UNITED STATES COPYRIGHT OFFICE

Long Comment Regarding a Proposed Exemption Under 17 U.S.C. § 1201

Hacking Policy Council reply comments
Ninth Triennial Proceeding, Class 4

March 19, 2024

Item A. Commenter Information.

The Hacking Policy Council is a group of experts dedicated to advancing good faith security research, AI red teaming, penetration testing, independent repair for security, and vulnerability disclosure and management.[1] In this proceeding, the Hacking Policy Council is represented by Harley Geiger, Counsel, Venable LLP.

Item B. Proposed Class Addressed.

Our reply comment follows up on our initial comments in support of the petition for a proposed exemption under Section 1201 of the Digital Millennium Copyright Act (DMCA) for Class 4: Computer Programs–Generative AI Research.[2]

The Copyright Office's Notice of Proposed Rulemaking states that the proposed classes are subject to

[1] Hacking Policy Council, https://hackingpolicycouncil.org.

[2] Hacking Policy Council comments, Ninth Triennial Section 1201 Proceeding, United States Copyright Office, Dec. 21, 2023, https://www.copyright.gov/1201/2024/comments/Class%204%20-%20Initial%20Comments%20-%20Hacking%20Policy%20Council.pdf.

further refinement based on the record.[3] In addition, the Copyright Office's form for a petition for a new exemption states that the petitioner need not "fully define the contours of an exemption class," and that "proponents will have the opportunity to further refine or expound upon their initial petitions during later phases of the rulemaking."[4] As noted in our initial comments, we believe that the petitioner did not fully define the contours of the proposed exemption, and that the Copyright Office should consider the exemption to apply to artificial intelligence (AI) trustworthiness research – which encompasses bias, discrimination, synthetic content, infringement, and other alignment issues not directly related to security.[5] As discussed below, the field of independent AI trustworthiness research faces adverse effects because of the absence of an exemption under DMCA Section 1201.

The proposed exemption would apply to a particular class of works: computer programs, which are a subcategory of literary works.[6] The proposed exemption would apply to a specific set of users: persons performing good faith research, as defined, under certain conditions. These are the same parameters that the Copyright Office uses to describe other classes of works and sets of users in existing exemptions.[7]

Accordingly, we request that the Copyright Office recommend that the Librarian of Congress adopt a new exemption to protect AI research, and that the Copyright Office clarify the extent to which AI research is already protected by existing DMCA Section 1201 exemptions.

Item C. Overview.

The Hacking Policy Council (HPC) supports the petition to protect independent testing of AI for bias and alignment ("trustworthiness"[8]) because we believe such testing is crucial to identifying and fixing algorithmic flaws to prevent harm or disruption. We further assert that good faith independent AI testing benefits both the public and the copyright system. However, technological access barriers and a lack of clear legal protection under DMCA Section 1201 adversely affect such research.[9]

[3] Copyright Office, Notice of proposed rulemaking, Exemptions to Permit Circumvention of Access Controls on Copyrighted Works, 88 F.R. 72013, 72024, Oct. 19, 2023.

[4] Jonathan Weiss, Petition for New Exemption Under 17 USC 1201, Copyright Office,9th Triennial Rulemaking, https://www.copyright.gov/1201/2024/petitions/proposed/New-Pet-Jonathan-Weiss.pdf (last accessed Mar. 13, 2023).

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

[6] 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.

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

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

[9] Hacking Policy Council, AI red teaming - Legal clarity and protections needed, Dec. 12, 2023, https://assets-global.website-files.com/62713397a014368302d4ddf5/6579fcd1b821fdc1e507a6d0_Hacking-Policy-Council-statement-on-AI-red-teaming-protections-20231212.pdf.

This reply comment focuses on three areas: 1) Asserted adverse effects on noninfringing uses; 2) Technological protection measures and methods of circumvention; and 3) Exemption language.

Item D. Asserted Adverse Effects on Noninfringing Uses.

