A Closer Look at the Existing Risks of Generative AI: Mapping the Who, What, and How of Real-World Incidents

Megan Li, Wendy Bickersteth, Ningjing Tang, Lorrie Cranor, Jason Hong, Hong Shen*, Hoda Heidari*

Carnegie Mellon University
{meganli, wbickers, ningjingt, lorrie, jasonh, hongs, hheidari}@andrew.cmu.edu

Abstract

Generative AI is applied in an ever-growing set of domains and tasks, leading to an expanding set of risks of harm impacting people, communities, society, and the environment. These risks may arise due to failures during the design and development of the technology, as well as during its release, deployment, or downstream usages and appropriations of its outputs. In this paper, building on prior taxonomies of AI risks, harms, and failures, we construct a taxonomy specifically for *Generative* AI failures and map them to the harms they're associated with in the real world. Through a systematic analysis of 499 publicly reported incidents of harm involving Generative AI, we describe *what* harms are reported and how often, *how* they tend to arise, and *who* they impact. We find that most reported incidents are associated with *use-related* failures but the harms are experienced by parties *beyond* the end user(s) of the system at fault, and that the landscape of Generative AI harms is distinct from that of traditional AI. Our work offers actionable insights to policymakers, developers, and end users. In particular, we call for the prioritization of non-technical risk and harm mitigation strategies that center responsible *use*, including public disclosures, AI literacy efforts, and careful regulatory stances.

Coded incident dataset —
https://doi.org/10.7910/DVN/NISSFZ

1 Introduction

Generative AI—defined as AI that produces novel output in the form of text, images, audio, or video (Weidinger et al. 2023)—has taken the world by storm, presenting both risks and benefits to humanity. While there is broad consensus on the need to mitigate the risks of Generative AI, a growing divide separates those who prioritize its *existential* threats from those who focus on its *existing* risks. Bostrom (2001) defines existential risks as "one where an adverse outcome would either annihilate Earth-originating intelligent life or permanently and drastically curtail its potential." We define existing risk as "one whose instances have been observed in the real world and caused harm to individuals, communities, organizations, or the environment."

Although some investment in assessing the existential risks of AI is essential for long-term planning (Hendrycks, Mazeika, and Woodside 2023; Karger et al. 2023; Kasirzadeh 2025), effective risk mitigation today requires understanding how the harms of Generative AI currently emerge. In this paper, we focus on characterizing the existing risks of Generative AI. We argue that addressing these risks is critical not only because they have already caused harm and are likely to persist, but also because they can accumulate over time into more severe harms.

The AI Risk Management Framework (AI RMF) developed by the U.S. National Institute of Standards and Technology (NIST) emphasizes risk *identification* and *prioritization* as critical steps in addressing AI risks (NIST 2023). A growing body of literature identifies the risks and harms of AI across domains and applications (Section 2.1). However, much of this work lacks grounding in real-world evidence and does not address *who* is affected or *how* the risk or harm arises, limiting its effectiveness in guiding risk prioritization and mitigation. In this work, we aim to fill this gap by providing a large-scale, empirical mapping of real-world incidents of harm associated with Generative AI systems, tracing each harm from its underlying sociotechnical *failure(s)* to its downstream impact on people and society.

The AI RMF recommends prioritizing risks based on their *projected impact* (NIST 2023). While assessing projected impact is complex, two key factors should guide this process: (1) the *prevalence* of the risk, or how frequently it causes harm in practice; and (2) *who* is at risk. This motivates our first two research questions. **RQ1:** What is the nature and prevalence of Generative AI harms observed in practice today? **RQ2:** Who is most at risk of these harms? To effectively inform risk mitigation strategies, it is imperative to identify the underlying issue(s) – which we refer to as *sociotechnical failure modes* (Section 2.2) – associated with the incident of harm. While prior work has proposed frameworks to identify where technology fails (Raji et al. 2022; Macrae 2022; Slattery et al. 2025; Carlson 2012; Leveson 2012), much of this work is focused on technical failures, is not grounded in real-world evidence, and does not consider how *Generative* AI may induce a unique landscape of failures and risks compared to other AI technologies. Identifying Generative AI failure modes and mapping them to real-world harms can guide more effective mitigation strate-

*These authors contributed equally.

gies, motivating our third research question. **RQ3:** What is the nature and prevalence of *sociotechnical failure modes* associated with observed harms of Generative AI?

To ground our investigation of these questions in real-world evidence, we follow recent work that uses repositories of publicly reported AI incidents as a critical source of empirical data for mapping and measuring AI harms (Lee et al. 2024; Abercrombie et al. 2024b; Pittaras and McGregor 2023; Hoffman and Frase 2023; Raji et al. 2022; Hutiri, Papakyriakopoulos, and Xiang 2024). We performed a systematic analysis of 499 publicly reported Generative AI incidents, identifying *what* harms occurred, *how* they materialized, and *who* experienced them. Through our analysis, we validated and revised prior taxonomies of AI harms (Figure 1) and developed a novel taxonomy of sociotechnical failure modes (Figure 3), which we present alongside a quantitative analysis of how they relate to each other. We find that most incidents in our analysis bring harm to stakeholders that did *not* interact directly with the Generative AI system, undermining the accountability frameworks assumed in many AI governance models. Additionally, in contrast to previous work that analyzes AI incidents more broadly (Velázquez et al. 2024), we find malicious use to be a prevalent failure mode associated with Generative AI harms.

Our dataset is likely not representative of real-world Generative AI harms in terms of scope or frequency. In particular, databases of publicly reported incidents may disguise the actual prevalence of harms due to factors such as reporting bias. Nonetheless, our study provides essential empirical evidence to ground discussions of AI risk management, helping researchers and policymakers prioritize existing risks, understand where harm mitigation efforts are most urgently needed, and gain insights into where *Generative* AI requires novel risk mitigation strategies (Section 5).

2 Related Work

We define harm as "an outcome negatively affecting people and their interests" and risk as "the magnitude and likelihood of harm" (NIST 2023). We view Generative AI risks and harms in the sociotechnical contexts where they arise; that is, we consider not only technical but also human and societal factors to play critical roles in determining the outcomes of Generative AI (Weidinger et al. 2023; Shelby et al. 2023).

2.1 Taxonomies of AI Risks and Harms

Risk and harm taxonomies provide a common vocabulary through which stakeholders can report, evaluate, and mitigate system failures (Koessler and Schuett 2023). There is some consensus on the broad domains of risk associated with AI. These include risks to the physical or psychological well-being of people, human rights and civil liberties, political and economic structures, society and culture, and the environment (Abercrombie et al. 2024b). Some taxonomies also consider multiple dimensions of harm, such as the level of severity, the scope of the impacted entities, and whether the harm is reversible (OECD 2023). For example, Hoffman and Frase (2023) proposed a two-dimensional taxonomy of AI harms with domain on one axis and level of realization on the other. Using this taxonomy, harm is described by both a domain category such as *Physical Health/Safety* or *Privacy* and whether the harm was realized or yet unrealized.[1]

Throughout this paper, we build on Weidinger et al.'s highly-cited risk taxonomy for Generative AI systems, developed through a literature review, and Abercrombie et al.'s harm taxonomy for AI and automated systems, developed through a combination of use case analysis, literature review, and annotation testing on publicly reported incidents (Weidinger et al. 2023; Abercrombie et al. 2024b). Our contribution goes beyond existing risk and harm taxonomies in two ways: first, we report the quantitative prevalence of different Generative AI harms based on real-world incidents, and second, we map harms to the stakeholders they affect and the underlying issues that precipitated them.

2.2 Sociotechnical Failure Modes

The harms of Generative AI do not materialize in a vacuum; they arise due to a variety of human and technical factors and their interactions (Macrae 2022). To describe these underlying dynamics of harm, we introduce the term *sociotechnical failure modes*, which we define as "the ways in which a technical system, its creators or users, or the various societal structures interacting with it bring risk or harm to some stakeholders." Most AI risk and harm taxonomies do not address the topic of *how* risks and harms surface (Turri and Dzombak 2023). As steps toward filling this gap, Raji et al. (2022) introduce a taxonomy of AI system failures and Slattery et al. (2025) organize AI risks in a *causal* taxonomy.

