Issue Brief

The Mechanisms of AI Harm

Lessons Learned from AI Incidents

Author
Mia Hoffmann

CSET CENTER *for* SECURITY *and* EMERGING TECHNOLOGY

October 2025

Executive Summary

With recent advancements in artificial intelligence—particularly, powerful generative models—private and public sector actors have heralded the benefits of incorporating AI more prominently into our daily lives. Frequently cited benefits include increased productivity, efficiency, and personalization. However, the harm caused by AI remains to be more fully understood. As a result of wider AI deployment and use, the number of AI harm incidents has surged in recent years, suggesting that current approaches to harm prevention may be falling short. This report argues that this is due to a limited understanding of how AI risks materialize in practice. Leveraging AI incident reports from the AI Incident Database, it analyzes how AI deployment results in harm and identifies six key mechanisms that describe this process (Table 1).

Table 1: The Six AI Harm Mechanisms

Intentional Harm	Unintentional Harm
• Harm by design • AI misuse • Attacks on AI systems	• AI failures • Failures of human oversight • Integration harm

A review of AI incidents associated with these mechanisms leads to several key takeaways that should inform AI governance approaches in the future.

1. **A one-size-fits-all approach to harm prevention will fall short.** This report illustrates the diverse pathways to AI harm and the wide range of actors involved. Effective mitigation requires an equally diverse response strategy that includes sociotechnical approaches. Adopting model-based approaches alone could especially neglect integration harms and failures of human oversight.

2. **To date, risk of harm correlates only weakly with model capabilities.** This report illustrates many instances of harm that implicate single-purpose AI systems. Yet many policy approaches use broad model capabilities, often proxied by computing power, as a predictor for the propensity to do harm. This fails to mitigate the significant risk associated with the irresponsible design, development, and deployment of less powerful AI systems.

3. **Tracking AI incidents offers invaluable insights into real AI risks and helps build response capacity.** Technical innovation, experimentation with new use cases, and novel attack strategies will result in new AI harm incidents in the

future. Keeping pace with these developments requires rapid adaptation and agile responses. Comprehensive AI incident reporting allows for learning and adaptation at an accelerated pace, enabling improved mitigation strategies and identification of novel AI risks as they emerge. Incident reporting must be recognized as a critical policy tool to address AI risks.

Table of Contents

Introduction

Due to widespread AI use and deployment, AI systems are increasingly implicated in harmful events. Just since the beginning of 2025, 279 new incidents have been added to the AI Incident Database (AIID), a nonprofit effort dedicated to tracking realized harm from AI deployment.[1] Since its launch in November 2020, the database has collected and indexed more than 1,200 incidents of harm or near misses involving algorithmic systems and AI.[*]

Clearly, more efforts are needed to prevent such AI harm. Preemptive harm prevention is the underlying goal pursued by most AI governance interventions, be it regulations like the European Union's AI Act, executive guidance like the Office of Management and Budget's memorandum M-25-21, or company frameworks like Anthropic's Responsible Scaling Policy.[2] Preventing harm effectively, however, requires a better understanding of how AI use leads to harmful outcomes in practice, rather than in theory.

AI incidents data provide valuable insights for understanding how AI systems can cause real harm. By collecting, indexing, and archiving reports from hundreds of real-world AI incidents, the AIID has created a treasure trove of data describing not only the myriads of harms AI systems have been implicated in, but also how these harms came to be. CSET previously leveraged this data to create an analytical framework that provides fundamental definitions and classification schemes for incident data analysis.[3] This framework was then used to annotate and classify more than 200 incidents from the database by incident type, harm category, and other dimensions.[4]

This past analytical work serves as the foundation for this report, which describes the variety of forces involved in AI harm. Leveraging the more than 200 reviewed cases of real-world harm from the database, it identifies six key "mechanisms of harm," which can be divided into intentional and unintentional harm (Table 2).

---

[*] Each incident in the AIID corresponds to one or more instances of harm, so the total number of discrete harm events captured in the database is higher than the number of incident IDs. While some incident IDs correspond to a single documented harm instance, others capture media-constructed accounts that aggregate related incidents into a single narrative.

