HACKING POLICY COUNCIL

Hacking Policy Council comments – Ninth Triennial Proceeding, Class 4

December 21, 2023

Item A. Commenter Information.

The Hacking Policy Council is a group of experts dedicated to advancing good faith security research, 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 comment supports the petition for a newly proposed exemption for Class 4: Computer Programs–Generative AI Research.[2]

Item C. Overview.

The Hacking Policy Council (HPC) supports the proposed exemption for generative artificial intelligence (AI) research under Section 1201 of the Digital Millennium Copyright Act (DMCA).[3] The exemption for generative AI research should not be limited to a narrow definition of "bias," but should encompass

[1] Hacking Policy Council, https://hackingpolicycouncil.org, (last accessed Dec. 21, 2023).
[2] 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 Dec. 12, 2023).
[3] Copyright Office, Notice of proposed rulemaking, Exemptions to Permit Circumvention of Access Controls on Copyrighted Works, 88 F.R. 72013, 72025, Oct. 19, 2023.

discrimination, and other harmful or undesirable outputs in AI systems.[4] This is more consistent with industry practices for AI red teaming, as well as the national priorities articulated in Executive Order 14110.[5]

By identifying and disclosing flaws so that they can be corrected, AI alignment research and AI red teaming are beneficial practices to help ensure the safety, trustworthiness, and fairness of generative AI systems. However, the 17 U.S.C. 1201(a)(1)(A) prohibition on circumvention of technological measures that control access to computer programs can restrict independent AI alignment research if permission of the computer program copyright holder is required to conduct such research.

HPC encourages the Register of Copyrights to clarify that some forms of research pertaining to AI bias and misalignment are already exempt under the good-faith security research exemption.[6] For example, if the researched bias or misalignment risks harm to confidentiality, integrity, or availability of information, or the physical safety of the users of the machines on which the AI system operates, or otherwise can be categorized as a security vulnerability finding, such research may qualify for the good-faith security research exemption.

However, it is appropriate to establish an exemption to protect research where the researched bias or misalignment may not directly affect security or safety (for example, manipulating a generative AI system to engage in racial or gender discrimination, or to produce synthetic child sexual abuse material). HPC further encourages the Register of Copyrights to consider adapting the good-faith security research exemption, in combination with definitions in Executive Order 14110, to establish an exemption for good-faith AI alignment research, such as:

> (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 alignment research.[7]
>
> (ii) 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.[8]

---

[4] "AI risk management calls for addressing many other types of risks[.]" NIST, Artificial Intelligence Risk Management Framework, AI Risks and Trustworthiness, pgs. 12, 39, Jan. 2023, https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf.

[5] 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.

[6] 37 CFR 201.40(b)(16).

[7] *Id.* at 201.40(b)(16)(i).

[8] See 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.

(iii) 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.[9]

(iv) For purposes of paragraph (i), "good-faith AI alignment research" means accessing a computer program solely for purposes of good-faith testing or investigation, of biased, discriminatory, 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.[10]

(v) Good-faith AI alignment 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.[11]

Item D. Technological Protection Measures and Methods of Circumvention.

Several generative AI alignment testing methods may be characterized as involving circumvention of technological protection measures to affect system behavior. For example, the copyright owner of the AI system may require a user account, the terms of which prohibit bypassing any protective measures or safety mitigations as a condition for permission to log in and use the system. By creating an account to access the system, an AI alignment researcher may be agreeing not to perform research. Such conduct was well documented in the *Sandvig v. Barr* proceedings in the context of research on algorithmic racial discrimination and the Computer Fraud and Abuse Act.[12]

Common AI alignment research techniques include bypassing guardrail programs or predefined rules that the AI system developers have established to align the system with human values, safeguard user interactions, prevent harmful or inaccurate system outputs, and protect against data extraction.[13] A range of attacks may be used to bypass guardrails. For example, "jailbreak prompts" are deliberately crafted inputs to bypass content safeguards and manipulate generative AI into creating harmful output, such as by directing the AI system to ignore previous instructions, or by escalating user privileges on the system.[14] Generative AI researchers may also circumvent automatic blocks on some inputs, as well as rate limits that restrict the volume or frequency of inputs to an AI system.