The preponderance of the evidence demonstrates that the absence of a clear exemption under DMCA Section 1201 for AI research adversely affects noninfringing uses, and is likely to adversely affect noninfringing uses in the 3-year period following this proceeding.[10] In recognition of these adverse effects, more than 350 AI researchers and experts recently called for a legal safe harbor for good faith AI testing. In doing so, the researchers note the gap in legal protections for AI bias and trustworthiness research as compared to good faith security testing, which is protected under DMCA Section 1201.[11]

Independent evaluators of AI systems allege that they have been penalized for good faith research, and express concern about liability as a direct consequence of prohibitions under DMCA Section 1201. Researchers further state that a legal safe harbor from both governments and private sector entities is necessary to mitigate these adverse effects and help provide protection for AI testing under, specifically, DMCA Section 1201.[12]

As an example, in 2024, independent good faith researchers claim to have tested generative AI services for outputs that potentially infringed copyright, and alleged they were suspended from the service after reporting their results. The researchers alleged that the service then changed its terms to threaten legal action for similar behavior, making no distinction between malicious terms of service violations and good faith research.[13]

[10] 17 USC 1201(a)(1)(C). *See also* Copyright Office, Eighth Triennial Proceeding, Recommendation of the Register of Copyrights, Oct. 2021, pg. 8, https://cdn.loc.gov/copyright/1201/2021/2021_Section_1201_Registers_Recommendation.pdf. "Such evidence must, on the whole, show that it is more likely than not that users of a copyrighted work will, in the succeeding three-year period, be adversely affected by the prohibition on circumvention in their ability to make noninfringing uses of a particular class of copyrighted works."

[11] Open Letter, A Safe Harbor for AI Evaluation and Red Teaming, Massachusetts Institute of Technology, https://sites.mit.edu/ai-safe-harbor (last accessed Mar. 13, 2024). "Independent evaluators fear account suspension (without an opportunity for appeal) and legal risks, both of which can have chilling effects on research."

[12] Longpre, Kapoor, Klyman, et. al, A Safe Harbor for AI Evaluation and Red Teaming, Mar. 5, 2024, pgs. 6-7, https://bpb-us-e1.wpmucdn.com/sites.mit.edu/dist/6/336/files/2024/03/Safe-Harbor-0e192065dccf6d83.pdf. "The legal safe harbor [...] would safeguard certain research from some amount of legal liability, mitigating the deterrent of strict terms of service and the threat that researchers' actions could spark legal action by companies (e.g. under US laws such as the CFAA or DMCA Section 1201)." *See also*, Longpre et. al, A Safe Harbor for AI Evaluation and Red Teaming, Knight First Amendment Institute at Columbia University, Mar. 5, 2024, https://knightcolumbia.org/blog/a-safe-harbor-for-ai-evaluation-and-red-teaming. "A legal safe harbor could provide assurances that AI companies will not sue researchers if their actions were taken for research purposes. In the U.S. legal regime, this would impact companies' use of [...] Section 1201 of the Digital Millennium Copyright Act (DMCA)."

[13] Gary Marcus and Reid Southen, Generative AI Has a Visual Plagiarism Problem, IEEE Spectrum, Jan. 6, 2024, https://spectrum.ieee.org/midjourney-copyright.

Independent researchers further claim technological protection measures controlling access to AI systems have a chilling effect on the field of research as a whole, to the detriment of the public.[14] Researchers point to potential liability under DMCA Section 1201, if those measures are circumvented, as having a chilling effect on independent AI research and responsible disclosure of algorithmic flaws.[15]

Item E. Technological Protection Measures and Methods of Circumvention.

HPC's initial comments identified several technological protection measures (TPMs) that may be circumvented in the course of good faith research into algorithmic flaws that produce bias, misalignment, or harmful content in generative AI systems.[16] Below, we provide additional detail on three of these TPMs: 1) Account requirements; 2) Rate limits; and 3) Algorithmic safeguards.

If the Copyright Office does not consider these measures to be technological measures that effectively control access to copyrighted works within the meaning of 17 USC 1201(a)(3), we request that the Copyright Office provide an affirmative declaration of this conclusion for clarity.

1. <u>Account requirements</u>

As noted in HPC's initial comments, and as highlighted by AI researchers, the copyright owner of the AI system often requires a user account to access the system.[17] An account login requirement is a technological protection measure that effectively controls access to computer programs, including AI systems. In many instances, the terms of the account prohibit generating undesirable content, bypassing any algorithmic protective measures (i.e., "guardrails"), or the use of automated tools as a condition for permission to access the system.[18] In addition, the terms often restrict users to one account per user. However, the terms generally do not distinguish good faith AI research from other terms of service violations.

[14] 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. "[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 preventative 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."

[15] Longpre, Kapoor, Klyman, et. al, A Safe Harbor for AI Evaluation and Red Teaming, Mar. 5, 2024, pg. 4-7, https://bpb-us-e1.wpmucdn.com/sites.mit.edu/dist/6/336/files/2024/03/Safe-Harbor-0e192065dccf6d83.pdf..