Table 2: The Six AI Harm Mechanisms

Intentional Harm	Unintentional Harm
• **Harm by design:** Harm caused by AI systems designed and developed for harmful purposes • **AI misuse:** Use of AI systems for harm against the developers' intentions • **Attacks on AI systems:** Harm resulting from AI behavior or (in)action caused by cyberattacks	• **AI failures:** Harm caused by AI errors, malfunctions, or bias • **Failures of human oversight:** Harm resulting from the failure of human-machine-teams • **Integration harm:** Harm resulting as an unintended consequence of deployment in a given context

The six mechanisms comprehensively describe the various pathways to harm found in the AIID. As such, they provide a richer understanding of how AI risks materialize in practice, which can help guide mitigation strategies.

These harm mechanisms may have immediate policy relevance for companies hoping to comply with regulations like the European Union's AI Act. The EU recently released a code of practice for general-purpose AI. This voluntary compliance tool was developed to help providers of general-purpose AI models adhere to the act's requirements, and lays out a comprehensive risk management process. Under it, developers must engage in risk modeling, described as "a structured process aimed at specifying pathways through which a systemic risk stemming from a model might materialize."[5] This framework of harm mechanisms, built on empirical evidence of real-world incidents, may serve as a useful starting point for this exercise.

Importantly, these diverse mechanisms need to be addressed through an equally wide range of mitigation strategies. Security practices may serve to alleviate risks of misuse and attack, but do nothing to address integration harms. Performance standards and testing protocols can reduce AI failures but won't mitigate limitations of human oversight. To prevent AI harm effectively, a diverse toolbox is required.

The following sections present harm incidents from the AIID to illustrate the six harm mechanisms and deepen readers' understanding of the variety of ways in which AI can cause harm. While this report does not intend to provide a comprehensive overview of mitigation techniques, it highlights measures that combat specific mechanisms where possible, and describes where more research is needed to find effective mitigations to particular challenges.

Methodology

Analysis for this report took place in two stages. The first stage involved the identification of the six harm mechanisms based on in-depth study of AI incidents. In previous research, CSET developed a standardized analytical framework for the study of AI harms.[6] The development of this framework involved an iterative process of incident review and framework adaptation through which the key elements required for the identification of real-world AI harm, their basic relational structure, and their definitions were identified. A central conceptual component of the framework is the "chain of harm," i.e., the series of events between AI deployment and the incident outcome that leads to harm. The chain of harm serves an essential function in the identification of AI harm: For there to be AI harm, there has to be a direct link between the behavior of the AI system and the harm that occurred.

The framework was applied to more than 200 incidents from the AIID, which were annotated and classified by incident type, harm category, and other variables.[7] This analytical process and the corresponding in-depth investigation of a large number of incidents showed that the chain of harm was characterized by a variety of forces that shaped how incidents unfolded: the harm mechanisms. While errors—both human and AI—played a role, so did intentionally harmful uses of AI systems, and on occasion the integration of AI into a specific deployment context was harmful on its own. The six harm mechanisms presented above were derived from the analysis of these repeated patterns of events in the chains of harm.

The second stage involved validating the derived mechanisms by categorizing a random set of 200 incidents from the AIID. This was necessary for two reasons. First, the sample of incidents to which the AI Harm Framework had originally been applied had not been randomly selected, and was thus not representative of incidents in the AIID at the time. Secondly, the number of incidents in the AIID had grown by over 50% since the beginning of the first stage in 2023. Validation was therefore essential to ensure the mechanisms remain applicable and relevant to a newer set of incidents that more accurately reflect the current level of technological innovation. The result of this exercise is shown in Figure 1 in the appendix.

Limitations

All frameworks and models are necessarily an abstract and simplified representation of reality. In the real world, harm mechanisms are often not as clear-cut as they appear in this report, and several mechanisms can be active simultaneously. Attacks on AI systems are often carried out to enable misuse. Model performance issues and human oversight failures occur at the same time. And systems designed for harm can fail and cause unintended or excessive harm.

Representing intentionality as a binary is useful to help distinguish the different actors that interventions need to target. However, in reality intentionality is a spectrum. Some incidents occur as truly unintended and unforeseeable consequences of AI use, and others are obviously intentional. But many sit in between, resulting from developers' and deployers' negligence to consider potential impacts, or even from reckless disregard of easily foreseeable harm. This fluidity creates edge cases that render the distinction between harm by design and integration harm difficult. Judging these cases with certainty would require insights into the developers' and deployers' decision-making and governance processes that are generally not available. Thus, unless there is evidence to the contrary, this report relies on the assumption that harm was unintentional when categorizing edge case incidents. Future work may address this limitation through a more disaggregated representation of intentionality.

