August 20, 2024

Ex Parte Meeting Concerning Class 4 – AI Research
Ninth Triennial Proceeding, Class 4
Docket No. 2023-5

Hacking Policy Council
Joint Academic Researchers

This memorandum summarizes the Aug. 14, 2024 ex parte meeting held between representatives of the U.S. Copyright Office, the Hacking Policy Council (HPC), and the Joint Academic Researchers regarding proposed Class 4 – AI Research – in the ninth triennial rulemaking proceeding. HPC was represented by Harley Geiger, and the Joint Academic Researchers were represented by Kevin Klyman and Shayne Longpre.

HPC and the Joint Academic Researchers support the proposed exemption for good faith AI research for trustworthiness purposes, not limited to security or safety. The meeting further clarified issues from the rulemaking record:

1. Technological protection measures.
2. Adverse effects.
3. Proposed exemption language.
4. Letters to the Copyright Office from U.S. officials.
5. The preponderance of the evidence.

Below we summarize the discussion.

1. Technological protection measures

During the meeting, we clarified the breadth of technological protection measures that are circumvented during the course of good faith AI trustworthiness research. We emphasized that mere terms of service are not a technological protection measure, and that loss of account access was only one of several technological protection measures that a researcher may encounter.

If the measures identified by the Hacking Policy Council and the Joint Academic Researchers are not "technological protection measures" within the meaning of the statute, we would welcome explicit clarification from the Copyright Office regarding this interpretation.

a. The role of terms of service

As discussed during the hearing, terms of service are not a technological protection measure. In this context, the role of terms of service is to confer authorization for use of software. Actions that contravene terms of service, such as independent good faith AI trustworthiness research, are often the basis for the imposition of technological protection measures that restrict the researcher's access to the protected works.

The proposed exemption for AI trustworthiness research would not change terms of service, nor the ability of any party to apply terms or technological protection measures to protected works.

b. Multiple TPMs at issue

As noted during the meeting, the statute defines technological protection measures controlling access to a protected work as a measure that requires the application of information, or a process or treatment, with the authority of the copyright owner to gain access to the work.[1] The statute further defines "circumvent a technological measure" to include merely avoiding, bypassing, deactivating, or impairing the technological measure without the authority of the copyright owner.[2] The record specifies several such measures that are circumvented during the course of AI research.

During the meeting, the Joint Academic Researchers reiterated several technological protection measures that may be circumvented as part of good faith AI trustworthiness research; such technological protection measures include loss of account access, but encompass other measures as well. Other technological protection measures the Joint Academic Researchers reiterated from comments and the hearing included:

i. Blocking model outputs (e.g. via a safety classifier or guardrails)
ii. Blocking user inputs or prompts (e.g., via a filter in the user interface)
iii. Account rate limits
iv. Limiting access to model or system outputs[3]

These TPMs are not theoretical, but are encountered by researchers in the field. The comments from both the Joint Academic Researchers and the Hacking Policy Council include numerous footnotes with

[1] 17 USC 1201(a)(3)(B).

[2] 17 USC 1201(a)(3)(A).

[3] Reply Comments of Joint Academic Researchers, Mar. 19, 2024, pgs. 5-7, https://www.copyright.gov/1201/2024/comments/reply/Class%204%20-%20Reply%20-%20Kevin%20Klyman%20et%20al.%20(Joint%20Academic%20Researchers).pdf.

specific examples of researchers encountering such TPMs.[4]

2. Adverse effects

During the meeting, HPC and the Joint Academic Researchers further clarified the existing record detailing the adverse effect of the lack of an exemption on research. The record indicates that noninfringing AI trustworthiness research is, or is likely to be, adversely affected by the prohibition against circumvention in the succeeding three-year period.[5]

We highlighted substantial adverse effects experienced by both individuals and the broader marketplace on good faith AI trustworthiness research. We noted the Copyright Office did not require litigation or cease-and-desist orders to demonstrate adverse effects for related exemptions in any previous rulemaking cycles.[6]

a. Broad marketplace chilling effect

The Joint Academic Researchers' comment included an open letter signed by more than 350 AI researchers and professionals. The letter stated that independent researchers feared legal risks due to trustworthiness research, and that these risks have chilling effects on trustworthiness research, due to a lack of legal protection on par with security research. This is a reference to, among other things, the security research exemption under DMCA Section 1201.[7] The large number of signatories demonstrates broad marketplace concern and high likelihood of chilling effect due to the lack of clear legal protection for good faith trustworthiness research. The letter further expresses support for a peer-reviewed position paper written by a number of the Joint Academic Researchers which expresses concern about

[4] See, for example, Reply Comments of the Hacking Policy Council, Mar. 19, 2024, footnote 24, page 5: "Note that we only conducted a small experiment due to the rate limit and account suspension risks upon repeated jailbreak attempts." Deng, Liu, Li, et. al, Masterkey: Automated Jailbreaking of Large Language Model Chatbots, Network and Distributed System Security Symposium, Feb. 26, 2024, pg. 13, https://www.ndss-symposium.org/wp-content/uploads/2024-188-paper.pdf.

