Poster: Towards Understanding Bug Bounties for AI/ML Open-Source Vulnerabilities in GitHub Repositories

Jessy Ayala and Joshua Garcia
Donald Bren School of Information and Computer Sciences
University of California, Irvine, USA

Abstract—**Recently, specialized bug bounty platforms, such as huntr, have emerged to incentivize the discovery of vulnerabilities in open-source software (OSS) that power AI (Artificial Intelligence) and ML (Machine Learning) systems. In this extended abstract, we present preliminary results from an empirical study on AI/ML OSS bug bounty reports, examining their characteristics, severities, and resolution patterns, e.g., in comparison to their non-AI/ML counterparts. Using scraping techniques, we were able to gather 6,427 OSS-based bug bounty reports, the largest dataset to date in recent work. Our study includes the following key findings: (1) AI/ML OSS vulnerabilities differ from what is expected based on expert-curated lists, e.g., OWASP Top Ten ML Security Risks; (2) 44.4% of CVE-IDs from CVE-assigned AI/ML vulnerabilities in our dataset are missing from the NVD, preventing alerts from being sent to all affected projects and advisory databases; and (3) almost half (49.5%) of AI/ML vulnerabilities with an assigned severity remain unfixed post-disclosure, while most non-AI/ML vulnerabilities are fixed (99.7%), indicating challenges in patch development. These findings provide early insight into the evolving security posture of AI/ML OSS and inform future research directions and tooling needs for securing AI/ML-based systems.**

I. Background and Research Questions

To study software vulnerability management aspects and practices of AI/ML-focused OSS projects, we center our study on one of the largest OSS ecosystems, i.e., GitHub. OSS platform security features [1], e.g., vulnerability reporting policies, and automated report notifications, i.e., through the huntr bug bounty program [2], help facilitate communication between the different actors during the triaging process [3].

After being acquired by Protect AI in mid-2023, huntr is an OSS bug bounty program specifically for AI/ML-focused GitHub repositories [3]. huntr pays vulnerability reporters for finding vulnerabilities in GitHub repositories and also pays project maintainers for fixing them, i.e., should the hunter not provide a fix upon validation of the vulnerablility from a maintainer or administrator. In particular, huntr offers higher bounties for AI/ML bug bounty reports than prior for general OSS-based reports. By doing so, huntr encourages vulnerability reporters to report vulnerabilities and project maintainers to provide the fixes promptly.

Prior work has investigated specific types of AI vulnerabilities that can be traced to OSS projects. For instance, Kathikar et al. mined models from Hugging Face, linked them to their underlying code bases on GitHub, and performed a large-scale vulnerability assessment of these repositories [4]. Other related work has explored qualitative perspectives of the general OSS vulnerability management process [5], [6], challenges related to both bug bounty report review [7], and OSS platform security features [6], but not in the AI/ML OSS space. Previous work has also focused on understanding bug bounty platforms, e.g., Luna et al. [8] investigated vulnerability reporter productivity across different platforms, and how to improve such platforms [9], [10], [11], [12], but not the aspects of AI/ML OSS vulnerabilities from bug bounty platforms.

To understand the aspects of GitHub AI/ML vulnerabilities from disclosed bug bounty reports, challenges faced during the vulnerability disclosure process for such reports, and their comparisons to general GitHub vulnerabilities, i.e., from other bug bounties, we investigate the following research questions:

RQ1: What are the most frequently reported vulnerabilities in open-source AI/ML projects?

RQ2: How do key stakeholders handle reported vulnerabilities in open-source AI/ML projects?

RQ3: How do reported vulnerabilities in open-source AI/ML projects compare to those reported in non-AI/ML projects?