The term "failure mode" is used in safety engineering contexts such as Failure Mode and Effects Analysis (FMEA), a framework for identifying and prioritizing failure modes by (1) listing out "the functions of a component/system or steps of a process," (2) identifying the failure modes—the "mechanisms by which each function or step can go wrong," and (3) identifying the *impact* and *cause* of each failure mode (Carlson 2012). Recent work proposes the application of safety engineering approaches including FMEA to assess and mitigate the social and ethical risks of machine learning systems (Rismani et al. 2023). However, our method of identifying failure modes using incident reports is distinct from FMEA in two ways: first, it is performed after the failure has occurred, making it more similar to a post-mortem risk assessment; second, because we are limited to publicly available data, we cannot examine the specific components, functions, or control actions of the system.

Another safety engineering framework that may be applicable to machine learning systems is System Theoretic Process Analysis (STPA), which involves mapping the components of a system and their interactions to identify potential sources of harm (Leveson 2012). As Rismani et al. (2023) note, STPA is better suited for capturing emergent phenomena rather than individual component behavior, congruent with increasing work toward thinking about AI failures in sociotechnical contexts (Tang et al. 2024; Weidinger et al. 2023; Shelby et al. 2023). While our analysis in this work

[1]Table 2 in the Appendices at https://arxiv.org/abs/2505.22073 provides a nonexhaustive overview of these taxonomies.

is less structured than what would constitute an application of STPA, we draw inspiration from the first two steps of the framework. First, we "[identify] losses via outlining stakeholders and their values" by identifying harms via first describing *who* is affected in each incident. Second, we "model the *control structure* of the full sociotechnical system" by situating failures within the three-layered framework introduced in Weidinger et al. (2023) to describe the sociotechnical system of Generative AI composed of *Capability*, *Human interaction*, and *Systemic impact* layers.

2.3 AI Incidents as Empirical Data

Prior work on identifying AI harms and failures highlights repositories of publicly reported AI incidents as sources of empirical data (Lee et al. 2024; Abercrombie et al. 2024b; Pittaras and McGregor 2023; Hoffman and Frase 2023; Raji et al. 2022; Hutiri, Papakyriakopoulos, and Xiang 2024; Velázquez et al. 2024). For example, Lee et al. (2024) systematically reviewed 321 publicly reported incidents to develop a taxonomy of AI privacy risks. Similarly, Raji et al. (2022) reviewed 283 publicly reported incidents to develop their taxonomy of AI system failures.

Three commonly cited repositories are the AI Incident Database (AIID); the OECD AI Incidents Monitor; and the AI, Algorithmic and Automation Incidents and Controversies (AIAAIC) repository. All of these repositories maintain harm taxonomies for incident annotation and categorization (McGregor 2021; OECD 2025; Abercrombie et al. 2024a). Each repository also provides their own definition of an *incident*; for example, the AIAAIC's definition is "a sudden known or unknown event (or 'trigger') that becomes public and which takes the form of a disruption, loss, emergency, or crisis" (Abercrombie et al. 2024a).

Although the reported-incident analysis approach has several key limitations due to factors including reporting bias (further discussed in 3.3), the assumption underpinning this approach is that these incidents reflect real-world harms arising in sociotechnical systems and hence serve as reasonable raw material on which to test and apply risk mapping frameworks. Incident repositories also play an important role in reducing risk in other safety-critical domains such as aviation and cybersecurity (Turri and Dzombak 2023).

2.4 How, What, Who: Mapping Failures to Harms to Stakeholders

Slattery et al. (2025) introduced a causal taxonomy of AI risks where each risk is mapped to the *entity* (human, AI, or other) that precipitated it, that entity's *intent*, and *when* (predeployment, post-deployment, or other) the risk arose. Their work contributes to an understanding of the sociotechnical factors that precipitate AI risks and harms, which may help develop novel harm mitigation strategies and identify the stakeholders best positioned to address specific risks. However, in addition to identifying only coarse-grained causal factors, their work does not identify *who* tends to experience each type of risk or harm, limiting its utility to prioritize risks and map accountability pathways.

To address this gap, Velázquez et al. (2024) applied a "what-where-who" framework to analyze 639 AI incident reports. They used large language models to operationalize their framework and automatically classify each incident based on *what* type of harm occurred, *where* it originated, and *who* was harmed. They found that the vast majority of incidents in their dataset harmed the individual directly interacting with the AI system. Hutiri, Papakyriakopoulos, and Xiang (2024) arrived at a similar conceptual framework to characterize the harms of speech generators, mapping harms from their *responsible* to *affected entities*. However, they found that most harms did *not* affect the individual who interacted with the speech generator. Taken together, these papers highlight the importance of mapping harms to the stakeholders they affect. They also suggest that there are critical differences between traditional AI harms and Generative AI harms in terms of who they affect and how they materialize.

In this work, we build on these insights to characterize the how, what, and who of Generative AI harms through a systematic manual analysis of 499 publicly reported incidents. We report the quantitative prevalence of the harms we identify (the *what*), the failure modes that precipitate them (the *how*), and the stakeholders who were harmed (the *who*). By grounding our analysis in real-world incidents, our work offers a novel, empirical understanding of the sociotechnical factors that contribute to Generative AI harms. Our results show how the landscape of Generative AI harms is distinct from traditional AI harms, reveal where pathways of accountability break down, and emphasize the potential impact of non-technical harm mitigation.

3 Methods

3.1 Using Case Studies to Construct Generative AI Risk and Failure Mode Taxonomies

To develop taxonomies of Generative AI harms and sociotechnical failure modes exhibited in publicly reported incidents, we performed a systematic review of Generative AI incident reports. We framed each incident as a case study, where the goal was to determine (1) the output *modality* of the Generative AI system at fault, (2) the *human stakeholders* who experienced harm, (3) the *nature* of the harm experienced, and (4) any and all *sociotechnical failure modes* that contributed to the harm. Table 1 provides further details for each of these four attributes. The third and fourth attributes are the building blocks for our taxonomies of harms and sociotechnical failure modes, respectively.

We sourced incident reports from both the AIID (McGregor 2021) and the AIAAIC (Abercrombie et al. 2024a). Both sources include incidents that involve a wide range of AI systems, so we manually filtered the union of the two databases to include only incidents where a Generative AI was at fault, resulting in our dataset of 499 Generative AI incidents. We did not use the OECD AI Incidents Monitor (OECD 2025) because it is automatically curated (whereas the incidents in the AIID and AIAAIC undergo manual review) and its definition of "incident" is broader than those used in the other two repositories. We scoped our study around incidents—where "incident" is defined as a single event—that have been manually submitted and reviewed.

We based our analysis of both harms and sociotechni-

Column	Description	Example entries
Modality	The output modality of the Generative AI system.	text, image, audio, video, multimodal
Affected stakeholders (up to 2)	The human stakeholders that suffered actual or worst-case scenario harm as a result of the Generative AI system.	end user, individual(s) beyond the end user, community, society
Harm (up to 2)	The actual or worse-case scenario harm experienced by the affected stakeholders.	Impersonation/identity theft, Pollution of information ecosystem
Sociotechnical failure mode (all that apply)	The technical or user issue that most proximally contributed to the harm.	Hallucination, Malicious use, Failure of safety guardrails

Table 1: Summary of Our Coding Schema

cal failure modes on previous literature and used an iterative coding process to categorize each of the four attributes described in Table 1. As a preliminary codebook for our analysis of Generative AI harms, we constructed a union of previous harm taxonomies developed by Abercrombie et al. (2024b) and Weidinger et al. (2023). As a preliminary codebook for Generative AI sociotechnical failure modes, we used Raji et al.'s taxonomy of AI system failures as a baseline and used Slattery et al.'s causal taxonomy of AI risks as a framework within which to brainstorm additional possible failure modes (Raji et al. 2022; Slattery et al. 2025). We then expanded and refined each of these codebooks through a systematic review of our incident dataset. Our codebooks are available at the link above Section 1.

