Coordinated Flaw Disclosure for AI: Beyond Security Vulnerabilities

Sven Cattell[*1]**, Avijit Ghosh**[*2]**, Lucie-Aimée Kaffee**[2]

[1]AI Village
[2]Hugging Face
sven@nbhd.ai, avijit@huggingface.co, lucie.kaffee@huggingface.co

Abstract

Harm reporting in Artificial Intelligence (AI) currently lacks a structured process for disclosing and addressing algorithmic flaws, relying largely on an ad-hoc approach. This contrasts sharply with the well-established Coordinated Vulnerability Disclosure (CVD) ecosystem in software security. While global efforts to establish frameworks for AI transparency and collaboration are underway, the unique challenges presented by machine learning (ML) models demand a specialized approach. To address this gap, we propose implementing a Coordinated Flaw Disclosure (CFD) framework tailored to the complexities of ML and AI issues. This paper reviews the evolution of ML disclosure practices, from ad hoc reporting to emerging participatory auditing methods, and compares them with cybersecurity norms. Our framework introduces innovations such as extended model cards, dynamic scope expansion, an independent adjudication panel, and an automated verification process. We also outline a forthcoming real-world pilot of CFD. We argue that CFD could significantly enhance public trust in AI systems. By balancing organizational and community interests, CFD aims to improve AI accountability in a rapidly evolving technological landscape.

1 Introduction

The landscape of AI accountability is rapidly evolving, as exemplified by The New York Times' lawsuit against OpenAI and Microsoft on December 27, 2023. This legal action alleges that ChatGPT was trained on the Times' published works without authorization, providing evidence of verbatim article generation (Mac 2023). As part of a growing trend of litigation against AI providers (O'Brien 2023; Authors Guild 2023; Fenwick & West LLP 2023; ACLU 2022), this case highlights the complexities of replicating and verifying such claims. The Times' lawsuit has garnered significant attention, partly due to its financial resources and media reach - advantages often unavailable to smaller affected parties. This underscores how such factors can shape public discourse and potentially drive improvements in AI systems through external pressure (Raji and Buolamwini 2022).

The emergence of unexpected model behaviors beyond intended scope and purpose emphasizes the need for robust frameworks to identify and address these issues effectively. This stands in contrast to the well-established Common Vulnerabilities and Exposures (CVE) ecosystem in cybersecurity, where software vulnerabilities are systematically cataloged and addressed with greater clarity and consensus.

Historically, AI accountability initiatives have relied heavily on periodic audits, emphasizing repetitive assessments but lacking a structured framework for reporting user-identified issues post-deployment. Recent policy efforts have begun to acknowledge this gap (European Commission 2020). This audit-centric paradigm is reflected in influential policies such as the U.S. Executive Order on AI (The White House 2023), the EU AI Act (European Parliament and Council 2024), and New York City's Local Law 144 (New York City Department of Consumer and Worker Protection 2021). However, this approach falls short when compared to the more comprehensive Coordinated Vulnerability Disclosure (CVD) processes standard in software security.

CVD plays a crucial role in enabling independent researchers to report newly identified vulnerabilities to affected vendors and the public (MultiMedia LLC 1999). This process facilitates transparent remediation before potential exploitation by malicious actors and has become a vital practice enshrined in government regulations and industry standards. Notably, the FDA mandates the implementation of CVD programs for medical device companies to enhance cybersecurity (U.S. Food and Drug Administration 2023).

While CVD has proven effective in traditional software security, its direct application to machine learning (ML) systems faces unique challenges. ML diverges from traditional software in two key aspects: 1) ML model issues must satisfy statistical validity thresholds, and 2) concerns related to trustworthiness and bias often extend beyond the typical scope of security vulnerabilities (Office of Science and Technology Policy 2022). Many critical bias reports discussed in social media and news may not align with the accepted definition of a vulnerability[1].

Recognizing this limitation in the conventional definition

---

[1]The Coordinated Vulnerabilities Enumeration program, managed by MITRE and central to the CVD ecosystem, defines a vulnerability as: *A problem in software, firmware, hardware, or service component resulting from a weakness that can be exploited, causing a negative impact on the confidentiality, integrity, or availability of the affected component or components.*

of a *vulnerability*, we propose expanding the terminology with a new term - *Flaw*. We define a flaw as *any unexpected model behavior that is outside of the defined intent and scope of the model design*.

To address the unique challenges posed by ML systems, we advocate for adapting the CVD process into a dedicated "Coordinated Flaw Disclosure" (CFD) framework. Tailored to ML's distinctive properties, CFD aims to formalize the recognition of valid issues in ML models through an adjudication process. This process seeks to balance the interests of vendors and the broader community, enhancing public trust by curating disclosed flaws into actionable reports. The CFD framework introduces several key innovations:

- Extended model cards that include detailed documentation of intent and scope, providing a baseline for flaw assessment.

- A dynamic scope expansion mechanism that allows for the inclusion of common, unforeseen uses of the model.

- An independent adjudication panel to mediate disputes and handle edge cases.

- An automated verification process to streamline issue reporting and reproduction.

In the subsequent sections of this paper, we explore specific aspects of this proposal. Section 2 provides essential background information, comparing periodic audits and ad hoc harm reporting while exploring emerging participatory auditing practices. Section 3 examines evolving trends in disclosure practices within ML systems and identifies their shortcomings. Section 4 outlines the proposed changes within the CFD framework to improve disclosure rigor. Section 5 explores the practical implementation of CFD, proposing the utilization of model cards to define scope and intent. We conclude by discussing limitations of our framework and future work which includes a real-world implementation pilot and the development of a Common Use Enumeration (CUE) system to standardize the tracking of AI model applications.

2 Background

In the following, we describe the background of our work. We will use the terminology of the glossary in Table 1.

2.1 Coordinated Disclosure Conventions in Cybersecurity

The cybersecurity community has cultivated a robust, 30-year history of well-established vulnerability disclosure practices, prominently featuring Coordinated Vulnerability Disclosure (CVD). This practice, acknowledged as a cornerstone of cybersecurity, facilitates collaboration between software vendors and vulnerability reporters to identify and address security flaws. Given the impossibility of eliminating all vulnerabilities during production, CVD establishes a crucial post-publication process for addressing these issues. Organizations, recognizing the limitations of in-house efforts, benefit from external parties and security researchers who disclose vulnerabilities through the CVD process. This collaboration allows affected organizations and the public to proactively address vulnerabilities, preempting exploitation by malicious actors. Moreover, the CVD process acts as a shield for those reporting vulnerabilities, demonstrating their good faith and avoiding legal conflicts (Cybersecurity Coalition 2019).

CVD's significance extends to government policies, guidance, and regulations. United States civilian government agencies (Cybersecurity and Infrastructure Security Agency 2020) and numerous federal IT contractors (116th Congress 2020) are mandated to adopt vulnerability disclosure policies. The Food and Drug Administration incorporates CVD into its pre- and post-market guidance for medical device security (U.S. Food and Drug Administration 2023). Key standards and guidelines from the National Institute of Standards and Technology (NIST), such as the NIST Cybersecurity Framework (National Institute of Standards and Technology 2018), endorse CVD as a core practice. The EU's NIS 2 Directive (European Parliament and Council 2022) requires critical infrastructure entities to establish CVD processes, and the recently passed EU Cyber Resilience Act (EUC 2023) mandates software manufacturers to adopt CVD processes. The European Union Agency for Cybersecurity further promotes national CVD legislation for its member states (European Union Agency for Cybersecurity (ENISA) 2022). Additionally, CVD features prominently in international standards and industry best practices on security.

CVD operates within a broader framework of programs that identify, index, and communicate security vulnerabilities, including the National Vulnerability Database, CVE system, and Common Vulnerability Scoring System (CVSS). These programs support diverse security operations such as vulnerability management, risk assessments, penetration testing, patch management, and threat intelligence (Cybersecurity Coalition 2019). CVE Numbering Authorities (CNA) are organizations external to MITRE that are authorized to issue CVEs within their scope, but MITRE acts as a "root" that handles disputes. For example vulnerabilities in open source projects managed by the Apache Software Foundation will be sent to them first and then, if needed, appealed to MITRE (Apache 2024). In the EU, the European Cybersecurity Scheme on Common Criteria (EUCC) is currently developed to cover information and communications technology (ICT) based on the international Common Criteria scheme (European Union Agency for Cybersecurity (ENISA) 2024).