[16] Hacking Policy Council comments, Ninth Triennial Section 1201 Proceeding, United States Copyright Office, Dec. 21, 2023, pg. 3, https://www.copyright.gov/1201/2024/comments/Class%204%20-%20Initial%20Comments%20-%20Hacking%20Policy%20Council.pdf.

[17] Longpre, Kapoor, Klyman, et. al, A Safe Harbor for AI Evaluation and Red Teaming, Mar. 5, 2024, pgs. 2-5, https://bpb-us-e1.wpmucdn.com/sites.mit.edu/dist/6/336/files/2024/03/Safe-Harbor-0e192065dccf6d83.pdf.

[18] *Id.*, pg. 5.

This TPM may be circumvented when a prospective user establishes an account to conduct independent AI research, despite restrictions on such activity in the account terms of service. This TPM may also be circumvented when a user's account has been suspended for conducting generative AI research, and the same user then establishes a new account to conduct additional generative AI research, as occurred during the Marcus and Southen research.[19]

This can result in an AI researcher facing penalties for violating terms of service by conducting good faith research. Penalties include, but are not limited to, access restrictions such as account suspension and user bans (blanket bans on email addresses, IP addresses, and credit cards), as well as legal liability.[20] As noted above, AI researchers cite potential legal liability under DMCA Section 1201 for circumventing account requirements as having an adverse effect on independent AI research.[21]

2. Rate limits

Rate limiting is a common technological measure for controlling access to a system if the user repeats the same action (such as providing input to the AI system) rapidly within a set period of time. Some generative AI bias and trustworthiness research relies on a high volume of inputs, sometimes automated, to identify and statistically validate algorithmic flaws.[22] Rate limits may be bypassed through such processes as IP address rotation and automated backoff measures, or by establishing another account.[23] However, exceeding rate limits or using automated tools without authorization can halt inputs or result in account suspension or other reprisals, which has had an adverse effect on independent generative AI research.[24]

3. Algorithmic safeguards

Algorithmic safeguards, commonly called "guardrails," are key technological measures designed to restrict AI systems from providing biased, unethical, inaccurate, or unsafe output, and to help ensure the AI system is trustworthy and in alignment with human values.[25] Guardrails are also customizable

[19] Gary Marcus and Reid Southen, Generative AI Has a Visual Plagiarism Problem, IEEE Spectrum, Jan. 6, 2024, https://spectrum.ieee.org/midjourney-copyright.

[20] Longpre, Kapoor, Klyman, et. al, A Safe Harbor for AI Evaluation and Red Teaming, Mar. 5, 2024, pg. 5, https://bpb-us-e1.wpmucdn.com/sites.mit.edu/dist/6/336/files/2024/03/Safe-Harbor-0e192065dccf6d83.pdf. "The cost of suspensions without refunds quickly tallies to hundreds of dollars, and creating new accounts is also not trivial, with blanket bans on credit cards and email addresses."

[21] *Id.*, pgs. 4-6.

[22] Ge et. al, MART: Improving LLM Safety with Multi-round Automatic Red-Teaming, Nov. 13, 2023, https://arxiv.org/pdf/2311.07689.pdf.

[23] Rollbar, How to Resolve ChatGPT Limit Errors, Jul. 31, 2023, https://rollbar.com/blog/chatgpt-api-rate-limit-error.

[24] 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. "Note that we only conducted a small experiment due to the rate limit and account suspension risks upon repeated jailbreak attempts."

[25] Rebedea, Dinu, et. al, NeMo Guardrails: A Toolkit for Controllable and Safe LLM Applications with Programmable Rails, Oct. 16, 2023, https://arxiv.org/pdf/2310.10501.pdf. *See also* Turing, Implementing Security Guardrails for LLMs, https://www.turing.com/resources/implementing-security-guardrails-for-llms (last accessed Mar. 13, 2024).

and may be arbitrarily imposed to automatically block certain inputs or filter output. Guardrails may be removed for certain users, such as internal software developers authorized by the AI system owner, so that those users can access features and outputs that would ordinarily be inaccessible to regular users.