Significant revisions to the codebook of harms. Weidinger et al. (2023) and Abercrombie et al. (2024b) consider privacy violations to be a specific harm rather than a broad category. Therefore, we introduced a *Privacy & Security* category and populated it using the taxonomy of AI privacy risks introduced in Lee et al. (2024). We also surfaced one harm, *Overburdening ecosystems*, not captured in our review of prior literature. Our final taxonomy of Generative AI harms has 41 harms organized into 12 categories (Figure 1).

Significant revisions to the codebook of sociotechnical failure modes. We first revised each AI system failure described in Raji et al. (2022) to be more specific to Generative AI. For example, we increased the granularity of Implementation Failures by deriving three failure modes from its descriptions of implementation issues: *Hallucination*, *Failure of commonsense reasoning*, and *Failure to comply with contextual norms*. We also noted that Raji et al.'s work is specifically concerned with engineering failures, but there were a large number of incidents with use-related failure modes in our dataset. Thus, we generated additional failure modes based on themes that arose in the data and informed by prior literature (Walkowiak and Potts 2024). Our final taxonomy of Generative AI failure modes has 14 failure modes organized into 4 categories (Figure 3).

3.2 Qualitative Analysis Procedure

The first and second authors collaborated to code the incident database with the four attributes summarized in Table 1. After developing the preliminary codebooks, the two authors convened for a training phase on a random 10% of the data. They then independently coded the remainder of the dataset, reconvening periodically to compare codes, discuss, and reach consensus for all disagreements.

The analysis was framed by three questions inspired by Slattery et al. (2025) and Hutiri, Papakyriakopoulos, and Xiang (2024): (1) Which human entities are mentioned in the incident summary? (2) Which entities were harmed and what is the nature of the harm? (3) Was the most proximal source of risk or harm primarily human or technical in nature? Was the harm intentional or unintentional? Did the risk materialize pre- or post-deployment?

In cases where there was insufficient information to determine whether actual harm occurred, the authors imagined the "worst-case scenario." For example, a chatbot convinced its user to euthanize their dog (Tangermann 2024). In the worst-case scenario, this choice was incorrect and the user suffered significant harm. The authors discussed and coded this incident as if this worst-case scenario were true. Additionally, the coders chose the two most *severe* harms in cases where there were more than two types of harm, where severity was measured by the scale of the harmed stakeholder and the magnitude of possible harm. The complete coded incident dataset is linked above Section 1.

3.3 Limitations

As with other work that relies primarily on databases of publicly reported incidents for empirical data, this paper has methodological limitations.

Some harms can have a cumulative nature. Both databases from which we sourced incident reports define an *incident* as a single *event* (the AIAAIC further specifies it to be "sudden") (McGregor 2021; Abercrombie et al. 2024a). This means that harms that do not have single precipitating events, but instead accumulate over time, are unlikely to surface in these databases. For example, many harms to society (e.g., the devaluation and loss of human creativity) or the environment can only be measured over the course of months or years. While it is possible that certain incidents are indicative of cumulative harms—for example, Generative AI systems trained on copyrighted art may produce copyright violations today but lead to the devaluation of human artists over time—we intentionally avoided this kind of speculation in our coding process as the landscape of Generative AI changes rapidly. Instead, we tried to be comprehensive in

our consideration of shorter-term harms.

Harms may be over- or under-represented due to reporting bias. Any database of publicly reported incidents will reflect some level of reporting bias. Concepts that are sensational or relatively well-understood by the public, such as deepfake impersonations, may be overreported. Concepts related to the design or development of Generative AI, such as data acquisition and cleaning or the implementation of safety guardrails, may be underreported, perhaps in part because industry is not incentivized or required to publicly report issues they surface during the design, development, and testing of their products. Additionally, the unsanctioned use of Generative AI in academic or workplace settings may be intentionally concealed and therefore infrequently reported despite high incidence in the real world.

The observational nature of incident reports makes establishing causation difficult. In general, we cannot infer causal relationships between failures and harms based on the information available for incidents in our dataset. Thus, our analysis is strictly correlative, but can help assess where more rigorous analyses (such as FMEA) should be prioritized.

Still, incident repositories serve as important starting points from which to map and estimate the frequency of materialized harms. Future partnerships with institutions such as MITRE, which is developing a confidential incident repository, may improve representativeness.

4 Findings

We report the Generative AI harms (Figure 1) and sociotechnical failure modes (Figure 3) surfaced in our incident dataset, their prevalence, who they affect, and how they relate to each other.[2] We also investigate whether particular output modalities are more often associated with certain types of harms and failure modes. While most of the harms we identified are well-defined and taxonomized in prior work (Weidinger et al. 2023; Abercrombie et al. 2024b; Lee et al. 2024), we introduce our taxonomy of sociotechnical failure modes, as well as our investigation of the relationships between failure modes, harms, and who they affect, as novel contributions. When reporting prevalence data, we report the *number of occurrences* of each harm or failure mode in our incident dataset. Because incidents were coded with up to two harms and as many failure modes as applicable, totals will be greater than the number of incidents.

4.1 What Generative AI Harms Surface in Publicly Reported Incidents?

Our taxonomy of Generative AI harms surfaced in our dataset is made up of 41 harms, each falling into exactly one high-level category (Figure 1). These harms and categories are largely identified in previous literature (Abercrombie et al. 2024b; Weidinger et al. 2023; Lee et al. 2024).

We surfaced one new harm, *Overburdening ecosystems*, and define it as "the pollution of a space or ecosystem intended for human productivity or creativity (e.g., creative material submission or job application portals) with Generative AI." This harm is closely associated with risks to human creativity and critical thinking described in prior literature (Abercrombie et al. 2024b). However, previous literature focuses on the cumulative societal implications of automating human labor and creativity, while Overburdening ecosystems captures the short-term impacts of synthetic content—often amounting to AI *slop*—on those tasked with distinguishing AI from human. For example, several magazines have reported an influx of "low-standard AI-generated" submissions, overwhelming their editors and leading them to close their online submission portals (McMillan 2023).

4.2 Who Experiences the Harms?

To categorize *who* experiences the harms of Generative AI, we define two types of stakeholders: *interacting* stakeholders, who directly and intentionally interact with the Generative AI system, and *non-interacting* stakeholders, who do *not* directly or intentionally interact with the Generative AI system. We further specify three interacting stakeholders: (1) *end user (individual)*, referring to an individual end user of a Generative AI system who is acting on their own interests; (2) *end user (organization)*, referring to an end user of a Generative AI system who is acting as or on behalf of an organization (e.g., a business); (3) *developer/deployer*, referring to the creators of a Generative AI system, including those developing foundation models and those carrying out downstream customizations; and four non-interacting stakeholders: (1) *individual(s) beyond the end user*, referring to individuals who are passive or third parties in the incident (e.g., the unknowing subject of a deepfake); (2) *organization(s) beyond the end user*, referring to organizations that are passive or third parties in the incident (e.g., a magazine unknowingly receiving AI-generated submissions); (3) *community*, referring to a group of people with a common interest, culture, or within a small geographic region (e.g., visual artists); and (4) *society*, referring to a large or disparate aggregate of people and their institutions.

Figure 2 illustrates the distribution of who was affected and highlights the category of harm that most frequently affected each type of stakeholder. Most publicly reported incidents bring harm to non-interacting stakeholders. For example, Autonomy harms, the most prevalent category of harm in our dataset, accrued to individuals or organizations beyond the end user of the Generative AI system in all but one case. Similarly, Political & Economic harms affected society in all but one case. The only category of risk that tended to affect interacting stakeholders was Reputational harms.

