PUBLIC SUBMISSION

Received: May 29, 2025
Tracking No. mb9-ky47-t6t1
Comments Due: May 28, 2025 **Submission Type:** Web

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-0199
Comment on FR Doc # 2025-07332

Submitter Information

Organization: News Corporation

General Comment

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.

Attachments

Comments of News Corporation - AI RD Strategic Plan

News Corp

May 29, 2025

Comments of News Corporation to the Office of Science and Technology Policy
Re: Development of a 2025 National Artificial Intelligence Research and Development Plan

News Corporation ("News Corp") submits these comments to the Office of Science and Technology Policy ("OSTP") and National Science Foundation in response to the request for information on the development of a 2025 National Artificial Intelligence ("AI") Research and Development ("R&D") Strategic Plan ("Strategic Plan").

As News Corp explained in its comments to the OSTP regarding its request for input on the development of an AI Action Plan, the U.S. lacks a clear policy for effectively protecting American made data from foreign adversaries. This leakage risks undermining U.S. AI competitiveness and national security because AI dominance requires the U.S. having the best access to key AI inputs: chips, energy, and data.

As with chips, data is an area where the U.S. should have a crucial advantage in the AI supply chain. Advancements in AI depend on access to vast amounts of high-quality data, however this is an area that is poorly understood because of the "black box" nature of commercial AI development. America cannot formulate a plan for dominating AI if there are gaps in understanding about supply chain components. As a first step then, the Strategic Plan should require research about the role that data plays across all stages of the AI life-cycle, addressing quantitative and qualitative matters such as, for example: how much data do American AI firms need access to for world leading AI; are there risks affecting continued data production—such as poor economic incentives—that may impede American AI leadership; what constitutes "quality" data; what added benefit does use of high-quality data by AI firms provide; and so on.

Second, the Strategic Plan should prioritize R&D about how to protect U.S. content from digital parasites and pirates, including foreign rivals. Chinese AI firms have identified accessing high-quality content as a means for circumventing export controls on chips. For example, 01.AI trained its AI model on mostly English-language content,[1] and both it and DeepSeek developed competitive models despite using older chips and less energy by prioritizing high-quality content.[2] China's race to access U.S. content assets explains why Chinese bots used are scraping the web at 25 times the rate of OpenAI and 3,000 times the rate of Anthropic.[3]

---

[1] 01.AI, *Yi: Open Foundation Models by 01.AI*, ARXIV (2024) https://arxiv.org/html/2403.04652v1.

[2] Eleanor Olcott, *Chinese AI groups get creative to drive down cost of models*, Financial Times (Oct 19, 2024) https://www.ft.com/content/0a6da1bb-2bda-40f3-9645-97877eb0947c?shareType=nongift ("Chinese AI players have been competing over the past year to develop the highest quality data sets"); 01.AI, above n 1 (2024) ("our data engineering principle is to promote quality over quantity for both pretraining and finetuning").

[3] Kali Hays, *TikTok's parent launched a web scraper that's gobbling up the world's online data 25 times faster than OpenAI*, Fortune (October 3, 2024) https://fortune.com/2024/10/03/bytedance-tiktok-bytespider-scraper-bot/.

Domestic firms are also leaking American made content to foreign rivals. California-based Common Crawl made copies of the American web, then marketed the content to third parties, including DeepSeek and 01.AI.[4] Both Chinese AI firms used Meta's Llama models, which were built using American made content without permission or compensation.[5] There is also a black-market trade in stolen American made content available to foreign competitors.[6]

News Corp invests heavily in protecting its digital content assets via state-of-the-art measures. However, technology firms are adopting increasingly sophisticated methods for accessing content, and many bots come from China. For example, despite News Corp's digital infrastructure investments, over a recent seven-day period, Huawei accessed one News Corp subsidiary's properties approximately 4.2 million times, and bots we attribute to China overall visited that subsidiary's websites almost 31 million times over the same seven days.

While the U.S. fails to protect its content assets, China protects its domestic content assets via the Great Firewall.[7] Where China may be on a path to achieve cyber sovereignty by protecting domestic content, America has no reciprocal policy.

Big Tech is unlikely to invest in data protection R&D because it is competing in a short-term arms race for acquiring content, failing to appreciate that its historical and ongoing theft of web content jeopardizes America's AI leadership by destroying the economic viability of content production—despite its necessity for AI—and means foreign rivals can compete by accessing the same content too. News publishers certainly do not have the resources to invest in such R&D.

President Trump has already laid the groundwork for the Strategic Plan accounting for protecting American content, emphasizing the need to stop foreign entities from "stealing our intellectual property".[8] Without decisive action to undertake research that gives effect to the President's prescient policy directive, the U.S. risks ceding its AI advantage to China.

---

[4] DeepSeek AI, *DeepSeek LLM Scaling Open-Source Language Models with Longtermism*, ARXIV (2024) https://arxiv.org/pdf/2401.02954v1, and 01.AI, above n 1 (2024), both acknowledging use of Common Crawl.

[5] DeepSeek AI, above n 4 (2024) ("we generally followed the architecture of LLaMA"); 01.AI, above n 1 (2024) ("where the [Yi model's] code is based on LLaMA").

[6] Such as "content contractors" that sell the underlying contents of a search index to AI firms.

[7] For example, China's 2021 Data Security Law imposes strict controls on data transfers outside China. See also: Stephanie Yang, *As China shuts out the world, internet access from abroad gets harder too*, Los Angeles Times (June 23, 2022) https://www.latimes.com/world-nation/story/2022-06-23/china-great-firewall-foreign-domestic-virtual-censorship ("academics … are finding it increasingly frustrating to penetrate China's cyber world").

[8] *Fact Sheet: President Donald J. Trump Encourages Foreign Investment While Protecting National Security*, The White House (February 21, 2025) https://www.whitehouse.gov/fact-sheets/2025/02/fact-sheet-president-donald-j-trump-encourages-foreign-investment-while-protecting-national-security/; *Artificial Intelligence for the American People*, Trump White House Archives https://trumpwhitehouse.archives.gov/ai/ ("The United States has long been a champion and defender of the core values of … respect for intellectual property".).