In the realm of machine learning systems, MITRE recognizes two ML CVEs: CVE-2019-20634 (CVE 2020) and CVE-2023-29374 (CVE 2023b). While CVE-2023-29374 involves a remote code execution with an LLM exploit, the vulnerability detailed by CVE-2019-20634 in Proofpoint's system is less straightforward from a security perspective, since creating an exploitable copycat ML model requires a ton of positive and negative examples and their classification scores. Skylight Cyber's discovery of a vulnerability in Cylance's malware detection system, although awarded a Vulnerability Note by CERT, did not receive a CVE (VU4 2019). Evaluating the impact of these ML vulnerabilities is complex, and the existing coordinated disclosure infrastructure lacks precedence for such cases.

Concept	Description
Vulnerability	A problem in software, firmware, hardware, or service component resulting from a weakness that can be exploited, causing a negative impact on the confidentiality, integrity, or availability of the affected component or components.
Flaw	Any unexpected model behavior that is outside of the defined intent and scope of the model design.
Intent	The stated purpose, goals, and intended use cases of an AI model or system, as defined by its creators or deployers. This includes the model's primary functions, target audience, and expected performance metrics.
Scope	The boundaries of an AI model's intended operation, including what the model is designed to do, what it is not designed to do, and known limitations or exclusions. It defines the range of inputs, outputs, and use cases that are considered valid for the model.
Coordinated Vulnerability Disclosure (CVD)	Process in which vulnerabilities will only be disclosed to the public after reporting to the software vendor and giving them sufficient time to address the vulnerability.
Common Vulnerabilities and Exposures (CVE)	A system managed by the MITRE Corporation under the National Cybersecurity FFRDC, offering a standard method for cataloging publicly known information security vulnerabilities and exposures.
Coordinated Flaws Disclosure (CFD)	The framework proposed in this paper to address risks derived from flaws in ML models, analogous to CVD.
Common Flaws and Exposures (CFE)	Analogous to CVE, the proposed system to report flaws of ML models.
Model Cards	First proposed by Mitchell et al. (2019), model cards standardize the documentation of ML models, including the intended use cases of a reported model.

Table 1: Glossary.

Moreover, socioeconomic flaws in ML models (Solaiman et al. 2023), not categorized as vulnerabilities (see Table 1), also demand reporting.

Recognizing the importance of CVD to security, there is a temptation to extend these structures to encompass the disclosure of ML issues. However, existing CVD processes, without substantial modifications, prove insufficient to address the diverse range of ML issues. The disparities between ML models and traditional software necessitate a tailored approach. Nonetheless, the fundamental value of CVD in cybersecurity underscores the urgency of establishing a CVD-like process for ML models.

2.2 Differences between Traditional Software and ML in Disclosure

Traditional software vulnerabilities, such as buffer overflows, SQL injections, cross-site scripting, etc., typically arise from errors or weaknesses in the code, allowing attackers to compromise key properties like confidentiality, integrity, and availability.

In contrast to traditional software vulnerabilities with clear remediation steps, addressing ML model issues may involve retraining the model, modifying its architecture, or redefining its purpose. Furthermore, due to the probabilistic and complex nature of ML models, judgments are often difficult as the models lack distinct exploited and non-exploited states (Barbierato and Gatti 2024).

Most consumer ML models are deployed as closed-source APIs (Allison 2023), posing challenges in tracking and indexing vulnerabilities, especially in Software as a Service

(SaaS) products. Versioning is not apparent to end users, and after remediation, the problem is not reproducible. Although these issues are recognized (Nichols 2021), CVEs for SaaS deployments are rare but have been issued (CVE 2023a). Continuous retraining and development add complexity to many AI systems.

These fundamental differences render existing vulnerability disclosure frameworks less applicable to ML model issues. Checking if an issue violates model integrity or intended functionality is less straightforward than identifying and replicating a software vulnerability due to ML models' lack of causal explanations about achieved predictions (Barbierato and Gatti 2024). In the following subsections, we delve deeper into the unique challenges posed by ML systems that existing vulnerability management frameworks cannot fully address.

ML Models are Expected to Make Mistakes   At its essence, the distinction in reporting known issues between traditional and ML-based systems lies in the expectation that ML systems will make mistakes. While these isolated incidents may be inconvenient for users, they are anticipated, and both the vendor and users should have processes in place to address this inherent limitation of machine learning. It is crucial to consider mistakes collectively, incorporating them fairly into an unbiased dataset from a statistical standpoint before deeming them as issues requiring attention.

Users anticipate that the spam model safeguarding their email will make occasional errors. A report stating that it failed to accurately classify an email provides limited

utility to users and vendors, often necessitating only a minor update to the training dataset. Even when scaled up with thousands of missed emails, it may not significantly impact processes or code. The report becomes valuable when it outlines an easily executable process to circumvent the spam filter. Detailed evasion process reports may introduce operational requirements that extend beyond dataset updates and scheduled retraining, impacting the model's functionality.

However, there are instances that are entirely unacceptable to the model developer, elevating them instantly from incidents to issues. This spectrum ranges from rare, not-even-once incidents that qualify as issues to billions of data points collectively pushing the needle into issue territory. Identifying where on this spectrum an incident falls necessitates a case-by-case evaluation and is subject to various potential failures.

Intent of ML Model is Hard to Quantify and Test Navigating ethical questions in disclosure adds complexity. Consider a model claiming demographic unbiasedness, verified by an independent auditor. If later discovered to be biased against an unexamined group, disclosure becomes crucial. Members of the affected group and users must adapt their use with this knowledge. The vendor deserves an opportunity to address and rectify the issue. Simultaneously, the independent auditor should enhance coverage for future models.

Yet, addressing ethical challenges, like bias, poses a drawback. Reports on such matters are challenging to assess. A report for a purportedly unbiased model may itself be biased, stemming from model mistakes or other biases. Distinguishing a biased report from legitimate ones becomes a challenge for external observers. While sampling biases can be refuted with a comprehensive model performance view, determining legitimacy may be impossible. Users need not alter behavior if the report is illegitimate, provided they already account for known error rates. The vendor and auditor need minimal behavioral changes, primarily preparing statements if the submitter opts to publish the report.

Ethical and Safety Questions are Hard to Judge A pertinent example illustrating the mismatch with traditional vulnerability disclosure involves Facebook's site integrity models (Noorshams, Verma, and Hofleitner 2020). These models address issues like hate speech, spam, and harassment, impacting the system. The intricate site integrity rulebook governs the trust and safety system, encompassing machine learning models and human reviewers. The complex objective function covers socioeconomic factors for billions across diverse social groups.

Evaluating if a model's behavior violates this intricate code, distinguishing it from an understandable mistake in the moderation system, presents challenges. Submitters may create reports that, while seeming damning, fail to reveal a flaw in the moderation system. Vendors might reject valid reports on unverifiable statistical grounds, asserting the system operates within parameters. Both scenarios hinder the public – one erodes trust in an effective system, the other

avoids fixing real issues.

2.3 Absence of Unified Product or Weakness Enumeration

Since the inception of the Common Vulnerabilities and Exposures (CVE) framework, an entire ecosystem has developed around it. In the National Vulnerability Database, a comprehensive CVE entry is now expected to encompass three key components:

1. **Common Platform Enumeration (CPE)** – Serving as a standardized naming scheme, CPE facilitates the identification of products that may contain vulnerabilities (Radack 2010).

2. **Common Weakness Enumeration (CWE)** – A collaborative initiative that meticulously enumerates weaknesses found in various products (MITRE 2024a).

3. **Common Vulnerability Scoring System (CVSS)** – An objective scoring system that quantifies the severity of a CVE through a numerical score (FIRST 2024).

These components, much like the CVE itself, significantly streamline the process of communicating and managing vulnerabilities within the cybersecurity landscape. The CPE allows organizations to quickly determine if their systems are susceptible to a particular CVE. Meanwhile, the CWE acts as a universal language for understanding the root causes of CVEs, thereby enhancing remediation efforts. Formally established in the mid-2000s, years after the CVE's introduction in 1999, these components have become integral to the overall CVE ecosystem.