These results suggest that stakeholders who tend to experience harm (third-party individuals and organizations, communities, and society) are not necessarily the same stakeholders thought to experience the benefits of Generative AI (the end users and creators). This represents a breakdown of one mechanism of accountability, whereby those *choosing* to participate in the benefits of a technology must also assume the associated risks of harm. Our results suggest that those choosing to participate in the benefits of Generative AI can do so while subjecting themselves to relatively little risk, while those who do *not* choose to participate experience the

[2]Harms and failure modes included in our complete codebooks but not applied during the coding process are omitted here.

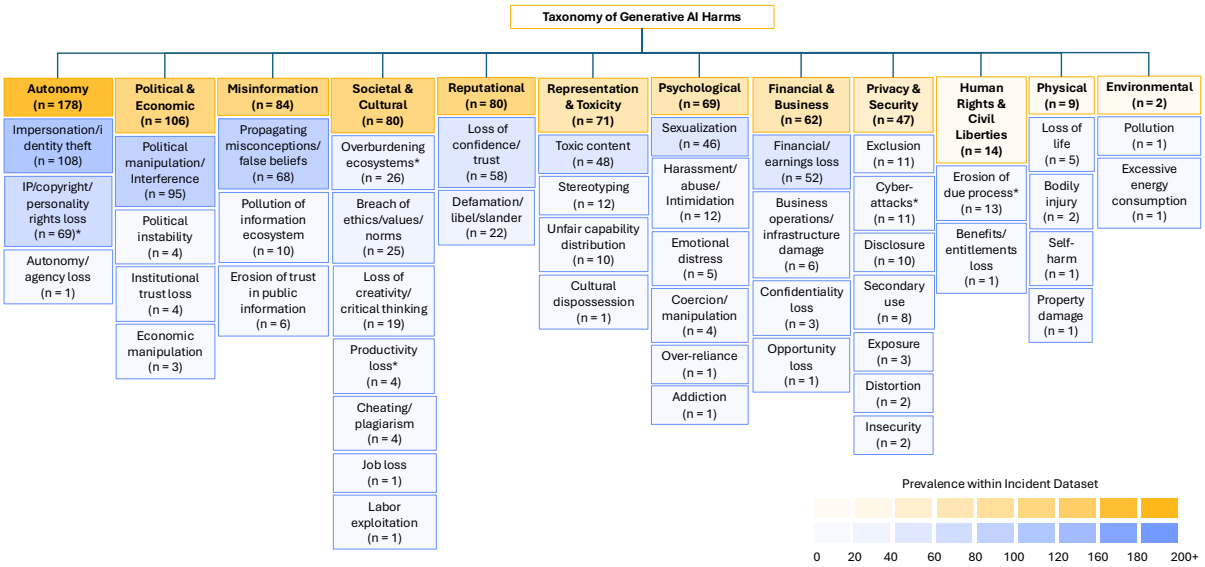

Figure 1: Overview of Taxonomy of Generative AI Harms. Darker colors indicate higher prevalence in our dataset. Harms that were heavily altered or created are indicated by an asterisk.

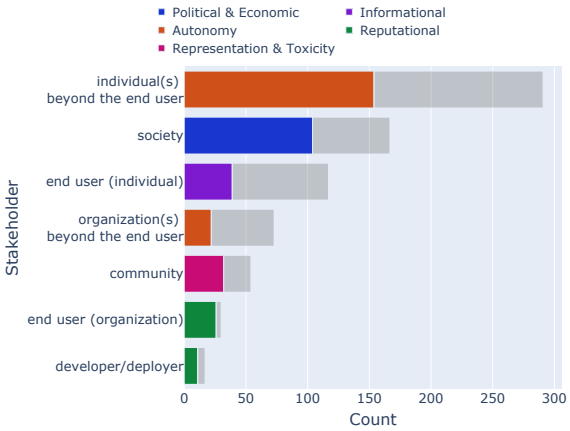

Figure 2: Distribution of harmed stakeholders, with the most prevalent category of harm to each stakeholder represented by the left (colored) side of the respective bar. The right (gray) side of each bar is all other categories of harm. Most incidents harmed non-interacting stakeholders.

vast majority of harm while being deprived of the benefits.

4.3 How do Generative AI Harms Surface in Publicly Reported Incidents?

As argued in Section 2.2, the real-world harms of Generative AI are associated with a variety of sociotechnical failures. Although we acknowledge that we often cannot infer causal relationships between failures and harms based on the information available for incidents in our dataset, our analysis offers a high-level idea of what failure modes exist and are prevalent, and what kinds of harm they tend to precipitate.

We organize the 14 Generative AI sociotechnical failure modes we identified in our dataset based on whether they arise due to the *design*, *development and evaluation*, *release*, or *use* of Generative AI. Certain failure modes may arise due to more than one of these stages, but we define each failure mode narrowly and classify it as accurately as possible. Furthermore, we situate our taxonomy of failure modes within Weidinger et al.'s sociotechnical framework for Generative AI safety evaluation, composed of Capability, Human interaction, and Systemic impact layers (Weidinger et al. 2023).

Design-related failure modes ($n = 112$). These are failure modes arising due to the *design* of a Generative AI system and cannot be addressed with safety features or improved model implementations. These exist largely at the Systemic impact layer, as they have to do with the broader systems into which the Generative AI system is integrated and are present regardless of the system's capabilities or human-AI interactions. We identify three design-related failure modes: (1) *Problematic data collection and use* ($n = 86$), referring to problems arising due to the indiscriminate nature of a system developers' collection and use of data, including preventing subjects (e.g., internet users) from giving their informed consent on how their data is used, (2) *Poor quality training data* ($n = 25$), referring to when the training data for a system reflects or exacerbates societal biases, overrepresents a certain type of content, or is polluted with poor quality information which is not sufficiently cleaned, and (3) *Resource demands* ($n = 3$), referring to the high cost of training and using Generative AI for inference in terms of both physical and human resources.

Development and evaluation-related failure modes ($n = 118$). These are failure modes arising due to the *development and evaluation* of a Generative AI system and exist largely at the Capability layer: that is, they determine whether the Generative AI system and its technical compo-

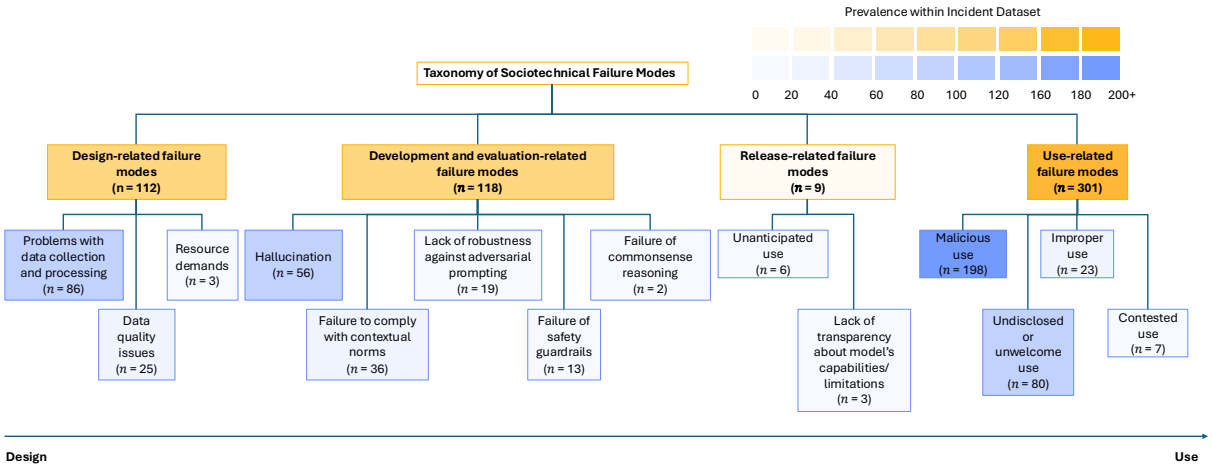

Figure 3: Overview of Taxonomy of Sociotechnical Failure Modes. Darker colors indicate higher overall prevalence in our dataset. Use-related failure modes are the most prevalent of the four categories.

nents are likely to exhibit problematic behaviors. We identify five development and evaluation-failure modes: (1) *Hallucination* ($n = 56$), referring to when a system makes false, misleading, or inaccurate claims as if they were facts, (2) *Failure of commonsense reasoning* ($n = 2$), referring to a system producing erroneous outputs resulting from a lack of intrinsic logic or commonsense reasoning, (3) *Failure of safety guardrails* ($n = 13$), referring to when a safety feature either produces a new problem or simply fails, (4) *Failure to comply with contextual norms* ($n = 36$), referring to when a system fails to meet user expectations (e.g., producing gibberish, parroting), and (5) *Lack of robustness against adversarial prompting* ($n = 19$), referring to when a system misbehaves as a result of adversarial prompting.

