HACKING POLICY COUNCIL

AI red teaming - Legal clarity and protections needed

December 12, 2023

The Hacking Policy Council (HPC) releases this statement to highlight the need for legal protections and clarity for testing artificial intelligence (AI) systems for alignment with norms and ethical values.

While organizations may be familiar with red teaming to test software for security, "AI red teaming" has a broader scope by testing AI systems for flaws and vulnerabilities that include security, bias, discrimination, and other harmful or undesirable outputs. By identifying and disclosing misalignment in AI systems so they can be corrected, AI red teaming is a beneficial practice to help ensure the security, safety, and trustworthiness of AI. The HPC is supportive of AI red teaming and foresees that additional legal clarity and protections will be needed to encourage information sharing and enable independent red teaming.[1]

HPC recommends:

1) **Develop consistent alignment goals for AI red teaming.** Governments should work with the private sector to develop consistent goals for AI alignment in the context of AI red teaming. This will enable AI red teaming to test for dissonance with those goals.

2) **Protect information sharing for AI alignment purposes.** Governments should ensure legal frameworks for security information sharing are adapted to encourage and protect information sharing for harmful, discriminatory, or undesirable outputs in AI systems.

3) **Prepare to receive misalignment disclosures.** Organizations should prepare to accept disclosures from independent AI red teamers. This may require adaptations to security vulnerability disclosure programs and handling processes to accommodate disclosures for harmful, discriminatory, or undesirable outputs in AI systems.

4) **Clarify legal protections for independent AI red teaming.** Governments should ensure legal protections for independent security research extend to independent AI red teaming performed in good faith.

[1] Center for Cybersecurity Policy and Law, AI Hacking in the White House's Executive Order, Nov. 1, 2023, https://www.centerforcybersecuritypolicy.org/insights-and-research/ai-hacking-in-the-white-houses-executive-order.

"AI red teaming" - testing more than security

The concept of testing AI for alignment with norms or ethical principles is reflected in emerging industry practices and government guidance, though different terms may be used to describe this process. Increasingly, the term "AI red teaming" is used to refer to testing AI systems, but other terms (such as, for example, alignment testing) could be used.

In the context of technology testing, the term "red teaming" is often used in connection with security. In this context, definitions of "red teaming" commonly center on emulating adversarial attacks and exploitation capabilities against an organization's security posture.[2] However, in recognition that attacks on AI systems can cause damage in a variety of ways, AI risk management is not limited to testing for security.[3] Accordingly, the term "AI red teaming" often refers to testing for security as well as harmful, discriminatory, or undesirable outputs.

Many organizations test their AI systems for security, reliability, and fairness using in-house and external teams.[4] In addition, the private sector offers services that test AI systems for biased, discriminatory, harmful, or undesirable outputs. For example, several companies manage "bias bounty" services that facilitate research and disclosure of AI systems for such flaws, using a structured approach similar to security bug bounties.[5] The generative red team events held by the AI Village also focus on testing AI systems for non-security flaws.[6]

The Biden Administration's Executive Order (EO) 14110 reflects the government's intention that AI red teaming should test for flaws that are not limited to security. The EO identifies AI red teaming as a key safeguard for both the private sector and government use of AI systems. Among other things, the EO requires companies developing certain foundational AI models to provide reports of AI red teaming tests to the federal government on an ongoing basis.[7] The EO defines "AI red-teaming" as a "structured testing effort to find flaws and vulnerabilities in an AI system, often in a controlled environment and in collaboration with developers of AI…most often performed by dedicated "red teams" that adopt adversarial methods to identify flaws and vulnerabilities, such as harmful or discriminatory outputs from an AI system, unforeseen or undesirable system behaviors, limitations, or potential risks associated with the misuse of the system."[8]