Lack of information can similarly impede the distinction between harm by design and misuse. Since outside observers generally cannot discern the underlying AI model in a system that is involved in harm, it is often unclear whether an existing AI model was misused or one was designed for this purpose.[8] Distinguishing between the two categories is nonetheless worthwhile, because mitigation requires different interventions based on which actor along the AI value chain intends to do harm; the model developer precipitates harm by design, whereas the AI system's user causes harm from misuse.[*] Even when they share some overlap, separation allows us to identify the appropriate mitigation measures to address each mechanism more effectively.

Finally, there are two limitations of the data source. Although the AIID represents the most comprehensive collection of AI incidents and harms to date, the distribution of incident types in the database does not necessarily reflect their prevalence in the real world. Since the database depends on journalistic reporting, it represents the practices and biases of the media ecosystem. As such, it overrepresents incidents that are attention-grabbing or associated with current societal debates. This suggests that less

[*] Distinguishing developers, deployers, and users is not always straightforward, and sometimes one entity occupies multiple roles. For example, ChatGPT is an AI system that is both developed and deployed by OpenAI. Individuals interacting with the chatbot are the users. In a scenario where OpenAI builds a customer service chatbot on top of their language model GPT-5 for a third party (e.g., an insurance company), the developer is still OpenAI but the deployer is the insurance company, and the company's customers are the chatbot's users.

spectacular mechanisms like integration harm might be underrepresented compared to instances of human misuse, which have driven much of the societal debate recently—particularly where it relates to generative AI systems.

Lastly, there are many harms from AI that cannot easily be captured in an incident database because they do not materialize in discrete instances. The consequences of AI energy consumption, the detrimental impact of AI overreliance on education, or the adverse effects of AI companions on human relationships are just a few examples of possible harms that rarely present as individual incidents. While worthy of analysis, these types of harms are not within the scope of this study.

AI Harm Mechanisms

Intentional Harm

Harm by Design

AI systems designed with the intention of causing harm represent the most straightforward of the six harm mechanisms. In this case, the developer designs the AI system to perform an inherently harmful task or to be used in harmful ways. Developers' and users' intentions to cause harm are generally aligned in incidents of harm by design, though as the following examples illustrate, types of harm by design systems can vary widely.

Some AI systems developed for defense and law enforcement, such as AI-enabled intelligence analysis and battlespace management systems used for targeting or autonomous weapons systems with AI-enabled navigation, computer vision, or terminal guidance, are obvious examples of harm by design systems. No malfunctions or misuse need to occur for harm to materialize when these AI systems are used, since harm is the intended outcome, both by developer and deployer. Militaries may appropriately use these systems against lawful combatants when deployed in accordance with the law of armed conflict. Recent conflicts involving Ukraine and Israel have reportedly seen AI-enabled systems capable of causing harm deployed in combat.[9]

But harm by design is also prevalent outside of law enforcement and defense contexts. Deepfake apps that allow users to maliciously create nonconsensual intimate imagery (NCII) abound. There are dozens of such incidents recorded in the AIID—far too many to describe individually.[10] The harm overwhelmingly affects women and girls, and as such, these incidents provide tangible evidence of a fast-growing form of gender-based digital violence.[11] While this is not a new problem nor even a new AI capability (image generators have been around since approximately 2017), incidents involving pornographic content have surged since AI "nudify" apps and pornographic video generators have become more widely available online.[12]

There are also more subtle forms of intentionally harmful algorithmic design. Online marketplaces such as Naver, Coupang, and Amazon India have been accused of engaging in unfair competitive practices through algorithmic manipulation.[13] The companies allegedly rigged the recommender systems and search algorithms powering their platforms to favor their own products and brands, boosting their market share and causing economic and financial harm to their competitors. Exploiting their dominance as an online market platform to promote their own brand as a vendor violates antitrust laws and, given the platforms' wide online reach and customer base, the overall impact and scale of harm from even minor manipulation can be significant.

Mitigating Harm by Design

In general, the choice of approach to addressing harm by design depends on whether the intended harm is considered socially acceptable or necessary. Prohibiting the development and deployment of AI systems for certain use cases can be an effective measure for use cases causing unacceptable harms.