II. Methodology

We organize our study around disclosed AI/ML-focused OSS bug bounty reports sourced from the huntr bug bounty program. However, huntr does not provide a method of knowing how many bug bounty reports are publicly disclosed, nor a list of projects with existing bug bounty reports. The huntr hacktivity page [13] provides a list of the 100 most recent publicly disclosed bug bounty reports, and disclosed bug bounty reports can be viewed per project using its respective author and GitHub repository name. Our collected bug bounty reports are scraped using the 100 most recent bug reports from each unique AI/ML-centered project on the huntr hacktivity page from 09/30/2023 to 02/25/2025, i.e., those disclosed after huntr was acquired by Protect AI, and from projects listed on the *bounties* page [2]. Further, we scraped non-AI/ML bug bounty reports from projects with disclosed reports dated from 09/01/2021 to 09/30/2023. Bug bounty reports date back to

August 2019, spanning 5.5 years. We also gather bug bounty reports that were marked as Informative, Not Applicable, Pending, Spam, or Self-closed, to gather a holistic view of reports that are not necessarily deemed as impactful to the security posture of projects by maintainers.

III. Preliminary Analyses and Findings

TABLE I
SEVERITIES FOR AI/ML OSS VULNERABILITIES AND 2022 NVD

Severity	NVD in 2022	Our data (AI/ML)	OSS AI models [4]
Low	14.7%	2.1%	6.8%
Medium	60.7%	25.7%	Not reported
High	24.6% (or Critical)	48.6%	36.0%
Critical	——	23.6%	Not reported

TABLE II
TOP 5 CWES FOR AI/ML OSS VULNERABILITIES

CWE-ID	CWE description	% of vulnerabilities
CWE-22	Path traversal '..filename'	11.3%
CWE-284	Improper access control	7.6%
CWE-400	Denial of service	5.9%
CWE-78	OS command injection	5.5%
CWE-79	Stored cross-site scripting	4.2%

TABLE III
REVIEW TURNAROUND TIMES FOR OSS AI/ML BUG BOUNTY REPORTS

Turnaround time	% of vulnerabilities
within a day	1.3%
within a week	0.9%
within 2 weeks	0.0%
within a month	1.1%
within 3 months	81.5%
within 6 months	12.6%
within 1 year	2.6%
after 1 year	0.2%

RQ1-1: Table I shows the severity distribution for (1) NVD entries in 2022, (2) AI/ML OSS vulnerabilities in our curated dataset, and (3) AI OSS vulnerabilities discovered in Hugging Face models, including their usages and forks in other OSS projects, discovered by Kathikar et al. [4]. Both our work and Kathikar et al. show a skew of High severity vulnerabilities in AI/ML OSS vulnerabilities when compared to general vulnerabilities, i.e., as demonstrated by the NVD'22 distribution. This demonstrates significant implications for the AI software supply chain and AI risk management more broadly.

RQ1-2: Table II reflects the diversity of vulnerability types, categorized by CWE-IDs, in disclosed AI/ML OSS bug bounty reports. Some of such CWE-IDs, e.g., CWE-400: Denial of service, are not consistent with expert-curated lists for common AI/ML vulnerabilities, e.g., OWASP Top Ten ML Security Risks. Based on this, we plan to develop a taxonomy to determine root causes and symptoms of vulnerabilities in our dataset, both with and without assigned severities, to further understand the nature of AI/ML OSS vulnerabilities.

RQ2: Table III shows bins of review turnaround times for AI/ML OSS vulnerabilities in our dataset. On average, the review turnaround time for such bug bounty reports is 85.7 days, which is in line what is generally expected for bug bounty report review, e.g., for professional organizations or companies (90 days) [14]. However, we find that 50.5% of such vulnerabilities fixed while the remaining 49.5% are not fixed, demonstrating the complexity of developing patches for AI/ML OSS vulnerabilities. Further, 44.4% of CVE-assigned AI/ML OSS vulnerabilties are missing from the NVD, which can be partially attributed due to NIST halting CVE ingestion in 2024 and the volume of vulnerabilities outpacing processing capacity [15]; raising concerns about lack of vulnerability awareness for affected projects due to NVD inactivity.