Release-related failure modes ($n = 9$). These are issues arising due to the *release* of a Generative AI system, which is typically accompanied by documentation or other user-facing communication. These exist largely at the Human interaction layer: that is, they determine whether the Generative AI system "perform[s] its intended function at the point of use." We identify two release-related failure modes: (1) *Lack of transparency about model's capabilities/limitations* ($n = 3$), referring to when a system's developer fails to provide clarity about the system's robustness, accuracy, or other performance metric in a manner that is comprehensible to its users, and (2) *Unanticipated use* ($n = 6$), referring to when a system's developer fails to anticipate a plausible use case of the system.

Use-related failure modes ($n = 301$). These are issues arising due to the use of the Generative AI system as a result of either user error or user intention. These exist largely at the Human interaction layer. We identify four use-related failure modes: (1) *Undisclosed or unwelcome use* ($n = 80$), referring to when the use of Generative AI in a particular context subverts an explicit or implicit expectation of human expertise, labor, or creativity, (2) *Contested use* ($n = 7$), referring to when it is unclear whether a Generative AI system was used in a certain context (e.g., resulting in false accusations of academic dishonesty), (3) *Improper use* ($n = 23$), referring to when a Generative AI system is used to complete tasks either requiring professional license/training or subject to industry standards and the user fails to properly review outputs, and (4) *Malicious use* ($n = 198$), referring to when a bad actor uses Generative AI to facilitate harm to others including through the spread of disinformation, fraud, defamation, nonconsensual sexualization, or security threats.

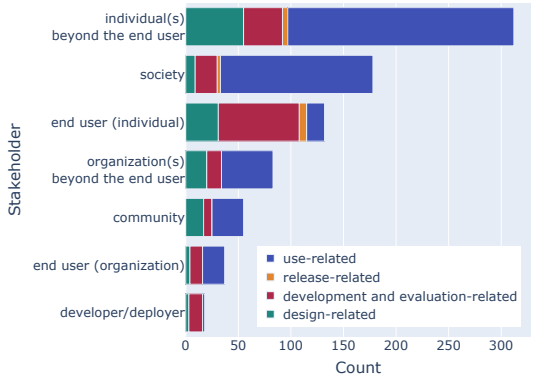

Figure 4: Bars are divided by the type of failure mode associated with harm to the respective stakeholder. Use-related failure modes were associated with the vast majority of harms to non-interacting stakeholders while Development and evaluation-related failure modes were most often associated with harms to interacting stakeholders.

Which failure modes affect each type of stakeholder? In Figure 4, we report the types of failure modes that contribute to the harms experienced by each type of stakeholder. For all types of non-interacting stakeholders, Use-related failure modes (specifically, Malicious use or Undisclosed or unwelcome use) most often contributed to the harms they ex-

perienced. Therefore, not only do most publicly reported incidents involve harms to stakeholders who did not choose to interact with Generative AI and rarely participate in the benefits, but the failure modes most often contributing to those harms are due to user intent. In particular, in cases of Malicious use, the apparent benefit that the Generative AI system offers to its user amounts to intentional harm to other parties. The failure mode most frequently associated with harms to interacting stakeholders was Hallucination, a Development and evaluation-related failure mode.

Thus, safety evaluations at the Capability layer mostly address harms to interacting stakeholders while evaluations at the Human interaction layer are necessary (but likely insufficient) to address harms to non-interacting stakeholders.

4.4 Which Failure Modes Are Associated With Each Type of Harm?

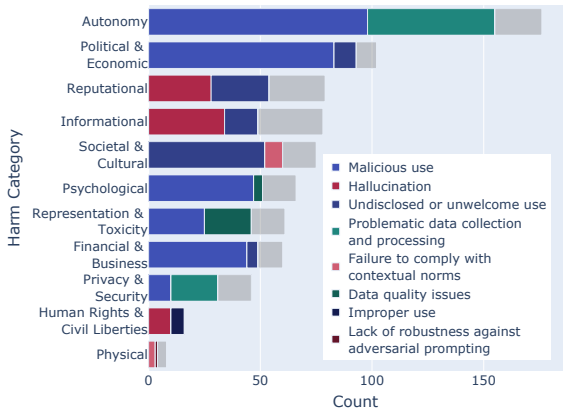

Figure 5: We indicate the two most prevalent sociotechnical failure modes associated with each category of harm with the colored sections of each bar; the gray sections are all other failure modes combined. We omit Environmental harms due to low frequency in our dataset.

As seen in Figure 5, some categories of harm are overwhelmingly associated with a single failure mode. We examined whether the incidents coded with each of these high-frequency combinations (e.g., Autonomy and Malicious use) converged on particular themes.

Autonomy, Political & Economic, Psychological, and Financial & Business harms were all most often associated with Malicious use. The incidents coded with any of these combinations largely involved deepfakes.

- Incidents coded with an Autonomy harm and Malicious use were largely cases of impersonation. For example, several companies were defrauded by deepfake impersonations of their C-suite executives (Stupp 2019; Franceschi-Bicchierai 2020) and both public and private individuals were defamed by their impersonations' offensive remarks (Cole 2019; Gilbert 2023).
- Incidents coded with a Political & Economic harm and Malicious use largely involved deepfakes deployed to create false narratives about political figures (Horton 2020; Liddell 2024), voter groups (Thruston 2024; Linly 2024), or current events (Reuters 2023; Bond 2023).
- Incidents coded with a Psychological harm and Malicious use were largely cases where pornographic material of an individual was created without their knowledge (Stern 2023; Rubin 2023).
- Incidents coded with a Financial & Business harm and Malicious use mostly involved deepfake-facilitated scams (Davies 2024; Al-Siabi 2023), fraud (Lo 2023), or extortion (Goodin 2023; Fischer 2021).

Incidents coded with Human Rights & Civil Liberties harms and Hallucination represented, in all but one case, when false information generated by AI was presented in a court of law (Durkee 2023; Hyde 2023). Incidents coded with Privacy & Security harms and Problematic data collection and use mostly involved violations of privacy rights due to Generative AI developers' data acquisition practices, including internet scraping (Davis 2024; Brittain 2023).

Some combinations were more heterogeneous; for example, the combination of Societal & Cultural harms and Undisclosed or unwelcome use included controversies due to the substitution of AI-generated outputs for human artists (Gault 2022; Codega 2022) and students' use of Generative AI to cheat (Proctor 2024a).

Co-occurring failure modes As Weidinger et al. (2023) argue, the Human interaction and Systemic impact layers of the sociotechnical system are critical in determining whether Generative AI harms materialize in the real world. Throughout our analysis, we found that this was demonstrated by several commonly co-occurring failure modes precipitating a particular type of risk. For example, when AI-generated false information was presented in a court of law (Durkee 2023; Hyde 2023), we identified both Hallucination and Improper use as associated failure modes – while the AI system produced the false information (Hallucination), the user was ultimately held at least somewhat responsible for not having verified the AI's output (Improper use). Both failure modes were necessary to have materialized the harm.