---

[9] *Id.* at Section 3(e).
[10] See 37 CFR 201.40(b)(16)(ii).
[11] *Id.* at 201.40(b)(16)(iii).
[12] *Sandvig v. Barr*, No. CV 16-1368 (JDB), 2020 WL 1494065 (D.D.C. Mar. 27, 2020). The United States Supreme Court cited this matter in *Van Buren v. United States*, 141 S. Ct. 1648 (2021).
[13] Daniel Kang et. al, Exploiting Programmatic Behavior of LLMs: Dual-Use Through Standard Security Attacks, Feb. 11, 2023, https://arxiv.org/abs/2302.05733. See also Agam Shah, Google Adds Guardrails to Keep AI in Check, Dark Reading, May 23, 2023, https://www.darkreading.com/cybersecurity-analytics/google-adds-guardrails-to-keep-ai-in-check.
[14] Xinue Shen et. al, "Do Anything Now": Characterizing and Evaluating In-The-Wild Jailbreak Prompts on Large Language Models, pgs. 1-3, 6 Aug. 7, 2023, https://arxiv.org/pdf/2308.03825.pdf.

Item E. Asserted Adverse Effects on Noninfringing Uses.

While the Librarian has established an exemption from 17 U.S.C. 1201(a)(1)(A) for independent security research performed "solely for purposes of good-faith testing, investigation, and/or correction of a security flaw or vulnerability," this exemption is limited to security and safety.[15] As a result, the exemption may not apply to circumventing software access controls for AI alignment research for some non-security or safety purposes that are still key for the underlying trustworthiness of the AI system. Enabling independent, good faith AI alignment research would help embrace diverse perspectives, promote impartial results, and promote a collaborative culture of ethical AI development, consistent with the broader goals of the United States as articulated in multiple policy initiatives. As with security research, limiting legal protections for AI alignment research to sources that have received authorization from the system owner would reduce the independence, volume, and diversity of testing.

Good-faith research into generative AI bias and misalignment is fair use and is not performed for the purpose of infringing, or enabling others to infringe, upon copyright. AI alignment research serves a socially beneficial purpose by evaluating and testing AI systems for algorithmic flaws that could cause harm, alerting stakeholders to these flaws so that they can be mitigated, and contributing to the advancement of computer science and the creation of better functioning AI systems. AI alignment research can also lead to the production of new creative works such as scientific publications, presentations, and educational material that discuss the research. The uses of AI alignment research are generally transformative – providing information about computer programs' susceptibility to bias and misalignment – rather than merely superseding the original copyrighted work.

The proposed class focuses on functional code, rather than expressive or imaginative work, by researching the algorithmic output of computer programs. In most instances of generative AI alignment research, it will not be necessary or desirable to reproduce more than small or *de minimis* portions of the copyrighted AI system in order to demonstrate the validity of the research. In addition, it is not uncommon for generative AI system vendors to assign ownership rights to the user for user input and related system output, which would reduce infringement concerns where the research reproduces the input and output.

Generative AI alignment research is highly unlikely to supplant the market for computer programs or generative AI systems. The research and the creative works produced by the research are of a wholly different nature than the AI systems subject to the research. Per the language proposed above, and consistent with the good-faith security research exemption, the research would be performed on lawfully obtained copies of the computer program. In addition, where generative AI alignment research leads to corrections of flaws, resulting in more trustworthy algorithms and AI systems, the value of the original work would be strengthened.

* * *

[15] 37 CFR 201.40(b)(16).

As the prevalence of AI systems continues to grow, so does the importance of testing for alignment with ethical principles. Extending protections for AI research and red teaming under DMCA Section 1201 not only fosters responsible development, but also promotes transparency, accountability, and trust. By addressing potential legal gaps and uncertainties, we can establish frameworks that improve and preserve AI alignment, ultimately safeguarding both technological advancements and societal interests.

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.