See also, for example, Reply Comments of the Hacking Policy Council, footnote 14, page 4: "[I]n response to the jailbreak threat, service providers have deployed a variety of mitigation measures. These measures aim to monitor and regulate the input and output of LLM chatbots, effectively preventing the creation of harmful or inappropriate content. [The] black-box nature of these services, especially their defense mechanisms, poses a challenge to comprehending the underlying principles of both jailbreak attacks and their preventive measures. As of now, there is a noticeable lack of public disclosures or reports on jailbreak prevention techniques used in commercially available LLM-based chatbot solutions." Deng, Liu, Li, et. al, Masterkey: Automated Jailbreaking of Large Language Model Chatbots, Network and Distributed System Security Symposium, Feb. 26, 2024, pg. 13, https://www.ndss-symposium.org/wp-content/uploads/2024-188-paper.pdf.

[5] 17 USC 1201(a)(1)(B).

[6] See, e.g., Copyright Office, Sixth Triennial Proceeding, Register's Recommendation, Oct. 2015, pg. 305, https://cdn.loc.gov/copyright/1201/2015/registers-recommendation.pdf.

[7] Reply Comments of Joint Academic Researchers, Mar. 19, 2024, pg. 32, https://www.copyright.gov/1201/2024/comments/reply/Class%204%20-%20Reply%20-%20Kevin%20Klyman%20et%20al.%20(Joint%20Academic%20Researchers).pdf.

liability under DMCA section 1201;[8] as the Joint Academic Researchers stated in their comments, "Our paper is based on the experiences of AI safety and security researchers of the chilling effect of potential legal liability if they attempt to bypass account restrictions and other technological protection measures."[9]

b. Individual adverse effects

In comments and during the hearings, the Hacking Policy Council and Joint Academic Researchers referenced that specific individuals experienced adverse effects to AI research due to lack of protection under DMCA Section 1201. As one example from the existing rulemaking record, the Joint Academic Researchers stated in their comment:

> "*In one case, a model owner banned an independent researcher's account after they claimed that a generative AI model readily creates copyrighted images, something they discovered in the course of their research. The model owner also banned the accounts that the researcher subsequently created and changed its terms to state 'If You knowingly infringe someone else's intellectual property, and that costs us money, we're going to come find You and collect that money from You. We might also do other stuff, like try to get a court to make You pay our legal fees.' The threat of legal liability for circumventing access restrictions imposed on research that is fair use as a result of terms of service violations is an example of the need for safe harbor under Section 1201*."[10]

3. Proposed exemption language

During the meeting, HPC and the Joint Academic Researchers addressed questions from the hearing related to the exemption language proposed by HPC.[11] We discussed the alignment of the proposed language with recognized standards, best practices, regulations, and DMCA specificity requirements.

For ease of reference, the proposed language is provided in Addendum I to this memorandum.

a. Trustworthiness

HPC's proposed exemption language uses the recognized term "trustworthiness" to encompass both security and non-security research objectives. Accordingly, the term covers research into prevention of AI bias, discrimination, infringement, synthetic content, and other potentially harmful outcomes.

[8] Reply Comments of Joint Academic Researchers, Mar. 19, 2024, pg. 19, https://www.copyright.gov/1201/2024/comments/reply/Class%204%20-%20Reply%20-%20Kevin%20Klyman%20et%20al.%20(Joint%20Academic%20Researchers).pdf.

[9] Reply Comments of Joint Academic Researchers, Mar. 19, 2024, pg. 12, https://www.copyright.gov/1201/2024/comments/reply/Class%204%20-%20Reply%20-%20Kevin%20Klyman%20et%20al.%20(Joint%20Academic%20Researchers).pdf.

[10] *Id.*, pg. 8.

[11] See Reply Comments of the Hacking Policy Council, Mar. 19, 2024, pg. 7, https://www.copyright.gov/1201/2024/comments/reply/Class%204%20-%20Reply%20-%20Hacking%20Policy%20Council.pdf.