In the realm of machine learning ethics, several independent efforts have emerged to develop a classification system for potential harms caused by ML models (Weidinger et al. 2022; Shelby et al. 2023; Newman 2023; Graziani et al. 2023). Concurrently, within the domain of ML security, initiatives like the OWASP Top 10 for Large Language Models (LLMs) (OWASP 2021a) and ATLAS (MITRE 2024b) have gained prominence. Despite these strides, the ML community has not yet reached a consensus on a universally accepted structured harm reporting taxonomy. This challenge, however, reflects a familiar trajectory seen in the evolution of the CWE project. Drawing inspiration from contributions such as the OWASP Top 10 project (OWASP 2021b) and notable endeavors like PLOVER (Christey 2005), the CWE project successfully integrated diverse initiatives within focused, multi-stakeholder communities. This historical precedent suggests that the establishment of a standardized taxonomy in the ML community is indeed achievable.

3 Disclosure Practices in Machine Learning Systems: Evolving Trends

This section examines trends in disclosure practices for machine learning (ML) systems and the work needed to harmonize these conventions. We analyze two axes: objective security and privacy flaw disclosures, and subjective ethical flaw disclosures, such as societal bias.

3.1 Security Disclosures for ML Systems

ML systems face three main types of disclosures: security vulnerability reports, model design failures, and error reports.

Infrastructure Reports These reports highlight security vulnerabilities in ML models, often in components around the model. Examples include CVE reports on Numpy (CVE 2019a), Tensorflow (CVE 2019b), and Pytorch (CVE 2023c). The OpenAI bug bounty on Bugcrowd addresses infrastructure flaws (OpenAI 2023). SafeTensors by Hugging Face (Hugging Face 2022) is an emerging effort to address reproducibility concerns.

An instance of a traditional vulnerability is DeepSloth (Hong et al. 2020), a Denial Of Service attack on LLMs. ProofPudding (MITRE 2019) was the first CVE issued for an issue caused by an ML model. The mitigation involved limiting API information (Oksuz, Halimi, and Ayday 2023).

Model Design Failure Reports Model design failures manipulate samples to induce changes in output. Unlike infrastructure reports, they rely on modifying model output. Examples include the "Bayesian Poisoning Attack" (The MIT 2004 Spam Conference) and the Cylance malware detector (Cyber 2019). The Malware Detection Evasion Competition (Anderson 2019) highlights such vulnerabilities.

Adversarial examples are often considered low priority (Szegedy et al. 2014). They can be defeated by simple techniques like jpeg compression (Aydemir, Temizel, and Temizel 2018) or blur & repair (Carlini et al. 2022a). While adversarial patches (Brown et al. 2018) are possible, they are hard to execute (Athalye et al. 2018). Marines bypassed an ML model with a simple attack (Scharre 2023b), highlighting that lower-hanging fruit is often more appealing (Apruzzese et al. 2022).

Error Reports Error reports stem from expected model errors or systematic errors amendable through data changes. These include:

- **One-off Errors:** Normal errors within the model's expected error rate.
- **Systematic Errors:** Errors from dataset biases, leading to harmful outputs. Examples include the Google image classifier's mislabeling (Scharre 2023a) and the ChatGPT divergence attack (Nasr et al. 2023).

Reports that Depend on Context Recent reports include prompt injection attacks (NVIDIA 2023), with vulnerabilities in LangChain receiving CVEs. These can be "model design failure reports" or "error reports" depending on the context. For instance, basic chatbots might not be vulnerable, but complex chatbots must have measures to address these risks (Burgess 2023). Proper disclosure is crucial for systems granting sensitive access, such as email or calendar functions.

3.2 Disclosure of Ethical Issues in ML Systems

Despite calls for audits from various stakeholders (Raji et al. 2022; Fjeld et al. 2020; Wilkinson et al. 2023), the identification and disclosure of ethical issues in machine learning (ML) systems have been ad hoc and reactive. Unlike traditional software systems with established practices for coordinated vulnerability disclosure, ML model issues often surface inconsistently. Without a mandated reporting mechanism, harms caused by flawed ML systems are often discovered post-damage (Scharre 2023a). Few proactive mechanisms exist for auditing these complex systems, identifying pre-deployment risks, and enabling responsible disclosure balancing organizational and community interests. However, approaches like cooperative audits (Wilson et al. 2021), participatory audits (Dennler et al. 2023), and AI red teaming (OpenAI 2023; Ganguli et al. 2022) are emerging as ML stakes grow. This section traces the history and trends in disclosing ethical issues in ML.

Reporting has primarily relied on affected communities raising complaints through social media, media coverage, lawsuits, and public pressure after issues arise. Few systematic mechanisms audit and identify model issues proactively. Examples include the racial bias in COMPAS recidivism prediction algorithms uncovered by investigative journalism (ProPublica 2016) and facial analysis systems misclassifying people of color until public backlash erupts (BBC News 2015). Harmful biases in language models generating toxic outputs against marginalized groups are often observed due to vigilance within directly impacted communities (Cosmos Magazine 2023; Abid, Farooqi, and Zou 2021).

The lack of coordinated discovery and disclosure processes and the closed-source nature of many AI systems mean users remain unaware of problems until they surface. Reactive reporting by harmed parties has made accountability an exception rather than the norm for ML systems. Researchers and practitioners emphasize the need for proactive, community-centered mechanisms for responsible disclosures. Providing legal protections for independent ML issue researchers, akin to protections for good faith security research, is essential (Hacking Policy Council 2023).

3.3 Participatory Approaches for Issue Discovery

In response to growing scrutiny around AI ethics and the limitations of ad hoc reporting, organizations are experimenting with participatory approaches to systematically discover model issues.

Bias Bounty Programs Bias bounty programs incentivize crowd-sourced identification of algorithmic problems by offering rewards and recognition for impactful findings. Early efforts include Twitter's algorithmic cropping bounty (Yee, Tantipongpipat, and Mishra 2021) and QueerInAI's queer bias bounty (Dennler et al. 2023). AI bug bounties are becoming common in large companies (Google 2023; Microsoft 2023).

Community Driven Audits Community-driven audits, petitions, and pressure campaigns are effective for advocacy groups to highlight urgent ML issues and hold organizations accountable (Cattell, Chowdhury, and Carson 2023; The New York Times 2023). Truly participatory approaches require investments in inclusion to lower barriers for directly affected groups (Dennler et al. 2023). Transparency in model development and deployment fosters constructive feedback.

However, relying on ML model owners for authorization can undermine researchers' independence (Friedler et al. 2023).

Crowdsourced AI Harm Databases   Standardized documentation of AI harms is nascent. Efforts like MITRE's ATLAS[2], AI Vulnerability Database (AVID)[3], and AI Incident Database (AIID)[4] are well-known. AIID collects news stories about AI incidents, while AVID assesses open-source models. These initiatives lack coordinated disclosure features and vendor rebuttal mechanisms. Bounty programs like Google's and Microsoft's show an embrace of disclosure at large model vendors (Google 2023; Microsoft 2023).

AI Red Teaming   AI red teaming involves organizations commissioning external experts to rigorously test and attempt to attack their ML systems, simulating real-world malicious actors (OpenAI 2023; Ganguli et al. 2022). Goals include pressure-testing models, surfacing overlooked issues, validating internal safety claims, and gaining new perspectives to strengthen systems before public release (Selbst 2021; Moss et al. 2021; Carlini et al. 2023; Zou et al. 2023). Effective red teams require diverse professional backgrounds, ideologies, regional representations, and skill sets.

While promising, these community-centered mechanisms are not yet the norm. Much work remains in formalizing scalable, sustainable, and equitable participatory practices for discovering and disclosing ML issues.

4   Proposed Changes to Recognition of a Disclosure

In traditional coordinated disclosure for software vulnerabilities, if a vendor rejects acknowledging a reported issue, the only recourse for the submitter is to publicly disclose the problem and advocate for recognition. While this can be effective in drawing attention to and resolving vulnerabilities, the dynamics change when dealing with ML-based systems.

ML disclosures often involve ethical considerations, particularly regarding political speech. The challenge lies in establishing a process that fairly evaluates submissions, distinguishing substantiated claims from unsupported ones aiming to serve the submitter's agenda. Accepting ethical issue reports based on sampling bias can erode public trust in the vendor, while rejecting valid issues supported by statistically sound data damages both public trust in the vendor and the entity responsible for the flawed rejection.

