hackerone

January 15, 2025

Working Group 3: Technical Risk Mitigation
European Commission
Rue de la Loi 200
1049 Brussels
Belgium

VIA ELECTRONIC SUBMISSION

Re: Second Draft of the General-Purpose AI Code of Practice

Dear Mr. Yoshua Bengio, Mr. Daniel Privitera, and Mr. Nitarshan Rajkumar:

HackerOne Inc. (HackerOne) submits the following comments in response to the Second Draft of the General-Purpose AI Code of Practice.[1] We appreciate the opportunity to contribute. We commend the work of the European Commission and the independent experts for developing this important framework on the responsible deployment of general-purpose AI.

HackerOne is the global leader in vulnerability elimination through continuous security testing. Its industry-leading HackerOne Platform combines AI with the expertise of the world's largest community of security researchers to deliver ongoing vulnerability discovery and management across the software development lifecycle. The platform offers bug bounty, vulnerability disclosure, pentesting, code audits, challenges, and AI red teaming.

We believe that securing AI systems is essential for establishing trust in their use and ensuring their safe deployment. In our comments below, we focus on key areas where we believe the draft Code of Practice can benefit from further clarification and enhancement, particularly in relation to risk management strategies, vulnerability management, and security assurance.

Alignment with Technical Standards

One area of particular importance is ensuring that the draft Code of Practice aligns with widely recognized technical standards, such as ISO/IEC 27001[2], NIST 800-53,[3] in addition to the RAND study. While we observe that some alignment with these standards is already present in the Measures, we strongly encourage the drafters to ensure that the final framework fully integrates these established technical standards. This would not only enhance the overall

[1] European Commission, *Second Draft of the General-Purpose AI Code of Practice*, January 15, 2024 https://digital-strategy.ec.europa.eu/en/library/second-draft-general-purpose-ai-code-practice-published-written-independent-experts.
[2] ISO/IEC 27001, Information technology – Cybersecurity and Privacy Protection – Information Security Management Systems – Requirements, International Standards Organization, Oct. 2022, https://www.iso.org/standard/27001.
[3] National Institute of Standards and Technology (NIST), Special Publication 800-53, Security and Privacy Controls for Information Systems and Organizations, Sept. 2020, https://csrc.nist.gov/pubs/sp/800/53/r5/upd1/final.

effectiveness of the Code but also help organizations meet global security standards, which is essential for widespread, trusted and secure adoption of AI technologies.

Commitment 12 (Security Mitigations), Measure 12.2 (Security Assurance):

HackerOne strongly agrees with several measures provided in Measure 12.2 - Security Assurance, and we encourage the independent experts to retain these measures in the Code of Practice. We believe that each of the following measures plays a crucial role in testing and protecting AI systems. From the draft Code, these include:

> **a) Frequent active red-teaming:** Red-teaming involves ethical hackers simulating real-world threats, typically to accomplish specific objectives such as exfiltrating data or disrupting operations. This exercise helps identify vulnerabilities that may not be discovered through conventional testing methods and ensures AI systems are resilient to adversarial attacks.
>
> **b) Secure communication channels for third parties to report security issues:** Ensuring secure communication channels, such as Vulnerability Disclosure Policies (VDPs), would provide a structured framework for receiving and responding to vulnerability reports. This would ensure that vulnerabilities are promptly identified, evaluated, and remediated in a secure and transparent manner.
>
> **c) Competitive bug bounty programs to encourage public participation in security testing:** Bug bounty programs (BBPs) incentive ethical hackers with monetary rewards to find vulnerabilities in an organization's system, providing an additional layer of security. BBPs are especially effective at uncovering vulnerabilities that automated scanners may miss.
>
> **d) Clear and public security whistleblower policies which prohibit retribution:** Whistleblower policies ensure that individuals who report security vulnerabilities are protected from retaliation, encouraging them to come forward with vital information. We specifically suggest including a provision that avoids legal retaliation against good faith security researchers, ensuring they are shielded from legal consequences when reporting vulnerabilities in a responsible manner.

Additionally, we recommend specifying that penetration testing is a key security assurance measure. While similar to red-teaming, penetration testing is distinct and focuses more specifically on breaching a system's security for the purpose of vulnerability identification.

Commitment 10 (Evidence Collection, Measure 10.2 (State-of-the-art model evaluations):

We fully support the commitment to evaluate AI models in order to address systemic risks using a range of suitable methodologies. Such risks include both security and non-security (i.e., safety and trustworthiness considerations), and we believe this should be explicitly clarified in Measure 10.2.

Measure 10.2 notes several risk evaluation methodologies, such as red-teaming and other adversarial testing, that are also outlined in Measure 12.2. We encourage the drafters to also reference non-security methodologies, such as bias bounties and independent non-security red-teaming in Measure 10.2:

- **Bias Bounties:** AI systems must undergo thorough testing and evaluation to address a spectrum of potential harms, extending beyond the traditional focus on security vulnerabilities. This includes evaluating models for non-security issues, such as bias, discrimination, fairness, trustworthiness, accuracy, and other adverse outcomes that may affect stakeholders. Bias bounty programs are a highly effective way to incentivize researchers to identify and address such issues within AI models. Just as BBPs are used to detect security vulnerabilities, bias bounties would provide rewards for discovering instances of biased outcomes in AI systems.

- **Independent Red-Teaming:** The methodologies of state-of-the-art model evaluations to address systemic risks should be extended beyond security. In addition to identifying security vulnerabilities, these evaluations must also encompass non-traditional risks, as described above. By broadening the scope of red-teaming to include both security and non-security risks, we can ensure that AI systems are thoroughly tested for a comprehensive range of potential harms, leading to more responsible and ethical AI deployment.

We believe this will help to clarify that model evaluations cover both security and non-security risks, and that the evaluation methodologies should encompass both types of risks. Explicitly clarifying that the scope of evaluation includes security and non-security testing will significantly strengthen the overall risk management framework.

Conclusion

HackerOne appreciates the opportunity to provide a response to the second draft of the AI Code of Practice. As the conversation around this topic continues to evolve, we would welcome the opportunity to further serve as a resource and provide insights on how to raise the standard for security in AI.

* * *

Respectfully Submitted,

Ilona Cohen
Chief Legal and Policy Officer
HackerOne