This terminology is consistent with key federal guidance and international standards – the NIST AI Risk Management Framework and ISO/IEC TS 5723:2022, respectively.[12] Such use of the term "trustworthiness" is also consistent with multiple White House Executive Orders[13] and other federal agency actions.[14] Such use of the term is also consistent with guidance from international and intergovernmental bodies such as OECD.[15]

b. AI definitions

HPC's proposed exemption language provides definitions for "artificial intelligence" or "AI," and "AI system." These definitions are consistent with current U.S. law and Executive Orders – see, e.g., 15 U.S.C. 9401(3) and EO 14110.

These definitions are also used in key federal guidance such as the NIST AI Risk Management Framework, as well as international standards such as ISO/IEC 22989:2022.[16] These definitions are also similar to guidance from intergovernmental bodies such as OECD,[17] as well as non-US laws such as the EU AI Act.[18]

c. DMCA specificity requirements

HPC's proposed language would apply to a particular class of works: computer programs on devices or machines on which an AI system operates. The Copyright Office has previously recognized that

[12] National Institute of Standards and Technology, Artificial Intelligence Risk Management Framework, Second Draft, Aug. 18, 2022, pgs. 10-11, https://www.nist.gov/system/files/documents/2022/08/18/AI_RMF_2nd_draft.pdf.

See also, International Standards Organization, ISO/IEC TS 5723:2022, Jul. 2022, https://www.iso.org/obp/ui/#iso:std:iso-iec:ts:5723:ed-1:v1:en.

[13] White House, Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence, Section 3(b), Oct. 30, 2023, www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy development-and-use-of-artificial-intelligence.

See also, White House, Executive Order on Promoting the Use of Trustworthy Artificial Intelligence in the Federal Government, Section 3, Dec. 3, 2020, https://trumpwhitehouse.archives.gov/presidential-actions/executive-order-promoting-use-trustworthy-artificial-intelligence-federal-government.

[14] See, e.g., U.S. Dept. of Health and Human Services, Trustworthy AI Playbook, Sep. 2021, https://www.hhs.gov/sites/default/files/hhs-trustworthy-ai-playbook.pdf.

[15] OECD Recommendations on AI (Amended), May 2, 2024, Section 1, https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449.

[16] International Standards Organization, ISO/IEC 22989:2022, Jul. 2022, Section 3, https://www.iso.org/obp/ui/en/#iso:std:iso-iec:22989:ed-1:v1:en.

[17] OECD Recommendations on AI (Amended), May 2, 2024, https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449.

[18] EU AI Act., OJ L 2024/1689, Jul. 12, 2024, Art. 3(1), https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:52021PC0206.

computer programs are a subcategory of literary works.[19] HPC's proposed class of works is narrower than other current exemptions.[20]

d. Safeguards

HPC's proposed exemption language incorporates several key safeguards designed to help ensure good faith AI trustworthiness research is carried out responsibly and maximizes public benefit. These safeguards are identical to provisions present in current exemptions to DMCA Section 1201. They include:

i. Lawfully acquired device or machine

ii. Solely for the purpose of good-faith research

iii. Carried out in an environment designed to avoid any harm

iv. Information derived from the activity is not used or maintained in a manner that facilitates infringement

v. Clarification that the exemption does not provide defense for liability under other laws[21]

4. Letters to the Copyright Office from U.S. officials.

During the meeting, HPC and the Joint Academic Researchers highlighted key issues raised by the letters to the Copyright Office from other U.S. officials in support of an exemption for AI trustworthiness research.

a. Letter from the Department of Justice

The letter from the Computer Crime and Intellectual Property Section of the Department of Justice (DOJ) unequivocally supports the proposed exemption for good faith AI trustworthiness research:

> "*Like good-faith computer security research that might circumvent technological protection measures used to protect copyrighted works but does not result and is not intended to result in infringement, CCIPS believes that good faith research on potentially harmful outputs of AI and*

[19] Copyright Office, Seventh Triennial Proceeding, Recommendation of the Acting Register of Copyrights, Oct. 2018, pgs. 289, https://cdn.loc.gov/copyright/1201/2018/2018_Section_1201_Acting_Registers_Recommendation.pdf.

[20] See, e.g., 37 CFR 201.40(b)(16).

[21] See 37 CFR 201.40(b)(16)(ii)-(iii).