To address these challenges, we propose a Coordinated Flaw Disclosure (CFD) process that incorporates:

1. **Extension of Model Cards:** Model Cards, traditionally used for complete systems, will be extended to provide detailed documentation of intent and scope.

2. **Trusted Independent Adjudication Panel:** A panel of independent adjudicators will be established to mediate disputes between the submitter and vendor when seeking official recognition.

---

[2]https://atlas.mitre.org/
[3]https://avidml.org/
[4]https://incidentdatabase.ai/

3. **Automated Verification through Submission Extension:** An extension of the submission concept will enable automatic verification of the potential issue's existence, enhancing the efficiency of the recognition process.

The main difference between this and the current CVE process is that it is based on the specific model or system card, not a universal definition of vulnerability. Bug bounty programs usually have a "scope" document associated with each vendor that dictates what is and is not reportable. The CVE process does not have this pre-defined scopes for a system; the definition of the vulnerability and community norms manage it. Vendors may argue for a lower CVSS score from MITRE if they believe the vulnerability is not relevant. However, there are flaws that are reputable in one model but not in another due to the creator's intent. Llama2 has an extensive model card (Meta 2023), but Mistral is meant to be a base model that is fine-tuned and embedded in another system, and the model card reflects that (AI 2023). Mistral's card is quite sparse and incomplete for a process such as this.

The main component is the trusted independent adjudication panel to govern the disputes that will arise between the submitter and vendor. This is a key part of the official CVD processes, as demonstrated by the "root" in the CVE Numbering Authority (CNA) rules for 3rd party CVE enumeration (Board 2020), the Bugcrowd appeals process (BugCrowd 2024), and the HackerOne Mediation process (Hackerone 2024). This will provide an outlet that fairly hears the submitter but also the vendor. If there's sensitive proprietary information from the vendor that refutes the issue, the adjudicator can review it and resolve the dispute without a public fight. However, if this data does not truly dispute the submitter, it can award the issue and thus start the mitigation process at the vendor.

We hope that this process will enable vendors to create safer and more secure models, and the public to understand them better. We expect the discovered issues that arise in the course of a model system's development to impact its Model Card to make it more precise, thus enabling everyone to better understand the capabilities of the system.

4.1   Complete Model/System Card

The vendor is required to provide a comprehensive model card for the entire system, detailing both the intent and scope. The intent encompasses the system's purpose and objectives, while the scope delineates the limitations and known issues associated with the system. This requirement aligns with the guidance established in the seminal work by Mitchell et al. (Mitchell et al. 2019) and is exemplified in various model cards, such as the Llama2 card (Meta 2023).

The intent section should include a clear mission statement for the system and outline the efficacy measures relevant to the model. These elements are expected to evolve as the model's performance and its contextual understanding improve. The scope section should address both current issues being addressed and those that are unresolved within a reasonable timeframe. These unresolved issues may be categorized into fundamental flaws that cannot be rectified and known flaws that are under ongoing development but

may take several years to resolve.

This approach effectively distinguishes between two aspects of "scope" commonly used in the security domain. While "intent" encompasses the inherent fallibility of the system, "scope" in security contexts often refers to vulnerabilities within the specified parameters. The intent is generally understood through product descriptions and community feedback. Given that models with identical input and output formats may serve different purposes, a clearly defined intent statement helps clarify the system's objectives and minimizes the need for extensive scope details. For instance, the Microsoft bounty program illustrates this distinction by excluding prompt injections from scope unless they access privileged information from other users.

A complete system card encompasses all components of the system, including both vulnerable models and those designed to mitigate such vulnerabilities. For example, integrating a diffusion model with an image classifier might address adversarial examples, with the model card specifying this capability as a system property rather than an attribute of the individual model (Carlini et al. 2022b). In this scenario, adversarial examples would be considered "in-scope" and reportable. Similarly, in the case of an LLM, deploying a hate speech detection model alongside the LLM could prevent the generation of hate speech, with the LLM model card noting the potential for hate speech, while the system card indicates that hate speech is effectively blocked. Another example includes Dalle2's use of demographic descriptors to enhance output diversity (OpenAI 2022).

We now further elaborate on what "Intent" and "Scope" mean in the context of our proposed framework.

Intent To make "intent" actionable, the system card should include a mission statement for the model system, and as many precise efficacy measures as possible.

Examples of Intent Mission statements:

1. From a text prompt, we produce images that are safe for work, safe for children, and free of demographic bias.

2. This malware model classifies Portable Executable (PE) binaries into malicious or benign. We consider Potentially Unwanted Applications, such as adware, to be malicious.

Examples of Intent measures:

1. We reproduce images of professionals in particular careers at the demographic rates of the US in this Text to Image model.

2. We have a 0.001 False Positive Rate and a 0.1 False Negative Rate on each customer's cumulative data for this malware model, if that customer has over 100 endpoints.

Scope The main purpose of Scope is to exclude known issues that the vendor has no intent or ability to resolve in the next year. There are intents, like Mistral's release of a base model for use in a larger system, that essentially exclude many reportable flaws. The two scope documents from the Google and Microsoft AI Bounty programs provide good examples of in and out-of-scope statements.

1. L2 norm adversarial examples are a known vulnerability in this image model; do not use for sensitive applications.

2. This is an LLM with no protections; prompt injections are out of scope.

3. Binary packing is a known vulnerability of this malware detection model, and we have no plans to build mitigations.

4.2 Adjudication Process

The vendor should not have sole authority over the recognition of an issue, especially when they are the only entity that can verify the issue's existence. Additionally, reports should be carefully vetted to ensure that the issue truly exists, is a valid violation of the model card, and is not a result of sampling bias in the provided report samples. We acknowledge that this is a significant deviation from the CNA process used to grant vendor CVEs and anticipate that it will be a topic of discussion.

The primary function of this committee is to distinguish between issues that involve valid violations of the model's intent and scope and those that do not. One challenge arises when an unbiased model used in hiring or image generation may not precisely align with a demographic split. If the model card does not allow for an error in this measurement, it would be considered a violation in the strictest sense. However, if the error is within industry standards, a report detailing this may not be officially recognized.

There are instances where the decisions made by the model may violate one of the stated intents of the model but align with another. Achieving complete political unbiasedness may be challenging, especially if one political group violates another policy far more frequently than another group.

4.3 Automatic Verification

Model errors are anticipated to constitute a significant proportion of reports in any database of AI system issues. A structured reporting mechanism, such as a telemetry system, offers several advantages to vendors. Primarily, it facilitates seamless integration of reported issues into the model's training and testing datasets. This approach also ensures that the reported behavior is verifiable and reproducible—a critical consideration given the potential for daily fluctuations due to model updates and continuous training processes.

The majority of substantive reports, as outlined in the background section, include demonstrative samples of the problematic behavior. Therefore, any system supporting Coordinated Flaw Disclosure should incorporate functionalities for automatic external verification and integration into the vendor's internal testing dataset. This approach significantly reduces the vendor's resource expenditure in issue remediation and provides a mechanism for ongoing verification of issue resolution.

Upon acceptance of a reported issue, the system should preserve the supporting data and implement periodic re-verification protocols. This is particularly crucial as issue resolution in one model iteration does not guarantee its absence in subsequent versions, potentially leading to regression and issue recurrence. By maintaining a robust verification system, vendors can ensure continued model integrity and swiftly address any re-emerging issues.

Coordinated Flaw Disclosure

General information
Information about the vendor, system, reporter.

Description of the Flaw
Description of the uncovered flaw. Could be a infrastructure report or a model design failure report. Should cover whether the flaw refers to a one-off error or systematic error.

Information about the model
Description of how this flaw contradicts the intent and scope reported in the model card of the system.

Sample Set	Output
Example sample	*Example model output*

Figure 1: A potential template for a CFE report by a user.

4.4 Dynamic Scope Expansion for Common Use

For general-purpose models[5] with potentially limitless applications, we propose a "Common Use" clause in the model card. This clause allows for the dynamic expansion of the model's intent and scope based on widespread adoption of certain use cases, even if they were not originally defined.