---

[2] See, e.g., National Institute for Standards and Technology (NIST), Glossary: Red Team, https://csrc.nist.gov/glossary/term/red_team, (last accessed Nov. 15, 2023).
[3] "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.
[4] See, e.g., OpenAI, OpenAI Red Teaming Network, Sep. 19, 2023, https://openai.com/blog/red-teaming-network. See also, OpenAI, Our approach to alignment research, Aug. 24, 2022, https://openai.com/blog/our-approach-to-alignment-research.
[5] See, e.g.,
[6] AI Village, Generative Red Team Recap, Oct. 12, 2023, https://aivillage.org/defcon%2031/generative-recap.
[7] White House, Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence, Section 4.2(a)(i)(C), 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.
[8] White House, Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence, Section 3(d).

Clarity and legal protections needed

It is positive that the public and private sectors have begun utilizing AI red teaming to keep AI systems functioning as intended, and that the White House recognizes the value of AI red teaming to avoid negative outcomes. However, because this emerging practice is broader than security, existing mechanisms for exchanging test results and protecting AI red teaming participants may need to adapt.

As enhancing cybersecurity grew as a priority, governments established legal frameworks for sharing cybersecurity information, which include liability protections for sharing, to encourage common defense and free collaboration. Public and private sector entities also established legal protections for good faith security research, in the form of vulnerability disclosure programs and statutory defenses to anti-hacking laws, to encourage independent testing to strengthen security. While many challenges remain, these advances help support the security of the technology ecosystem and protect individuals.

As AI systems are more broadly adopted, avoiding misalignment of AI systems with ethical principles is likewise growing as a priority. It is important for private sector organizations and government agencies to clarify how existing protections and processes apply to AI red teaming, and to extend protections and processes where there is a gap. As with security, frameworks of protection for AI red teaming can avoid confusion, set expectations, facilitate information sharing, and encourage robust testing that will help maintain the trustworthiness of AI systems.

Below, the Hacking Policy Council suggests four broad actions to provide legal protection and clarity for AI red teaming.

1) Develop consistent alignment goals for AI red teaming

Encouraging AI red teaming as a valuable tool in bolstering trustworthiness, safety, and security demands a collaborative effort between government and industry. Governments should work with the private sector to develop consistent goals for AI alignment in the context of AI red teaming. This will better enable AI red teaming to test for dissonance with agreed-upon security, safety, and trustworthiness principles established in national, international, or industry guidelines, rather than testing against principles developed by a single organization.[9]

In this process, it will be helpful to delineate when harms or alignment principles fall within the domain of security, and when they extend beyond security. Some AI misalignment flaws may be characterized as security weaknesses because of their impact to users. For example, AI bias that causes some customers to inequitably pay more for a service may not qualify as a security weakness, but discriminatory AI output that leads to improper alteration of medical records and misdiagnoses may be associated with a loss of data integrity and fall within the scope of security weaknesses. However, this continues to be a gray area, leaving the scope of protections for security testing and information sharing unclear, as described below.

[9] Heidy Khlaaf, Toward Comprehensive Risk Assessments and Assurance of AI-Based Systems, Trail of Bits, pgs. 11-13, 24, Mar. 7, 2023, https://www.trailofbits.com/documents/Toward_comprehensive_risk_assessments.pdf.

Developing consistent alignment goals would be a useful outgrowth of upcoming NIST efforts. EO 14110 directs NIST to establish appropriate guidelines, procedures and processes to enable developers of AI "to conduct AI red-teaming tests to enable deployment of safe, secure and trustworthy systems."[10] The guidelines and processes established under NIST should leverage existing work, such as NIST's AI Risk Management Framework[11] and its guidance on identifying and managing AI bias,[12] as well as international best practices.