Role of open-source models As noted in prior work, the technical ecosystem that facilitates the creation of AI-generated non-consensual intimate imagery (AIG-NCII) is built upon open-source models (Gibson et al. 2024; Ding and Suresh 2025). Our work both emphasizes the real-world impact of this finding – our dataset included many incidents of non-consensual deepfake pornography – and corroborates it: in cases where the nudification tool was known, we found that it was built upon an open-source model such as Stable Diffusion (Ly 2024; Milmo 2023; Maiberg 2024).

Effect of output modality Among the Generative AI systems associated with the incidents in our dataset, *text* was the most prevalent output modality ($n = 219$), followed by *image* ($n = 93$), *video* ($n = 76$), *audio* ($n = 49$), and *multimodal* ($n = 62$).[3] The prevalence of particular sociotechni-

[3]Throughout the coding process, we considered *code* a separate modality, but ultimately merged it with text due to a low number of incidents ($n = 3$).

cal failure modes and harm categories was highly dependent on the output modality of the Generative AI system at fault. Hallucination was a prevalent failure mode for systems producing text while Malicious use was the most prevalent failure mode for systems producing any other output modality.

5 Discussion

Our findings offer a quantitative overview of the harms and sociotechnical failure modes of Generative AI systems deployed in the real world. Here, we discuss the benefits of eliciting not only the harms, but also *who* they affect and the *sociotechnical failures* that underlie them when conducting case studies of real-world incidents. We conclude with several recommendations for key AI stakeholders.

5.1 Implications of Our Findings

Previous taxonomies of AI and Generative AI harms have identified, categorized, and suggested the implications of a wide range of harms (Abercrombie et al. 2024b; Weidinger et al. 2023; Hutiri, Papakyriakopoulos, and Xiang 2024; Zeng et al. 2024; Shelby et al. 2023; Critch and Russell 2023; Solaiman et al. 2024; Liu et al. 2024; Bird, Ungless, and Kasirzadeh 2023; Barnett 2023; Vidgen et al. 2024; Hoffman and Frase 2023; Pittaras and McGregor 2023; Slattery et al. 2025). Separately, work has been done to identify *failure modes* of AI (Raji et al. 2022; Slattery et al. 2025). However, most previous work lacks a real-world understanding of how these harms and failure modes relate to each other. In this paper, we develop a *relational* understanding of harms, failure modes, and affected stakeholders. This coordinated approach lends itself to mapping pathways of accountability and identifying where they break down, helping inform which stakeholders are best positioned to address each harm or risk.

The uneven distribution of benefits and harms of Generative AI across stakeholders erodes pathways of accountability. In particular, our findings suggest that the balance of benefits and harms of Generative AI differs greatly between its interacting and non-interacting stakeholders. We find that the touted benefits of Generative AI (e.g., increased productivity) accrue mostly to its interacting stakeholders while its non-interacting stakeholders bear many of the most critical and prevalent harms, such as harms to their autonomy, psychological wellbeing, and civil liberties. This has two main implications. First, interacting stakeholders tend not to need to assume substantive risk of harm for their choice to participate in the benefits of Generative AI, eroding one path of accountability. Second, ethical and safety considerations at the Human interaction and Systemic impact layers of Generative AI sociotechnical systems are highly necessary; however, because non-interacting stakeholders affected by failures at these layers often must rely on indirect and inefficient pathways of recourse – for instance, those harmed by the copyright issues resulting from Generative AI developers' data acquisition practices are suing the relevant developers (Bruell 2023; Wininger 2024; Picchi 2023) – developers are rarely incentivized to conduct them. Thus, ethics and safety improvements at these layers likely must be either incentivized or enforced by external oversight bodies. We see this in other domains where this lopsided distribution of benefits and risks is present, such as data privacy, where those who collect data benefit while assuming little risk, and those whose data are collected rarely benefit but assume substantive risk.

Certain cumulative harms, which, as noted in Section 3.3, are unlikely to surface in analyses of publicly reported incidents may accrue largely to the end users of Generative AI. This may change the balance of harms and benefits that end users assume over time. These harms include the deterioration of critical thinking skills due to over-reliance on Generative AI systems, psychological addiction, and psychological isolation (Abercrombie et al. 2024b).

Generative AI changes the landscape of AI harms. As discussed in Section 2.4, Velázquez et al. (2024) applied a framework similar to ours in their analysis of 639 AI incident reports. While their analysis includes incidents involving a wide range of AI systems, we focused our analysis strictly on incidents involving *Generative* AI. Their work shares some key findings with this paper: for example, they found that Autonomy harms were the most prevalent type of harm in their dataset, which we also found (Figure 1). However, they also came to several conclusions that contrast with our findings. For example, they found that the vast majority of incidents in their dataset harmed the individual directly interacting with the AI system, while we found the opposite: the overwhelming majority of incidents in our dataset harmed stakeholders that did *not* directly interact with the Generative AI system (Figure 2). They also found that most harms were non-malicious while Malicious use was the single most prevalent failure mode in our analysis (Figure 3). We argue that these contrasting findings are due to differences between how *Generative* AI and traditional AI harms manifest in the real world; as noted in Section 2.4, this argument is corroborated by the findings of Hutiri, Papakyriakopoulos, and Xiang (2024).

These differences, in concert with our finding that Generative AI systems with different output modalities tend to produce different types of harm via different failure modes (Section 4.4), suggest that effective AI governance is necessarily application-specific.

5.2 Recommendations

As argued in Section 5.1, the distribution of benefits and harms of Generative AI across stakeholders creates an incentive structure that leads to the neglect of non-interacting stakeholders. Therefore, we call on academic, industry, legal, and policy actors to prioritize and advocate for non-interacting stakeholders' interests. We outline three recommendations that aim to address some of the most prevalent and/or severe harms in our analysis, including Autonomy, Psychological, and Human Rights & Civil Liberties harms.

Basic AI literacy can reduce certain risks of harm. Harm is often associated with not just one, but two or more failures in combination (Section 4.4). This suggests that preventing just one of the precipitating failures may have prevented these harms altogether. Capitalizing on this insight,

we advocate for work toward promoting basic AI literacy in the general public, which may help minimize Use-related failures. In particular, several incidents associated with Improper use demonstrate that members of the general public do not understand the capabilities, limitations, or risks of Generative AI. For example, when a Canadian lawyer cited hallucinations in a divorce case, she claimed that she did not realize that ChatGPT could hallucinate (Proctor 2024b).

Because Generative AI applications are often marketed as consumer products, the general public is a critical audience and participant in conversations about Generative AI governance. Yet, many members of the public lack the knowledge and vocabulary to describe the behavior of Generative AI applications. We argue that it is critical to communicate current understandings of the capabilities, limitations, and harms of Generative AI in ways that are comprehensible to the general public. Some work is being done on consumer-facing *safety labels* (Chia 2024), which aim to encourage developers and deployers to "be transparent with users by providing information on how the generative AI models and apps work." This is an important step, but it is insufficient: many of the harms of Generative AI affect those who do *not* interact directly with it, and hence are unlikely to see these labels. Taxonomies of harms and failure modes are important artifacts for the general public, but are currently largely directed toward expert audiences. Hence, an important area of future research is to create publicly accessible and digestible information about Generative AI failure modes and harms. In the near future, we aim to distill both taxonomies developed in this paper into a non-technical and concise document and assess its usability for members of the public.

The current regulatory landscape enables the development of Generative AI tools specifically for abusive purposes, especially for the creation of AIG-NCII. The EU AI Act incentivizes open-source development by reducing the associated obligations of documentation and disclosure (EU 2024). As Kapoor et al. (2024) caution, however, while open-source models offer benefits such as increasing innovation, they also pose greater risks of misuse. Our findings corroborate this concern (Section 4.4). Although we acknowledge that, when limited to publicly available data, it is difficult to confidently determine whether an incident involved an open or closed model, we conjecture that the majority of egregious applications (for example, websites enabling the creation of AIG-NCII) are built on open-source models. This suggests that more research centered on real-world evidence of the supposed risks and benefits of open-source foundation models is necessary to inform regulatory stances.