Initially, individual companies should strive to identify and incorporate common uses of their models. However, as the field matures, this process could evolve into a centralized effort, potentially managed by a cross-vendor coalition or an independent body. Such an approach could lead to the development of a Common Use Enumeration (CUE), complementing existing systems like the Common Weakness Enumeration (CWE).

To implement this approach:

1. **Define a threshold for "common use":** Establish quantitative criteria (e.g., X% of users or Y number of deployments) to determine when a use case becomes 'common'.

2. **Regular review process:** Implement a periodic review (e.g., quarterly or bi-annually) to identify emerging common uses of the model.

3. **Scope expansion:** For uses meeting the defined threshold, expand the model card to include these as in-scope for flaw reporting and mitigation.

4. **Allow reporting:** Enable the submission of issues related to these expanded use cases through the CFD process.

The Common Use clause acknowledges that while general-purpose models may have vast potential applications, safety and responsible development require some boundaries. By allowing for dynamic scope expansion, we create a

---

[5]also called foundation models or frontier models, some of these models are additionally open-weight with permissive licensing and have led to an ecosystem of downstream use cases, such as Meta's LLAMA family of models (https://llama.meta.com/).

framework that can evolve with the actual use patterns of the model, enhancing the overall effectiveness of the CFD process. As the AI field continues to advance, the development of a centralized, independent CUE system could further streamline this process and promote industry-wide consistency in addressing common uses of AI models. We further discuss a CUE system in the Future Work section of this paper.

5 Issue Reporting Governance

In the following, we describe the process of submitting a CFD. The implementation of the process depends on the vendors, however, their clear advantage of formalizing the process is similar to the adoption of the previous CVD process; vulnerabilities are explored before they can be exploited. Where CFD-related issues are identified, we can assume that it is in the best interests of the vendors to be notified about those to be able to resolve them in a timely manner. Our proposed CFE issuance process (Section 5.2) takes this into account, and proposes a process to resolve potential issues between reporters and vendors.

5.1 Submitter Interaction

Users identify potential issues believed to violate the scope and create a submission with:

1. The sample set (even a single sample is acceptable)

2. The output(s) the model should not generate.

The report argues that the sample output pairs violate the scope. After submission, the report is sent to the vendor for acceptance or rejection. If rejected, the user can appeal to the Adjudicators.

We propose a reporting template in Figure 1. The submitter should provide information about the vendor, system, and the person reporting the issue. The description of the uncovered flaw should include whether this is a infrastructure or model design failure report, and whether it is a one-off error, or a systematic error. For more details on the error types, see Section 3.1. To give the vendor context about how to address the issue and a possibility to prioritize the report, the report should refer to the model card of the model. Finally, the template should have the possibility to submit sample set and according output, depending on the size either as a table or as structured in a common data format.

5.2 CFE Issuance Process

An overview of the CFE issuance process can be seen in Figure 2. Upon submission of a flaw, the following steps are followed:

1. **Triage**: The vendor or an authorized person reviews the report to ensure the sample and output pairs satisfy the report.

   Note: If not, the submission is rejected on technical grounds.

2. **Review**: If the report and sample/output pairs align, the vendor determines if the report is within Scope and violates the model Intent.

Note: If not, the submission is rejected as either out-of-scope or in-intent.

Caveat: If the data supports the report but is incomplete or flawed, the vendor can reject on statistical grounds or request additional data.

3. **Adjudication**: If rejected for any reason, the Submitter can appeal to the adjudicator panel.

The vendor may reject the submission on statistical grounds to the adjudicator panel. If the submitter-supplied sample/output pairs are statistically biased, and an unbiased sample set would not show a violation of the model card, the adjudicator panel may request supporting data from the vendor to validate the rejection. The adjudicator and vendor can jointly assess whether the vendor-supplied data is too sensitive for the submitter and decide on further steps. An example of this cooperation is once again the "divergence attack" paper by Google (Nasr et al. 2023). The authors followed a standard 90 day responsible disclosure practice[6] in their private disclosure of the bug to OpenAI and publicly reported the bug only after the 90 days were up.

5.3 Adjudication Process for Edge Cases

While the CFD process primarily focuses on issues within the defined scope and intent of a model, the dynamic nature of AI systems, particularly general-purpose models, necessitates a mechanism for addressing edge cases. These edge cases often arise from common uses that were not initially anticipated in the model card. To address this, we propose an extended adjudication process:

1. **Initial Triage:** When a report is submitted that falls outside the stated intent but appears to be within common use, it is flagged for special review.

2. **Common Use Assessment:** The adjudication panel first determines if the reported use case meets the criteria for 'common use' as defined in the Dynamic Scope Expansion clause (Section 4.4).

3. **Harm Evaluation:** If the use case is deemed common, the panel assesses the severity and prevalence of the reported harm.

4. **Vendor Consultation:** The panel consults with the vendor to understand any technical constraints or considerations related to the reported issue.

5. **Decision:** Based on these factors, the panel may:
   - Recommend immediate inclusion of the issue in the CFD process
   - Suggest updating the model card to include this use case in future versions
   - Advise on interim mitigation strategies
   - Determine that the issue, while outside of scope, warrants public disclosure due to its potential impact

6. **Follow-up:** The panel ensures that any decisions are implemented, including updates to the model card, CFD database, or public disclosures as necessary.

---

[6]https://not-just-memorization.github.io/extracting-training-data-from-chatgpt.html#responsible-disclosure
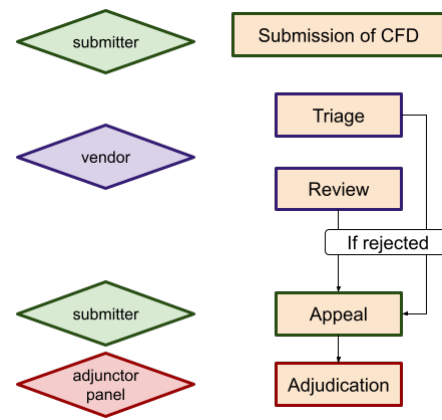

Figure 2: Flowchart of the CFE issuance processes described in Section 5.2.

This extended process balances the need for clear boundaries with the reality of evolving AI applications, providing a structured approach to handling edge cases that may otherwise fall through the cracks of traditional disclosure processes.

5.4 Adjudicator Panel

Current policy proposals focus on providing audits by external experts (Hartmann et al. 2024). These audits have a very wide range of tasks and possible restrictions to apply. In the CFD process, these already planned panels could also act as adjudicator panels in the CFD processes. It would give the audits a starting point about contested issues, and could be a starting point for the implementation of the CFD process. It is important to emphasise that we expect the vast majority of CFD to be resolved between the reporters and AI providers directly as it is the case with the CVE processes (Red Hat 2022; MITRE Community Board 2014), keeping the workload of the adjudicator panel light.

6 Conclusion

This paper introduces a Coordinated Flaw Disclosure (CFD) framework for machine learning and AI systems, addressing algorithmic flaws through transparent intent and scope definitions via extended Model Cards. The CFD framework balances submitter and vendor interests, fostering trust and transparency through structured issue reporting governance. By considering both quantitative and qualitative evidence, it captures a holistic view of AI system flaws. We believe widespread adoption of this framework can bolster public trust in AI systems by promoting clear processes and collaborative engagement. As machine learning evolves, the CFD framework offers a structured, adaptable approach to navigate the changing landscape of algorithmic flaws, benefiting both vendors and the public.

6.1 Future Work

While our proposed CFD framework lays the groundwork for improving AI accountability and transparency, there

are several areas where further development and real-world testing are needed. We outline two key areas for future work: an implementation pilot to test the CFD framework in practice, and the development of a Common Use Enumeration system to standardize the tracking of AI model applications.