2) Protect information sharing for AI alignment purposes

As with security, facilitating information sharing for AI alignment can help organizations stay aware of the full range of risks and threats to AI systems and collaborate on defensive measures, making AI systems safer, more reliable, and more trustworthy.[13]

Governments should clarify and adapt cybersecurity information sharing frameworks to encourage and protect information sharing related to AI alignment. This should include legal protections for sharing information to protect against harms identified under EO 14110, including harmful or discriminatory outputs from an AI system, unforeseen or undesirable system behaviors, limitations, or potential risks associated with the misuse of the system.[14] Information sharing for AI alignment should also be informed by forthcoming guidance on AI red teaming to be produced by NIST as required under EO 14110, as well as guidance and best practices from other recognized standards bodies.[15]

For example, the Cybersecurity Information Sharing Act of 2015 (CISA 2015) was enacted to encourage robust sharing of cybersecurity information in order to aid in common defense and awareness. CISA 2015 enables private sector entities and state and local government agencies to share and receive cyber threat indicators and defensive measures for cybersecurity purposes "*notwithstanding any other provision of law*."[16] Though subject to certain sharing procedures and privacy safeguards, this framework exempts cybersecurity information sharing from coverage under civil and criminal laws, including regulatory and enforcement actions, antitrust, privacy restrictions, and open records laws. In addition, CISA 2015 ensures that cybersecurity information sharing does not act as a waiver of privilege, and that the shared information can be treated as proprietary information.[17]

---

[10] White House, Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence, Section 4.1(a)(ii).
[11] NIST, AI Risk Management Framework 1.0, Jan. 2023, https://airc.nist.gov/AI_RMF_Knowledge_Base/AI_RMF.
[12] NIST, Special Publication 1270, Towards a Standard for Identifying and Managing Bias in Artificial Intelligence, Mar. 2022, https://nvlpubs.nist.gov/NISTpubs/SpecialPublications/NIST.SP.1270.pdf.
[13] Group of Seven (G7), Hiroshima Process International Code of Conduct for Organizations Developing Advanced AI Systems, pgs. 3-5, Sep. 7, 2023, https://www.mofa.go.jp/files/100573473.pdf.
[14] White House, Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence, Section 3(d).
[15] *Id.*, Section 4.1(a)(ii).
[16] 6 USC 1503(c)(1), 1507(k)(1).
[17] 6 USC 1505(b)-(d). See also, Department of Homeland Security and Department of Justice, Guidance to Assist Non-Federal Entities to Share Cyber Threat Indicators and Defensive Measures with FEderal Entities under the Cybersecurity Information Sharing Act of 2015, pgs. 16-21, Oct. 2020, https://www.cisa.gov/sites/default/files/publications/Non-Federal%20Entity%20Sharing%20Guidance%20under%20the%20Cybersecurity%20Information%20Sharing%20Act%20of%202015_1.pdf.

Statutory definitions for "cyber threat indicators," "defensive measures," and "cybersecurity purposes" are limited to security information.[18] This encompasses, for example, disclosure of security vulnerabilities and exploits for the purpose of protecting an information system from a cybersecurity threat. However, it is unclear the extent to which these definitions can also encompass disclosure of, for example, information demonstrating algorithmic flaws that enable misuse of AI systems to self-replicate, produce erroneous or discriminatory decisions, or produce child pornography images. As a result, it is unclear whether the liability protections provided under CISA 2015 extend to the full range of AI red teaming conducted for alignment, but not cybersecurity, purposes.

3) Prepare to adapt vulnerability disclosure programs to include AI misalignment disclosures

Many private and public sector organizations have programs to receive and assess security vulnerability disclosures from internal and external sources.[19] In addition, several third party coordinators offer assistance in the disclosure and reporting of security vulnerabilities.[20] However, the scope of many such programs is limited to disclosure of security vulnerabilities.

As they integrate AI systems, organizations should consider working to leverage these vulnerability disclosure programs to receive and assess disclosures regarding AI systems' alignment with ethical principles. This could include providing a channel for external actors to disclose a discovery that, for example, an organization's AI deployment makes decisions with discriminatory results or can produce child pornography images. Such misalignment disclosures could then be routed to the organization's internal team best equipped to validate and mitigate the misalignment, which may be different than the organization's security team. The same industry norms on providing time to mitigate before public disclosure, and avoiding retaliation for good faith disclosures, should eventually apply to AI misalignment disclosures as they do for security vulnerability disclosures.