RQ3: Upon initial analyses, we compare turnaround times shown in Table III to those for general OSS vulnerabilities in Ayala et al. [16]. We find that turnaround times for OSS AI/ML vulnerabilties are greater turnaround times for OSS non-AI/ML vulnerabilities, i.e., using Mann-Whitney and effect size tests, implying that AI/ML OSS vulnerabilties are resolved slower, i.e., take longer, than for non-AI/ML OSS vulnerabilities. Comparing fix rates, we find that 99.7% of non-AI/ML OSS vulnerabilities from the same dataset are fixed, while 50.5% of AI/ML OSS vulnerabilities are fixed; further reflecting vulnerability complexity in AI/ML counterparts. We also plan to compare bounty amounts between such reports to determine statistical significance in disclosure and fix payouts.

References

[1] GitHub, "Code security documentation," https://docs.github.com/en/code-security, 2017.

[2] A. Nygate, "Huntr," https://huntr.com, 2020.

[3] Huntr, "Participation guidelines," https://huntr.com/guidelines, 2020.

[4] A. Kathikar, A. Nair, B. Lazarine, A. Sachdeva, and S. Samtani, "Assessing the vulnerabilities of the open-source artificial intelligence (ai) landscape: A large-scale analysis of the hugging face platform," in *2023 IEEE International Conference on Intelligence and Security Informatics (ISI)*, 2023, pp. 1–6.

[5] D. Wermke, N. Wöhler, J. H. Klemmer, M. Fourné, Y. Acar, and S. Fahl, "Committed to trust: A qualitative study on security & trust in open source software projects," in *2022 IEEE Symposium on Security and Privacy (SP)*, 2022, pp. 1880–1896.

[6] J. Ayala, Y.-J. Tung, and J. Garcia, "A mixed-methods study of open-source software maintainers on vulnerability management and platform security features," 2024.

[7] J. Ayala, S. Ngo, and J. Garcia, "A deep dive into how open-source project maintainers review and resolve bug bounty reports," in *2025 IEEE Symposium on Security and Privacy (SP)*, 2024.

[8] D. Luna, L. Allodi, and M. Cremonini, "Productivity and patterns of activity in bug bounty programs: Analysis of hackerone and google vulnerability research," in *Proceedings of the 14th International Conference on Availability, Reliability and Security*, 2019, pp. 1–10.

[9] S. Atefi, A. Sivagnanam, A. Ayman, J. Grossklags, and A. Laszka, "The benefits of vulnerability discovery and bug bounty programs: Case studies of chromium and firefox," in *Proceedings of the ACM Web Conference 2023*, 2023, pp. 2209–2219.

[10] A. Y. Ding, G. De Jesus, and M. Janssen, "Ethical hacking for boosting iot vulnerability management: A first look into bug bounty programs and responsible disclosure," in *Proceedings of the 8th International Conference on Telecommunications and Remote Sensing*, 2019.

[11] M. Zhao, A. Laszka, and J. Grossklags, "Devising effective policies for bug-bounty platforms and security vulnerability discovery," *Journal of Information Policy*, vol. 7, pp. 372–418, 2017.

[12] T. Walshe and A. C. Simpson, "Coordinated vulnerability disclosure programme effectiveness: Issues and recommendations," *Computers & Security*, vol. 123, p. 102936, 2022.

[13] Huntr, "Hacktivity," https://huntr.com/bounties/hacktivity/, 2021.

[14] HackerOne, "Program metrics," https://docs.hackerone.com/en/articles/8490865-program-metrics, 2024.

[15] J. Ma, "Delayed vulnerability analysis puts america at a cybersecurity disadvantage," https://www.fdd.org/analysis/policy_briefs/2025/03/21/delayed-vulnerability-analysis-puts-america-at-a-cybersecurity-disadvantage/, 2025.

[16] J. Ayala, Y.-J. Tung, and J. Garcia, "Investigating vulnerability disclosures in open-source software using bug bounty reports and security advisories," *arXiv preprint arXiv:2501.17748*, 2025.

Towards Understanding Bug Bounties for AI/ML OSS Vulnerabilities in GitHub Repositories

Jessy Ayala and Joshua Garcia – *University of California, Irvine*
jessya1@uci.edu joshug4@uci.edu

In open-source software (OSS), the number of vulnerabilities has gone ↑↑↑

- *Bug bounty reports* can be submitted for AI/ML open-source projects using bug bounty platforms, e.g., **`huntr`**, for project maintainers to review vulnerabilities

There is a **lack of** literature **understanding AI/ML vulnerabilities in OSS projects**.