Many techniques for generative AI trustworthiness research involve identifying methods for bypassing guardrails.[26] One such technique is the use of "jailbreak prompts." A jailbreak prompt is input designed to bypass generative AI system guardrails so that the system does something it was programmed to avoid, such as generating biased, discriminatory, infringing, or harmful output.[27] This may include, for example, injecting code into a prompt, or elevating a researcher's user privileges so that the researcher has fewer restrictions on their use of the AI system.[28]

However, the terms of many generative AI system providers restrict bypassing guardrails and do not distinguish between good faith research and misuse by bad actors.[29] As a result, bypassing guardrails without authorization may lead to account suspension or other reprisals, which has had an adverse effect on independent generative AI research.[30]

Item F. Exemption language.

As emphasized in HPC's initial comments, the purpose of good faith AI research is to identify, assess, and correct algorithmic flaws and thereby help strengthen the trustworthiness of AI systems.[31] Such research is fair use, contributes to the advancement of computer science, and leads to the production of new creative works.[32] We note there is no evidence that the proposed exemption would result in increased copyright infringement or piracy.

[26] Kang et. al, Exploiting Programmatic Behavior of LLMs: Dual-Use Through Standard Security Attacks, Feb. 11, 2023, https://arxiv.org/pdf/2302.05733.pdf.

[27] Zou et. al, Universal and Transferable Adversarial Attacks on Aligned Language Models, Dec. 20, 2023, https://arxiv.org/pdf/2307.15043.pdf.

[28] Shen et. al, "Do Anything Now": Characterizing and Evaluating In-The-Wild Jailbreak Prompts on Large Language Models, Aug. 7, 2023, pgs. 1-3, 6, https://arxiv.org/pdf/2308.03825.pdf.

[29] Longpre, Kapoor, Klyman, et. al, A Safe Harbor for AI Evaluation and Red Teaming, Mar. 5, 2024, pgs. 4-5, https://bpb-us-e1.wpmucdn.com/sites.mit.edu/dist/6/336/files/2024/03/Safe-Harbor-0e192065dccf6d83.pdf.

[30] Deng, Liu, Li, et. al, Masterkey: Automated Jailbreaking of Large Language Model Chatbots, Network and Distributed System Security Symposium, Feb. 26, 2024, pg. 4, https://www.ndss-symposium.org/wp-content/uploads/2024-188-paper.pdf. "[W]e observe that repeated unsuccessful jailbreak attempts [result] in account suspension, making it infeasible to conduct extensive trial experiments."

[31] Hacking Policy Council comments, Ninth Triennial Section 1201 Proceeding, United States Copyright Office, Dec. 21, 2023, pg. 4, https://www.copyright.gov/1201/2024/comments/Class%204%20-%20Initial%20Comments%20-%20Hacking%20Policy%20Council.pdf.

[32] Copyright Office, Eighth Triennial Proceeding, Recommendation of the Register of Copyrights, Oct. 2021, pg. 303, https://cdn.loc.gov/copyright/1201/2021/2021_Section_1201_Registers_Recommendation.pdf. "In prior rulemakings, the Office has consistently found that exemptions to allow noninfringing analysis of computer programs are likely to promote the availability of copyrighted works."

HPC proposes an exemption for AI research that is modeled closely on existing exemption language approved by the Copyright Office and leverages existing definitions in US statutes and Executive Orders. The exemption is drafted to promote public benefit and prevent misuse. The proposed exemption would apply to a particular class of works: computer programs on devices or machines on which an AI system operates.[33] The proposed exemption would apply to a specific set of users: persons performing good-faith AI trustworthiness research, as defined, under certain conditions. These are the same parameters that the Copyright Office uses to describe other classes of works and sets of users in existing exemptions.[34]

Accordingly, HPC requests that the Librarian provide an exemption for the following class:

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.[35]

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.[36]

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.[37]

[33] 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. “The Acting Register agrees that the requested expansion appropriately describes a particular class within the meaning of the statute. As proponents correctly note, computer programs are a subcategory of literary works.”

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

[35] *Id.* at 201.40(b)(16)(i). *See also* National Institute of Standards and Technology, Artificial Intelligence Risk Management Framework, Second Draft, Aug. 18, 2022, pgs. 10-12, https://www.nist.gov/system/files/documents/2022/08/18/AI_RMF_2nd_draft.pdf.

[36] *Id.* at 201.40(b)(16)(ii).

[37] 15 USC 9401(3). *See also* 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.

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.[38]

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.[39]

* * *

Thank you for your consideration. If we can be of additional assistance, please contact Harley Geiger, coordinator of the Hacking Policy Council, at hgeiger@venable.com.

[38] *Id.* at Section 3(e).

[39] 37 CFR 201.40(b)(16)(iii).