4) Clarify and extend legal protections for independent AI red teaming

Legal protections for independent security research should extend to independent AI red teaming performed in good faith. As guidance for this activity is developed, we encourage governments to consider how to enable red teaming in collaboration with solicited testers who have obtained formal authorization to test an AI system for alignment with ethical principles, as well as with independent or unsolicited researchers who have not obtained such authorization.[21] Enabling independent research would help embrace diverse perspectives, promote impartial results, and promote a collaborative culture of alignment. 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.

---

[18] 6 USC 1501(6), 1501(7), 1501(4).

[19] See e.g., Open Bug Bounty, Bug Bounty and Security Vulnerability Disclosure Programs List, https://www.openbugbounty.org/bugbounty-list (last accessed Nov. 21, 2023).

[20] See, e.g., Cybersecurity and Infrastructure Security Agency, Report to CISA, https://www.cisa.gov/report (last accessed Nov. 21, 2023).

[21] G7, Hiroshima Process International Code of Conduct for Organizations Developing Advanced AI Systems, pgs. 1-4, Sep. 7, 2023, https://www.mofa.go.jp/files/100573473.pdf.

Governments should clarify the extent to which legal protections for independent security research apply to AI red teaming, and extend legal protections where there is a gap. Areas where governments should consider clarifying or updating legal protections include, but are not limited to

a) **National and state computer crime charging policies.** The United States Department of Justice (DOJ) issued a charging policy for the Computer Fraud and Abuse Act (CFAA) to decline prosecution of good faith security research while also restricting activity inconsistent with good faith, such as extortion.[22] Similar legal frameworks have been adopted in non-US legal contexts, such as the Belgian vulnerability reporting policy to the Centre for Cybersecurity Belgium.[23] These charging policies are aimed at testing and disclosure in connection with security vulnerabilities, and are less likely to apply to good faith AI red teaming for non-security purposes. These charging policies could be extended to decline prosecution for good faith AI red teaming.

b) **Section 1201 of the Digital Millennium Copyright Act (DMCA)**. DMCA Section 1201 provides for renewable exemptions from restrictions against circumventing software access controls without authorization of the copyright owner.[24] These exemptions include independent security research performed "solely for purposes of good-faith testing, investigation, and/or correction of a security flaw or vulnerability[.]"[25] However, it is unlikely that the exemption universally applies to circumventing software access controls for non-security AI red teaming purposes, such as to identify harmful or discriminatory outputs. HPC encourages the Copyright Office to establish a separate DMCA Section 1201 exemption for good faith AI red teaming that is not covered by the existing good faith security research exemption. An AI red teaming exemption could build upon the language of the good faith security research exemption, as well as EO 14110's definition of AI red teaming, to protect this activity.

* * *

As the prevalence of AI systems continues to grow, so does the importance of testing for alignment with ethical principles. Extending legal protections for AI red teaming 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.

For more information, contact Harley Geiger, Hacking Policy Council Coordinator, at HLGeiger@Venable.com.

[22] Department of Justice, 9-48.000 - Computer Fraud and Abuse Act, https://www.justice.gov/jm/jm-9-48000-computer-fraud#:~:text=The%20Department%20will%20not%20charge,authority%20to%20grant%20such%20authorization (last accessed Nov. 17, 2023).
[23] Centre for Cybersecurity Belgium, New Legal Framework for Reporting IT Vulnerabilities, Feb. 15, 2023, https://ccb.belgium.be/en/news/new-legal-framework-reporting-it-vulnerabilities.
[24] 17 USC 1201(a)(1)(C).
[25] 37 CFR 201.40(b)(16).