We argue that the current regulatory landscape enables the creation of AIG-NCII. While there has been some work to regulate the spread of AIG-NCII, such as the Take It Down Act (Cruz and Klobuchar 2025), these measures still place the burden of removal on the victims (Ding and Suresh 2025). Therefore, minimizing the *creation* of AIG-NCII is critical to minimize its psychological harms. Gibson et al. (2024) surface several viable solutions, including, as we have discussed, critical analysis of the benefits of open-source models, as well as consideration of access restrictions and tracking of downstream usage. They also note that in the absence of any regulation requiring verification of deepfake subjects' consent or age, many nudification tools do not even mention these issues. Overall, the current lack of regulation of the open-source technical ecosystem enables the existence of tools specifically positioned to be abusive.

Provenance data tracking and synthetic content detection can reduce some risks of harm. The NIST AI 100-4 (NIST 2024) proposes provenance data tracking and synthetic content detection as strategies to reduce risks posed by synthetic content, including AIG-NCII and misinformation. However, some researchers find that these solutions are incomplete or misaligned. Ding and Suresh (2025) argue that just because AIG-NCII is detectable does not mean it is harmless, and Kapoor and Narayanan (2025) argue that the real bottleneck for misinformation is *distribution* rather than creation – thus, fears surrounding the proliferation of AI-generated misinformation were overblown.

Nonetheless, we argue that data tracking and synthetic content detection can reduce some risks from synthetic content. In particular, researchers have cautioned that Generative AI could still meaningfully change the landscape of online misinformation, although in more incremental ways than originally anticipated (Sanderson, Messing, and Tucker 2024). We corroborate this argument: we found several incidents where Generative AI was deployed to generate convincing misinformation, causing significant harm. For example, an AI-generated image of an explosion at the Pentagon sparked a dip in the US stock market (Bond 2023). In addition, we surfaced many incidents of deepfake-facilitated scams (Davies 2024; Al-Siabi 2023), fraud (Lo 2023), and extortion (Goodin 2023; Fischer 2021). These incidents have caused significant financial and psychological harm to their victims, and we argue that they can be effectively mitigated by data tracking and synthetic content detection.

6 Conclusion

In this paper, we set out to develop an understanding of real-world Generative AI harms, the sociotechnical failures that precipitate them, and who they affect. We ground our work in a systematic review of 499 publicly reported Generative AI incidents. We find that many incidents result in harm to those who do *not* interact directly with Generative AI (such as the unknowing subject of a deepfake), but are often associated with failures due to those who *do* interact directly with Generative AI (the end users and developers). Our findings also suggest that the landscape of Generative AI harms is meaningfully different from that of traditional AI systems: for example, malicious use may be a far more prevalent failure mode for Generative AI systems than for traditional AI systems. Given these insights, we make several recommendations for key AI stakeholder groups to address Generative AI harms, including through public disclosures and education, regulatory stances on open-source development, and provenance data tracking and synthetic content detection.

Acknowledgments

This work was supported in part by Meta, a Carnegie Mellon University Rales Fellowship, the Block Center for Technology and Society at Carnegie Mellon University, and the CMU-NIST Cooperative Research Center on AI Measurement Science & Engineering (AIMSEC). Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not reflect the views of funding entities.

References

Abercrombie, G.; Benbouzid, D.; Giudici, P.; Golpayegani, D.; Hernandez, J.; Noro, P.; Pandit, H.; Paraschou, E.; Pownall, C.; Prajapati; et al. 2024a. AIAAIC - AIAAIC Repository.

Abercrombie, G.; Benbouzid, D.; Giudici, P.; Golpayegani, D.; Hernandez, J.; Noro, P.; Pandit, H.; Paraschou, E.; Pownall, C.; Prajapati, J.; Sayre, M. A.; Sengupta, U.; Suriyawongkul, A.; Thelot, R.; Vei, S.; and Waltersdorfer, L. 2024b. A Collaborative, Human-Centred Taxonomy of AI, Algorithmic, and Automation Harms. arXiv:2407.01294.

Al-Siabi, N. 2023. Someone Deepfaked Joe Rogan to Sell a Male Enhancement Product. *Futurism*.

Barnett, J. 2023. The ethical implications of generative audio models: A systematic literature review. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, 146–161.

Bird, C.; Ungless, E.; and Kasirzadeh, A. 2023. Typology of risks of generative text-to-image models. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, 396–410.

Bond, S. 2023. Fake viral images of an explosion at the Pentagon were probably created by AI. *NPR*.

Bostrom, N. 2001. Existential Risks: Analyzing Human Extinction Scenarios and Related Hazards. *Journal of Evolution and Technology*.

Brittain, B. 2023. Google hit with class-action lawsuit over AI data scraping. *Reuters*.

Bruell, A. 2023. New York Times Sues Microsoft and OpenAI, Alleging Copyright Infringement. *Wall Street Journal*.

Carlson, C. 2012. *Effective FMEAs: Achieving Safe, Reliable, and Economical Products and Processes Using Failure Mode and Effects Analysis*. John Wiley & Sons, Inc. ISBN 9781118007433.

Chia, O. 2024. 'Safety labels' that clearly indicate AI risks and testing on the cards in Singapore. *The Straits Times*.

Codega, L. 2022. Tor Tried to Hide AI Art on a Book Cover, and It Is a Mess. *Gizmodo*.

Cole, S. 2019. A Site Faking Jordan Peterson's Voice Shuts Down After Peterson Decries Deepfakes. *Vice*.

Critch, A.; and Russell, S. 2023. TASRA: a Taxonomy and Analysis of Societal-Scale Risks from AI. arXiv:2306.06924.

Cruz, T.; and Klobuchar, A. 2025. S.4569 - 118th Congress (2023-2024): TAKE IT DOWN Act | Congress.gov | Library of Congress.

Davies, M. 2024. Michael Mosley 'deepfake' warning as late doctor's likeness used in sham footage. *Daily Record*.

Davis, W. 2024. OpenAI transcribed over a million hours of YouTube videos to train GPT-4. *The Verge*.

Ding, M. L.; and Suresh, H. 2025. The Malicious Technical Ecosystem: Exposing Limitations in Technical Governance of AI-Generated Non-Consensual Intimate Images of Adults. arXiv:2504.17663.

Durkee, A. 2023. Ex-Trump 'Fixer' Michael Cohen Admits He Accidentally Used Google Bard To Put Fake Cases Into Legal Filing. *Forbes*.

EU, T. 2024. High-level summary of the AI Act | EU Artificial Intelligence Act.

Fischer, D. 2021. Teen charged with extorting official with explicit photos. *AP News*.

Franceschi-Bicchierai, L. 2020. Listen to This Deepfake Audio Impersonating a CEO in Brazen Fraud Attempt. *Vice*.

Gault, M. 2022. An AI-Generated Artwork Won First Place at a State Fair Fine Arts Competition, and Artists Are Pissed. *Vice*.

Gibson, C.; Olszewski, D.; Brigham, N. G.; Crowder, A.; Butler, K. R. B.; Traynor, P.; Redmiles, E. M.; and Kohno, T. 2024. Analyzing the AI Nudification Application Ecosystem. arXiv:2411.09751.

Gilbert, D. 2023. High Schoolers Made a Racist Deepfake of a Principal Threatening Black Students. *Vice*.

Goodin, D. 2023. FBI warns of increasing use of AI-generated deepfakes in sextortion schemes. *ArsTechnica*.

Hendrycks, D.; Mazeika, M.; and Woodside, T. 2023. An Overview of Catastrophic AI Risks. arxiv:2306.12001 [cs].