Implementation Pilot at DEF CON 32   In collaboration with key stakeholders from government, industry, freelance security researchers, and policy experts, we are organizing the Generative Red Team 2 (GRT2)[7] event at DEF CON 32. This event serves as a comprehensive test of the CFD framework's robustness and applicability in real-world scenarios. Key aspects of GRT2 that align with and test our CFD framework include:

- **Model Cards:** Participants receive detailed model cards specifying the intent and scope of the target AI system, mirroring our proposed extended documentation approach.
- **Comprehensive Flaw Reports:** Using the UK AI Safety Institute's Inspect AI framework[8], participants prepare statistically valid reports on model flaws, moving beyond single-example vulnerabilities.
- **Dedicated Submission Platform:** Reports are submitted through a platform, demonstrating the feasibility of specialized CFD infrastructure.
- **Vendor Interaction:** AI system vendors review and respond to submitted reports, simulating real-world disclosure dynamics.
- **Independent Adjudication:** A panel of independent experts mediates disputes between submitters and vendors, testing the efficacy of our proposed adjudication process.
- **Transparency:** All data generated during the event will be made public, aligning with our goals for increased transparency in AI flaw disclosure.

GRT2 is designed not only to validate the CFD framework but also to stress-test it, identifying potential weaknesses and areas for improvement. The results and lessons learned from GRT2 will be invaluable in refining the CFD framework.

Common Use Enumeration (CUE) System   To address the challenge of tracking and standardizing common uses of AI models, we propose the development of a Common Use Enumeration (CUE) system. This system would serve as a centralized repository for documenting and categorizing the various applications of AI models, particularly focusing on general-purpose or foundation models that may have a wide range of unforeseen uses.

Proposed Structure of the CUE System

1. **Hierarchical Classification:** The CUE system would organize AI model uses into a hierarchical structure, similar to the Common Weakness Enumeration (CWE) system. This would allow for broad categories (e.g., "Natural Language Processing", "Computer Vision") with increasingly specific subcategories.

2. **Unique Identifiers:** Each documented use case would receive a unique identifier (e.g., CUE-001) for easy reference and tracking.

3. **Metadata Fields:** Each entry would include relevant metadata such as description of the use case, associated model types, potential risks or ethical considerations, known implementations or case studies, and related CUEs or CWEs.

4. **Version Control:** The system would maintain version history to track how use cases evolve over time.

Management and Governance   We propose the establishment of a multi-stakeholder governing body to oversee the CUE system, modeled after established multi-stakeholder processes in the CVE ecosystem. This body would manage a standardized submission process, regular updates, and ensure open access to the CUE database. Vendors could integrate CUE identifiers into their model cards, use the system for reporting new uses and risk assessment, and access an API for integrating CUE data into their development and monitoring tools.

The development of a Common Use Enumeration system represents a significant undertaking that would require collaboration across the AI industry, academia, and regulatory bodies. However, such a system could greatly enhance our ability to track, understand, and responsibly manage the diverse applications of AI models in society.

6.2   Limitations

While the CFD framework presents a significant step forward, several challenges remain:

- **Model Cards Standardization:** The lack of standardized model cards poses a challenge for widespread CFD adoption. Establishing standards for model cards is crucial for effective implementation, as emphasized by the NIST AI Risk Management Framework (Singh 2021).
- **Dynamic ML Datasets:** Tracking and managing evolving datasets used in model training presents complex challenges. Unlike software dependencies tracked by SBOMs (CISA 2021; Corman 2022), monitoring dynamic ML datasets lacks established tools. Recent incidents, such as the LAION-5B dataset retraction (Schuhmann et al. 2022; LAION.ai 2023; Thiel 2023), highlight the need for robust dataset management in machine learning.
- **Contextual Challenges:** The context-specific nature of AI systems complicates the development of standardized weakness and use enumerations. For instance, while a system might be theoretically vulnerable to "Prompt Injection" (MITRE 2024b), its specific application might mitigate this risk. Identical ML models used for different purposes add complexity, making model documentation crucial for understanding relevant weaknesses and uses across various contexts.

Addressing these limitations requires ongoing collaboration between AI developers, security researchers, and domain experts to create flexible, adaptive systems that accurately capture the nuances of AI weaknesses and uses across diverse applications.

---

[7]https://grt.aivillage.org/announcement
[8]https://inspect.ai-safety-institute.org.uk/

7 Research Ethics and Social Impact

Our interdisciplinary team, consisting of AI ethicists and security researchers, navigates the intersection of AI ethics and cybersecurity, drawing inspiration from established practices like the Common Vulnerabilities and Exposures (CVE) process. We recognize the ethical imperative of mitigating harms associated with vulnerability reporting, aligning with the broader goal of enhancing model transparency and promoting the development of safer AI systems.

7.1 Ethical Considerations

In adherence to ethical research principles and community norms, we disclose potential challenges and ethical considerations associated with our work. No human subject experiments were conducted, and therefore an IRB was not required.

7.2 Author Positionality

As authors with a foundation in privacy and cybersecurity, we have witnessed the positive outcomes of the Coordinated Vulnerability Disclosure (CVD) process for software projects, both small and large. This proposal advocates for extending the CVD process to the field of machine learning research, aiming to bridge the gap and introduce machine learning researchers to the benefits of a familiar and effective process. Our positionality as proponents of CVD in this context stems from our understanding of its commercial advantages and positive impact on software security.

7.3 Adverse Impact

In the majority of cases, the CVD process operates smoothly without adverse consequences. However, we acknowledge instances where a vendor may reject an issue without engagement, addressing it quietly without public acknowledgment. While such occurrences are limited, they emphasize the need for a framework that encourages responsible disclosure and engagement, ensuring transparency and accountability in the machine learning community.

References

2019a. CVE-2019-6446. Available from MITRE, CVE-ID CVE-2019-6446.

2019b. CVE-2022-23560. Available from MITRE, CVE-ID CVE-2022-23560.

2019. Cylance Antivirus Products Susceptible to Concatenation Bypass. Available from CERT, Vulnerability Note VU#489481.

2020. CVE-2019-20634. Available from MITRE, CVE-ID CVE-2019-20634.

2023a. CVE-2023-1304. Available from MITRE, CVE-ID CVE-2023-1304.

2023b. CVE-2023-29374. Available from MITRE, CVE-ID CVE-2023-29374.

2023c. CVE-2023-43654. Available from MITRE, CVE-ID CVE-2023-43654.

2023. EU Cyber Resilience Act. https://digital-strategy.ec.europa.eu/en/policies/cyber-resilience-act.

116th Congress. 2020. Internet of Things Cybersecurity Improvement Act of 2020. https://www.congress.gov/116/plaws/publ207/PLAW-116publ207.pdf.

Abid, A.; Farooqi, M.; and Zou, J. 2021. Persistent anti-muslim bias in large language models. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, 298–306.

ACLU. 2022. ACLU v. Clearview AI — American Civil Liberties Union — aclu.org. https://www.aclu.org/cases/aclu-v-clearview-ai. [Accessed 23-01-2024].

AI, M. 2023. https://mistral.ai/news/announcing-mistral-7b/. https://mistral.ai/news/announcing-mistral-7b/.

Allison, F. 2023. OpenAI Vendor Lock-in: The Ironic Story of How OpenAI Went from Open Source to "Open Your Wallet". https://www.lunasec.io/docs/blog/openai-not-so-open/.

Anderson, H. 2019. Machine Learning Static Evasion Competition. https://github.com/endgameinc/malware_evasion_competition. Accessed: 17 Jan 2024.

Apache. 2024. ASF Project Security for Committers. https://www.apache.org/security/committers.html.

Apruzzese, G.; Anderson, H. S.; Dambra, S.; Freeman, D.; Pierazzi, F.; and Roundy, K. A. 2022. "Real Attackers Don't Compute Gradients": Bridging the Gap Between Adversarial ML Research and Practice. arXiv:2212.14315.

Athalye, A.; Engstrom, L.; Ilyas, A.; and Kwok, K. 2018. Synthesizing Robust Adversarial Examples. arXiv:1707.07397.

Authors Guild. 2023. Authors Guild Supports Nonfiction Writers in Lawsuit Against OpenAI - The Authors Guild — authorsguild.org. https://authorsguild.org/news/ag-supports-nonfiction-writers-in-lawsuit-against-openai/. [Accessed 23-01-2024].

Aydemir, A. E.; Temizel, A.; and Temizel, T. T. 2018. The Effects of JPEG and JPEG2000 Compression on Attacks using Adversarial Examples. arXiv:1803.10418.

Barbierato, E.; and Gatti, A. 2024. The Challenges of Machine Learning: A Critical Review. *Electronics*, 13(2): 416.

BBC News. 2015. Google apologises for Photos app's racist blunder. *BBC News*.