> *similar algorithmic systems should be similarly exempted from the DMCA's circumvention provisions."*[22]

The DOJ letter also clearly acknowledges that such research may not be protected by existing exemptions to Section 1201, and recommends an exemption to protect such research:

> *"While the existing exemption for computer security research covers many types of research focused on the security and integrity of AI models, we recognize that it may not be sufficiently broad in its current form to exempt research that falls outside of "security" concerns. Therefore, we recommend that the Copyright Office consider clarifying the existing exemption to ensure its application to good-faith security research regarding AI systems and other, similar, algorithmic models, but also consider how best to clarify or amend the existing exemptions to cover good-faith research into bias and other harmful and unlawful outputs of such systems."*[23]

b. Letter from Senator Warner

Senator Warner's letter calls on the Copyright Office to ensure any exemption includes clear indicia of good faith.[24] We agree, and it is for this reason that HPC's proposed exemption language includes the safeguards outlined above in subsection 3.d of this memorandum.

The letter from Senator Warner further supports protection of good faith AI trustworthiness research under DMCA Section 1201:

> *"I urge the Copyright Office to consider expanding the existing good-faith security research exemption to cover both security and safety flaws or vulnerabilities, where safety includes bias and other harmful outputs. [...] This research into bias and other harmful outputs is essential to ensuring public safety and equity while enabling continued innovation, public trust, and adoption of AI. Therefore, it is crucial that we allow researchers to test systems in ways that demonstrate how malfunctions, misuse, and misoperation may lead to an increased risk of physical or psychological harm."*[25]

5. The preponderance of the evidence

During the meeting, we discussed the totality of the rulemaking record as reflecting the need for an exemption by a preponderance of the evidence. On the whole, the evidence shows that it is more likely

[22] Letter from U.S. Department of Justice Computer Crime and Intellectual Property Section, Apr. 15, 2024, pg. 4, https://www.copyright.gov/1201/2024/USCO-letters/Letter%20from%20Department%20of%20Justice%20Criminal%20Division.pdf.

[23] *Id.*

[24] Letter from Senator Mark Warner, May 24, 2024, pg. 2, https://www.copyright.gov/1201/2024/USCO-letters/Senator%20Warner%20DMCA%20AI%20Exemption%20Letter%20-%2024%20May%202024.pdf.

[25] *Id.*, at pg. 1.

than not that independent researchers will, in the succeeding three-year period, be adversely affected by the prohibition on circumvention in their ability to perform noninfringing good faith AI trustworthiness research.[26]

* * *

HPC and the Joint Academic Researchers thank the Copyright Office for the meeting. Please let us know if we can be of further assistance.

[26] U.S. Copyright Office, Section 1201 of Title 17 Report, Jun. 2017, pg. 8, https://www.copyright.gov/policy/1201/section-1201-full-report.pdf.

ADDENDUM I: PROPOSED EXEMPTION LANGUAGE

Below is the exemption language proposed by the Hacking Policy Council:[27]

i. Computer Programs, where the circumvention is undertaken on a lawfully acquired device or machine on which an AI system operates, or is undertaken on a computer, computer system, or computer network on which an AI system operates with the authorization of the owner or operator of such computer, computer system, or computer network, solely for the purpose of good-faith AI trustworthiness research.

ii. For purposes of paragraph (i), "good-faith AI trustworthiness research" means accessing a computer program solely for purposes of good-faith testing or investigation of bias, discrimination, infringement, or harmful outputs in an AI system, where such activity is carried out in an environment designed to avoid any harm to individuals or the public, and where the information derived from the activity is used primarily to promote the trustworthiness of the AI system, and is not used or maintained in a manner that facilitates copyright infringement.

iii. For purposes of paragraph (i), the term "artificial intelligence" or "AI" has the meaning set forth in 15 U.S.C. 9401(3): a machine-based system that can, for a given set of human-defined objectives, make predictions, recommendations, or decisions influencing real or virtual environments.

iv. For purposes of paragraph (i), the term "AI system" means any data system, software, hardware, application, tool, or utility that operates in whole or in part using AI.

v. Good-faith AI trustworthiness research that qualifies for the exemption of this section may nevertheless incur liability under other applicable laws, including without limitation the Computer Fraud and Abuse Act of 1986, as amended and codified in title 18, United States Code, and eligibility for that exemption is not a safe harbor from, or defense to, liability under other applicable laws.

* * *

[27] See Reply Comments of the Hacking Policy Council, Mar. 19, 2024, https://www.copyright.gov/1201/2024/comments/reply/Class%204%20-%20Reply%20-%20Hacking%20Policy%20Council.pdf.