Hoffman, M.; and Frase, H. 2023. Adding Structure to AI Harm: An Introduction to CSET's AI Harm Framework. *Center for Security and Emerging Technology*.

Horton, J. 2020. US 2020 Election: Does Joe Biden support defunding the police? *BBC*.

Hutiri, W.; Papakyriakopoulos, O.; and Xiang, A. 2024. Not my voice! a taxonomy of ethical and safety harms of speech generators. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, 359–376.

Hyde, J. 2023. LiP presents false citations to court after asking ChatGPT. *The Law Society Gazette*.

Kapoor, S.; and Narayanan, A. 2025. We Looked at 78 Election Deepfakes. Political Misinformation is not an AI Problem. *AI Snake Oil*.

Kapoor, S.; et al. 2024. Position: on the societal impact of open foundation models. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

Karger, E.; Rosenberg, J.; Jacobs, Z.; Hickman, M.; Hadshar, R.; Gamin, K.; and Tetlock, P. 2023. Forecasting Existential Risks: Evidence from a Long-Run Forecasting Tournament. *Forecasting Research Institute*.

Kasirzadeh, A. 2025. Two Types of AI Existential Risk: Decisive and Accumulative. arxiv:2401.07836 [cs].

Koessler, L.; and Schuett, J. 2023. Risk assessment at AGI companies: A review of popular risk assessment techniques from other safety-critical industries. arXiv:2307.08823.

Lee, H.-P.; Yang, Y.-J.; Von Davier, T. S.; Forlizzi, J.; and Das, S. 2024. Deepfakes, phrenology, surveillance, and more! a taxonomy of ai privacy risks. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, 1–19.

Leveson, N. 2012. *Engineering a Safer World: Systems Thinking Applied to Safety*. The MIT Press. ISBN 9780262298247.

Liddell, J. 2024. Trump posts AI-generated image of Harris speaking at DNC with communist flags. *The Independent*.

Linly, Z. 2024. GOP Pollster's Viral AI Fake Photos Of Republican 'Black Voters' Spotlight Election Misinformation Fears. *News One*.

Liu, Y.; Yao, Y.; Ton, J.-F.; Zhang, X.; Guo, R.; Cheng, H.; Klochkov, Y.; Taufiq, M. F.; and Li, H. 2024. Trustworthy LLMs: a Survey and Guideline for Evaluating Large Language Models' Alignment. arXiv:2308.05374.

Lo, C. 2023. Hong Kong police arrest 6 in crackdown on fraud syndicate using AI deepfake technology to apply for loans. *South China Morning Post*.

Ly, J. 2024. Open Foundation Models: Implications of Contemporary Artificial Intelligence. *Center for Security and Emerging Technology*.

Macrae, C. 2022. Learning from the failure of autonomous and intelligent systems: Accidents, safety, and sociotechnical sources of risk. *Risk analysis*, 42(9): 1999–2025.

Maiberg, E. 2024. Samsung-Backed AI Image Generator Produces Nonconsensual Porn. *404 Media*.

McGregor, S. 2021. Preventing repeated real world AI failures by cataloging incidents: The AI incident database. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 15458–15463.

McMillan, R. 2023. AI Junk Is Starting to Pollute the Internet. *Wall Street Journal*.

Milmo, D. 2023. AI-created child sexual abuse images 'threaten to overwhelm internet'. *The Guardian*.

NIST. 2023. Artificial intelligence risk management framework (AI RMF 1.0). *National Institute of Standards and Technology*, 100–1.

NIST. 2024. Reducing risk posed by synthetic content an overview of technical approaches to digital content transparency.

OECD. 2023. Stocktaking for the development of an AI incident definition. *OECD Artificial Intelligence Papers*.

OECD. 2025. AIM: The OECD AI Incidents Monitor, an evidence base for trustworthy AI.

Picchi, A. 2023. George R.R. Martin, John Grisham and other major authors sue OpenAI, alleging "systematic theft". *CBS News*.

Pittaras, N.; and McGregor, S. 2023. A Taxonomic System for Failure Cause Analysis of Open Source AI Incidents. In *Proceedings of the Workshop on Artificial Intelligence Safety 2023*.

Proctor, J. 2024a. Air Canada found liable for chatbot's bad advice on plane tickets. *CBC*.

Proctor, J. 2024b. B.C. lawyer reprimanded for citing fake cases invented by ChatGPT. *CBC*.

Raji, I. D.; Kumar, I. E.; Horowitz, A.; and Selbst, A. 2022. The fallacy of AI functionality. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 959–972.

Reuters. 2023. Kremlin: fake Putin address broadcast on Russian radio stations after 'hack'. *Reuters*.

Rismani, S.; Shelby, R.; Smart, A.; Jatho, E.; Kroll, J.; Moon, A.; and Rostamzadeh, N. 2023. From plane crashes to algorithmic harm: applicability of safety engineering frameworks for responsible ML. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 1–18.

Rubin, A. 2023. Teens exploited by fake nudes illustrate threat of unregulated AI. *Axios*.

Sanderson, Z.; Messing, S.; and Tucker, J. A. 2024. Misunderstood mechanics: How AI, TikTok, and the liar's dividend might affect the 2024 elections. *Brookings*.

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.

Slattery, P.; Saeri, A. K.; Grundy, E. A. C.; Graham, J.; Noetel, M.; Uuk, R.; Dao, J.; Pour, S.; Casper, S.; and Thompson, N. 2025. The AI Risk Repository: A Comprehensive Meta-Review, Database, and Taxonomy of Risks From Artificial Intelligence. arXiv:2408.12622.

Solaiman, I.; et al. 2024. Evaluating the Social Impact of Generative AI Systems in Systems and Society. arXiv:2306.05949.

Stern, M. 2023. A College Girl Found Deepfake Porn of Herself Online. Who Did It Shocked Her. *Rolling Stone*.

Stupp, C. 2019. Fraudsters Used AI to Mimic CEO's Voice in Unusual Cybercrime Case. *Wall Street Journal*.

Tang, N.; Zhi, J.; Kuo, T.-S.; Kainaroi, C.; Northup, J. J.; Holstein, K.; Zhu, H.; Heidari, H.; and Shen, H. 2024. AI Failure Cards: Understanding and Supporting Grassroots Efforts to Mitigate AI Failures in Homeless Services. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, 713–732.

Tangermann, V. 2024. AI CEO Proud of Chatbot for Convincing Woman to Euthanize Her Dog. *Futurism*.

Thruston, J. 2024. AI images of Donald Trump with black voters spread before election. *The Times*.

Turri, V.; and Dzombak, R. 2023. Why we need to know more: Exploring the state of AI incident documentation practices. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, 576–583.

Velázquez, J. D. M.; Šćepanović, S.; Gvirtz, A.; and Quercia, D. 2024. Decoding Real-World Artificial Intelligence Incidents. *Computer*, 57(11): 71–81.

Vidgen, B.; et al. 2024. Introducing v0.5 of the AI Safety Benchmark from MLCommons. arXiv:2404.12241.

Walkowiak, E.; and Potts, J. 2024. Generative AI, Work and Risks in Cultural and Creative Industries. *SSRN Electronic Journal*.

Weidinger, L.; Rauh, M.; Marchal, N.; Manzini, A.; Hendricks, L. A.; Mateos-Garcia, J.; Bergman, S.; Kay, J.; Griffin, C.; Bariach, B.; Gabriel, I.; Rieser, V.; and Isaac, W. 2023. Sociotechnical Safety Evaluation of Generative AI Systems. arXiv:2310.11986.

Wininger, A. 2024. China's Beijing Internet Court Recognizes Personality Rights in Generative AI Case. *The National Law Review*.

Zeng, Y.; Klyman, K.; Zhou, A.; Yang, Y.; Pan, M.; Jia, R.; Song, D.; Liang, P.; and Li, B. 2024. AI Risk Categorization Decoded (AIR 2024): From Government Regulations to Corporate Policies. arXiv:2406.17864.