Board, C. 2020. CVE Numbering Authority (CNA) Rules.

Brown, T. B.; Mané, D.; Roy, A.; Abadi, M.; and Gilmer, J. 2018. Adversarial Patch. arXiv:1712.09665.

BugCrowd. 2024. Your Guide to Hacking with Bugcrowd and Why You Should Get Started Today. https://www.bugcrowd.com/resources/report/your-guide-to-hacking-with-bugcrowd-and-why-you-should-get-started-today/.

Burgess, M. 2023. The security hole at the heart of Chatgpt and bing.

Carlini, N.; Nasr, M.; Choquette-Choo, C. A.; Jagielski, M.; Gao, I.; Awadalla, A.; Koh, P. W.; Ippolito, D.; Lee, K.; Tramer, F.; et al. 2023. Are aligned neural networks adversarially aligned? *arXiv preprint arXiv:2306.15447*.

Carlini, N.; Tramer, F.; Dvijotham, K. D.; Rice, L.; Sun, M.; and Kolter, J. Z. 2022a. (certified!!) Adversarial robustness for free! *arXiv preprint arXiv:2206.10550*.

Carlini, N.; Tramer, F.; Krishnamurthy, D.; and Kolter, J. Z. 2022b. Certified! Adversarial Robustness for Free.

Cattell, S.; Chowdhury, R.; and Carson, A. 2023. AI Village at DEF CON Announces Largest-Ever Public Generative AI Red Team. https://aivillage.org/generative%20red%20team/generative-red-team/.

Christey, S. 2005. PLOVER - Preliminary List Of Vulnerability Examples for Researchers. *NIST Workshop Defining the State of the Art of Software Security Tools*.

CISA. 2021. RE: A Proposal for Identifying and Managing Bias within Artificial Intelligence (Spec. Pub. 1270).

Corman, J. 2022. Statement of Joshua Corman : For Senate Health, Education, Labor & Pensions Committee (HELP).

Cosmos Magazine. 2023. Busting homophobic, anti-queer bias in AI language models. *Cosmos Magazine*. Accessed on December 29, 2023.

Cyber, S. 2019. Cylance, I Kill You! https://skylightcyber.com/2019/07/18/cylance-i-kill-you/. Accessed: 17 Jan 2024.

Cybersecurity and Infrastructure Security Agency. 2020. Binding Operational Directive 20-01: Develop and Publish a Vulnerability Disclosure Policy. https://www.cisa.gov/binding-operational-directive-20-01.

Cybersecurity Coalition. 2019. Policy Priorities for Coordinated Vulnerability Disclosure and Handling. https://www.cybersecuritycoalition.org/reports/white-paper-policy-priorities-for-coordinated-vulnerability-disclosure-and-handling.

Dennler, N.; Ovalle, A.; Singh, A.; Soldaini, L.; Subramonian, A.; Tu, H.; Agnew, W.; Ghosh, A.; Yee, K.; Peradejordi, I. F.; et al. 2023. Bound by the Bounty: Collaboratively Shaping Evaluation Processes for Queer AI Harms. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, 375–386.

European Commission. 2020. Assessment List for Trustworthy Artificial Intelligence (ALTAI) Self-Assessment. url.

European Parliament and Council. 2022. Directive (EU) 2022/2555 of the European Parliament and of the Council. https://eur-lex.europa.eu/eli/dir/2022/2555.

European Parliament and Council. 2024. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations. https://eur-lex.europa.eu/eli/reg/2024/1689/oj.

European Union Agency for Cybersecurity (ENISA). 2022. Coordinated Vulnerability Disclosure Policies in the EU. https://www.enisa.europa.eu/publications/coordinated-vulnerability-disclosure-policies-in-the-eu.

European Union Agency for Cybersecurity (ENISA). 2024. Cybersecurity Certification Framework. https://www.enisa.europa.eu/topics/certification/cybersecurity-certification-framework.

Fenwick & West LLP. 2023. Artist's Copyright Claim against Stability AI Moves Forward Following a First-of-Its Kind Ruling — JD Supra — jdsupra.com. https://www.jdsupra.com/legalnews/artist-s-copyright-claim-against-4630588/. [Accessed 23-01-2024].

FIRST. 2024. Common Vulnerability Scoring System SIG. https://www.first.org/cvss/.

Fjeld, J.; Achten, N.; Hilligoss, H.; Nagy, A.; and Srikumar, M. 2020. Principled artificial intelligence: Mapping consensus in ethical and rights-based approaches to principles for AI. *Berkman Klein Center Research Publication*, (2020-1).

Friedler, S.; Singh, R.; Blili-Hamelin, B.; Metcalf, J.; and Chen, B. J. 2023. AI Red-Teaming Is Not a One-Stop Solution to AI Harms: Recommendations for Using Red-Teaming for AI Accountability. https://datasociety.net/library/ai-red-teaming-is-not-a-one-stop-solution-to-ai-harms-recommendations-for-using-red-teaming-for-ai-accountability/.

Ganguli, D.; Lovitt, L.; Kernion, J.; Askell, A.; Bai, Y.; Kadavath, S.; Mann, B.; Perez, E.; Schiefer, N.; Ndousse, K.; et al. 2022. Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned. *arXiv preprint arXiv:2209.07858*.

Google. 2023. Google's reward criteria for reporting bugs in AI products — security.googleblog.com. https://security.googleblog.com/2023/10/googles-reward-criteria-for-reporting.html. [Accessed 23-01-2024].

Graziani, M.; Dutkiewicz, L.; Calvaresi, D.; Amorim, J. P.; Yordanova, K.; Vered, M.; Nair, R.; Abreu, P. H.; Blanke, T.; Pulignano, V.; et al. 2023. A global taxonomy of interpretable AI: unifying the terminology for the technical and social sciences. *Artificial intelligence review*, 56(4): 3473–3504.

Hackerone. 2024. Hacker Mediation. https://docs.hackerone.com/en/articles/8466617-hacker-mediation.

Hacking Policy Council. 2023. AI Red Teaming - Legal Clarity and Protections Needed. https://assets-global.website-files.com/62713397a014368302d4ddf5/6579fcd1b821fdc1e507a6d0_Hacking-Policy-Council-statement-on-AI-red-teaming-protections-20231212.pdf.

Hartmann, D.; de Pereira, J. R. L.; Streitbörger, C.; and Berendt, B. 2024. Addressing the Regulatory Gap: Moving Towards an EU AI Audit Ecosystem Beyond the AIA by Including Civil Society. *CoRR*, abs/2403.07904.

Hong, S.; Kaya, Y.; Modoranu, I.; and Dumitras, T. 2020. A Panda? No, It's a Sloth: Slowdown Attacks on Adaptive Multi-Exit Neural Network Inference. *CoRR*, abs/2010.02432.

Hugging Face. 2022. Safetensors: Simple, safe way to store and distribute tensors. hhttps://huggingface.co/docs/safetensors/en/index. [Accessed 23-01-2024].

LAION.ai. 2023. Safety Review For LAION 5B. https://laion.ai/notes/laion-maintanence/.

Mac, R. 2023. The Times Sues OpenAI and Microsoft Over A.I. Use of Copyrighted Work — nytimes.com. https://www.nytimes.com/2023/12/27/business/media/new-york-times-open-ai-microsoft-lawsuit.html. [Accessed 23-01-2024].

Meta. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models. https://ai.meta.com/research/publications/llama-2-open-foundation-and-fine-tuned-chat-models/.

Microsoft. 2023. Microsoft AI Bounty — MSRC — microsoft.com. https://www.microsoft.com/en-us/msrc/bounty-ai. [Accessed 23-01-2024].

Mitchell, M.; Wu, S.; Zaldivar, A.; Barnes, P.; Vasserman, L.; Hutchinson, B.; Spitzer, E.; Raji, I. D.; and Gebru, T. 2019. Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency*, 220–229.

MITRE. 2019. NVD - CVE-2019-20634 — nvd.nist.gov. https://nvd.nist.gov/vuln/detail/CVE-2019-20634. [Accessed 23-01-2024].

MITRE. 2024a. Common Weakness Enumeration: A community developed list of common software and hardware weakness types. https://cwe.mitre.org/index.html.