We want to investigate:

What aspects of AI/ML OSS reports make them unique from other vulnerabilities?

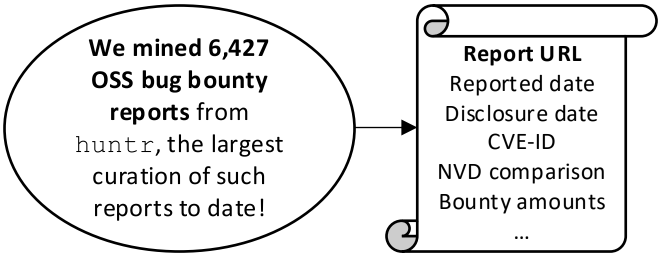
How are they handled by OSS maintainers? E.g., review times and fix rates.

How do they compare to non-AI/ML reports? E.g., bounty payouts and types.

RQ1 What are the most frequently reported vulnerabilities in open-source AI/ML projects?

RQ2 How do key stakeholders handle reported vulnerabilities in open-source AI/ML projects?

RQ3 How do reported vulnerabilities in open-source AI/ML projects compare to those reported in non-AI/ML projects?



We mined 6,427 OSS bug bounty reports from `huntr`, the largest curation of such reports to date!

Report URL
Reported date
Disclosure date
CVE-ID
NVD comparison
Bounty amounts
…

Highlighted Preliminary Findings from Mining AI/ML OSS Bug Bounty Reports

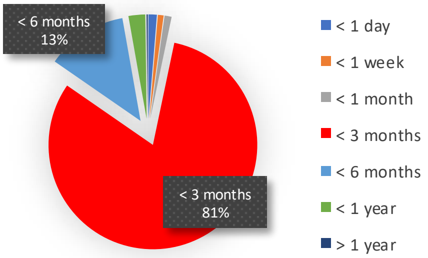
SEVERITIES FOR AI/ML OSS VULNERABILITIES AND 2022 NVD

Severity	NVD in 2022	Our data (AI/ML)	OSS AI models
Low	14.7%	2.1%	6.8%
Medium	60.7%	25.7%	Not reported
High	24.6% (or Critical)	48.6%	36.0%
Critical	——	23.6%	Not reported

TOP 5 CWES FOR AI/ML OSS VULNERABILITIES

CWE-ID	CWE description	% of vulnerabilities
CWE-22	Path traversal '..filename'	11.3%
CWE-284	Improper access control	7.6%
CWE-400	Denial of service	5.9%
CWE-78	OS command injection	5.5%
CWE-79	Stored cross-site scripting	4.2%

REVIEW TURNAROUND TIMES FOR OSS AI/ML BUG BOUNTY REPORTS



CWEs in AI/ML OSS vulnerabilities do not line up with expected, expert-curated lists, e.g., OWASP Top Ten ML security risks. **There is need for a taxonomy.**

AI/ML OSS vulnerabilities are skewed towards High severity, especially when compared to overall vulnerabilities, e.g., NVD'22 severities, consistent with prior work in studying OSS AI model vulnerabilities.

Reviewing AI/ML OSS vulnerabilities is time consuming! They take longer to review vs non-AI/ML counterparts based on statistical test results.

Despite being consistent with recommended review turnaround times (86 days vs 90 days), **49.5% of the vulnerabilities in our dataset remain unpatched**. Only 0.03% of non-AI/ML counterparts are unpatched!

Upon initial inspection, **AI/ML OSS vulnerabilities pay more!** They also have a greater proportion of Informative or N/A reports vs non-AI/ML counterparts, possibly indicating **more critique by reviewers**.

What's next?

Further analyzing review rates and bounties with statistical techniques

Closer looking at how popular projects are handling reports, e.g., who is involved?

Further analyzing reports to develop a taxonomy for AI/ML OSS vulnerabilities