, and text-based works, an honor the industry has had for over four hundred years. With careful consideration of the risks and strong federal oversight and enforcement, we believe AI can be a tool for good for the country and its people.

AAP appreciates the opportunity to provide its views on this important inquiry and looks forward to participating in further consultations on these critical issues.

Sincerely,

J. Carl Maxwell
Senior Vice President, Public Policy

*This document is approved for public dissemination. The document contains no business-proprietary or confidential information. Document contents may be reused by the government in developing the 2025 National AI R&D Strategic Plan and associated documents without attribution.*