MITRE. 2024b. MITRE — ATLAS; — atlas.mitre.org. https://atlas.mitre.org/. [Accessed 23-01-2024].

MITRE Community Board. 2014. Re: procedure for penalizing or revoking CNA status? https://cve.mitre.org/data/board/archives/2014-11/msg00000.html.

Moss, E.; Watkins, E. A.; Singh, R.; Elish, M. C.; and Metcalf, J. 2021. Assembling accountability: algorithmic impact assessment for the public interest. *Available at SSRN 3877437*.

MultiMedia LLC. 1999. MS Windows NT Kernel Description. http://web.archive.org/web/20080207010024/http://www.808multimedia.com/winnt/kernel.htm.

Nasr, M.; Carlini, N.; Hayase, J.; Jagielski, M.; Cooper, A. F.; Ippolito, D.; Choquette-Choo, C. A.; Wallace, E.; Tramèr, F.; and Lee, K. 2023. Scalable extraction of training data from (production) language models. *arXiv preprint arXiv:2311.17035*.

National Institute of Standards and Technology. 2018. Framework for Improving Critical Infrastructure Cybersecurity. https://nvlpubs.nist.gov/nistpubs/cswp/nist.cswp.04162018.pdf.

New York City Department of Consumer and Worker Protection. 2021. Automated Employment Decision Tools, 2021/144. https://www.nyc.gov/site/dca/about/automated-employment-decision-tools.page.

Newman, J. 2023. A taxonomy of trustworthiness for artificial intelligence. *CLTC: North Charleston, SC, USA*.

Nichols, S. 2021. Why cloud bugs don't get CVEs, and why it's an issue. https://www.techtarget.com/searchsecurity/news/252508948/Why-cloud-bugs-dont-get-CVEs-and-why-its-an-issue.

Noorshams, N.; Verma, S.; and Hofleitner, A. 2020. Ties: temporal interaction embeddings for enhancing social media integrity at facebook. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 3128–3135.

NVIDIA. 2023. Securing LLM Systems Against Prompt Injection — NVIDIA Technical Blog — developer.nvidia.com. https://developer.nvidia.com/blog/securing-llm-systems-against-prompt-injection/. [Accessed 23-01-2024].

O'Brien, M. 2023. Photo giant Getty took a leading AI image-maker to court. Now it's also embracing the technology — apnews.com. https://apnews.com/article/getty-images-artificial-intelligence-ai-image-generator-stable-diffusion-a98eeaaeb2bf13c5e8874ceb6a8ce196. [Accessed 23-01-2024].

Office of Science and Technology Policy. 2022. Blueprint for an AI Bill of Rights: Making Automated Systems Work For The American People. https://www.whitehouse.gov/wp-content/uploads/2022/10/Blueprint-for-an-AI-Bill-of-Rights.pdf.

Oksuz, A. C.; Halimi, A.; and Ayday, E. 2023. AUTOLYCUS: Exploiting Explainable AI (XAI) for Model Extraction Attacks against White-Box Models. arXiv:2302.02162.

OpenAI. 2022. Reducing bias and improving safety in DALL·E 2. https://openai.com/blog/reducing-bias-and-improving-safety-in-dall-e-2.

OpenAI. 2023. Bug Bounty: OpenAI — Bugcrowd — bugcrowd.com. https://bugcrowd.com/openai. [Accessed 23-01-2024].

OpenAI. 2023. GPT-4V(ision) system card. https://openai.com/research/gpt-4v-system-card.

OWASP. 2021a. OWASP Top 10 for Large Language Model Applications. https://owasp.org/www-project-top-10-for-large-language-model-applications/.

OWASP. 2021b. OWASP Top Ten. https://owasp.org/www-project-top-ten/.

ProPublica. 2016. Machine Bias: Risk Assessments in Criminal Sentencing.

Radack, S. 2010. Security Content Automation Protocol (SCAP): Helping Organizations Maintain and Verify the Security of Their Information Systems. *ITL Bulletin*.

Raji, I. D.; and Buolamwini, J. 2022. Actionable Auditing Revisited: Investigating the Impact of Publicly Naming Biased Performance Results of Commercial AI Products. *Communications of the ACM*, 66(1): 101–108.

Raji, I. D.; Xu, P.; Honigsberg, C.; and Ho, D. 2022. Outsider oversight: Designing a third party audit ecosystem for ai governance. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, 557–571.

Red Hat. 2022. Red Hat RootCNA Appeals Process. https://access.redhat.com/articles/red_hat_rootcna_appeals_process.

Scharre, P. 2023a. Google apologises for Photos app's racist blunder. https://www.bbc.com/news/technology-33347866. Accessed: 17 Jan 2024.

Scharre, P. 2023b. Marines Outwitted an AI Security Camera by Hiding in a Cardboard Box and Pretending to Be Trees. https://www.cnas.org/press/in-the-news/marines-outwitted-

an-ai-security-camera-by-hiding-in-a-cardboard-box-and-pretending-to-be-trees. Accessed: 17 Jan 2024.

Schuhmann, C.; Beaumont, R.; Vencu, R.; Gordon, C.; Wightman, R.; Cherti, M.; Coombes, T.; Katta, A.; Mullis, C.; Wortsman, M.; Schramowski, P.; Kundurthy, S.; Crowson, K.; Schmidt, L.; Kaczmarczyk, R.; and Jitsev, J. 2022. LAION-5B: An open large-scale dataset for training next generation image-text models. arXiv:2210.08402.

Selbst, A. D. 2021. AN INSTITUTIONAL VIEW OF ALGORITHMIC IMPACT. *Harvard Journal of Law & Technology*, 35(1).

Shelby, R.; Rismani, S.; Henne, K.; Moon, A.; Rostamzadeh, N.; Nicholas, P.; Yilla-Akbari, N.; Gallegos, J.; Smart, A.; Garcia, E.; et al. 2023. Sociotechnical harms of algorithmic systems: Scoping a taxonomy for harm reduction. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, 723–741.

Singh, S. 2021. RE: A Proposal for Identifying and Managing Bias within Artificial Intelligence (Spec. Pub. 1270).

Solaiman, I.; Talat, Z.; Agnew, W.; Ahmad, L.; Baker, D.; Blodgett, S. L.; Daumé III, H.; Dodge, J.; Evans, E.; Hooker, S.; et al. 2023. Evaluating the social impact of generative ai systems in systems and society. *arXiv preprint arXiv:2306.05949*.

Szegedy, C.; Zaremba, W.; Sutskever, I.; Bruna, J.; Erhan, D.; Goodfellow, I.; and Fergus, R. 2014. Intriguing properties of neural networks. arXiv:1312.6199.

The MIT 2004 Spam Conference. 2004. How to beat an adaptive/Bayesian spam filter.

The New York Times. 2023. When Hackers Descended to Test A.I., They Found Flaws Aplenty.

The White House. 2023. Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence. https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/.

Thiel, D. 2023. Identifying and Eliminating CSAM in Generative ML Training Data and Models.

U.S. Food and Drug Administration. 2023. Cybersecurity in Medical Devices: Quality System Considerations and Content of Premarket Submissions. https://www.fda.gov/media/119933/download.

Weidinger, L.; Uesato, J.; Rauh, M.; Griffin, C.; Huang, P.-S.; Mellor, J.; Glaese, A.; Cheng, M.; Balle, B.; Kasirzadeh, A.; et al. 2022. Taxonomy of risks posed by language models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 214–229.

Wilkinson, D.; Crawford, K.; Wallach, H.; Raji, D.; Rakova, B.; Singh, R.; Strohmayer, A.; and Zuckerman, E. 2023. Accountability in Algorithmic Systems: From Principles to Practice. CHI EA '23. New York, NY, USA: Association for Computing Machinery. ISBN 9781450394222.

Wilson, C.; Ghosh, A.; Jiang, S.; Mislove, A.; Baker, L.; Szary, J.; Trindel, K.; and Polli, F. 2021. Building and auditing fair algorithms: A case study in candidate screening.

In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 666–677.

Yee, K.; Tantipongpipat, U.; and Mishra, S. 2021. Image cropping on twitter: Fairness metrics, their limitations, and the importance of representation, design, and agency. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2): 1–24.

Zou, A.; Wang, Z.; Kolter, J. Z.; and Fredrikson, M. 2023. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*.