In contexts where harm by design is deemed acceptable, such as defense and law enforcement functions, the goal of governance is not to prevent harm entirely but to reduce it to what is necessary in a clearly defined and contestable framework. Institutional policy can help ensure the responsible deployment of the technology in order to prevent excessive harm and negligent use or abuse. Generally, organizations should establish AI governance principles that clearly define the circumstances and conditions under which reliance on autonomous or AI-powered decisions and actions is acceptable. A solid accountability framework with clearly assigned roles and responsibilities can ensure that any decision-making that leads to harm is transparent and tractable. Institutional oversight bodies, assuming sufficient independence and transparency, should then be authorized to investigate and audit AI-supported decision-making and actions taken when violations of those policies and frameworks occur or are suspected.

AI Misuse

AI systems that have not been developed with the explicit goal of doing harm can still be misused for that purpose. Compared to the harm by design mechanism, where the intent to harm lies with both the developer and user, in cases of AI misuse the intent to harm lies with the user or operator of the AI system only. Note that AI models can also be misused for non-malicious purposes, such as using AI to do homework.[14] While this may cause users to unintentionally harm themselves—for example by detrimentally affecting their own learning—this section is only concerned with intentionally harmful, malicious misuse.[15]

Both specialized algorithmic systems and general-purpose AI models are prone to malicious misuse, although incidents involving the latter are more common in the AIID.

General-purpose AI, including large language models and text-to-image generators, perform many different tasks well, which makes them particularly easy to misuse for a range of purposes.

For example, in 2023, users of the online forum 4chan created hateful and violent voice impersonations of celebrities using ElevenLabs' voice synthesis AI model.[16] More recently, Microsoft and OpenAI reported on how state-sponsored hackers from North Korea, Iran, Russia, and China had misused ChatGPT for phishing and social engineering attacks targeting defense, cybersecurity, and cryptocurrency sectors.[17] Other investigations revealed that ChatGPT had been misused by cyber criminals to create malware and other malicious software.[18]

Specialized AI systems, which generally serve a single particular purpose, can also be harmfully misused. Ranking online search results provides a troubling example. Malicious actors can be especially effective at exploiting search result ranking with AI systems that exhibit full or very high levels of autonomy, misusing them to achieve harmful, nefarious outcomes.

For instance, antisemitic online groups tagged images of portable ovens on wheels with the label "Jewish baby stroller".[19] As a result, if users searched for the term "Jewish baby stroller", Google's algorithm ranked images of the ovens at the top of the search results. This was a direct exploitation of the image search algorithm's functionality, which works by matching the words in a query to the words that appear next to images on a webpage. The strategy succeeded particularly well because of a "data void" related to the search term: Because the product "Jewish baby stroller" doesn't actually exist, the only results available were the offensive images, which were then promoted by the algorithm.[20]

Malicious users have carried out similarly coordinated activities to trigger content moderation algorithms into removing marginalized creators' social media posts, a tactic known as "adversarial reporting."[21] Because content on social media sites is sometimes automatically removed when a sufficiently high number of users flag a post, regardless of whether or not the post actually violates any policies, right-wing trolls have strategically reported posts by influencers belonging to minority groups on TikTok in order to trigger the platform's content moderation algorithm.[22] Even if TikTok's appeal and review process finds that the video did not violate community guidelines, penalties become more severe the more frequently a creator's posts are flagged, and can range from content removal to account deletion. Automated content moderation systems are thus exploited to effectively censor marginalized communities online.

The Limits of Technical Mitigations of Misuse Risk

Developers of generative AI models can take steps to control model outputs so as to limit the generation of harmful content.[23] Risk-based release strategies that restrict access to particularly capable or advanced models can further help address misuse risks.[24] Assessing a model's propensity for misuse requires evaluating whether it can perform a given malicious task (plausibility) and, if so, how well it can do so (performance).[25] Red-teaming has emerged as a popular method to uncover misuse plausibility across a wide range of domains, and to surface where additional safeguards are needed.[26] Performance assessments should include benchmark evaluations and experiments, and focus on the marginal utility of the model compared to existing modes for task execution.[27]

Putting in place comprehensive safeguards is exceptionally challenging since it is difficult for developers to anticipate all potential (mis)uses of their model. Most importantly, such interventions at the model level will not necessarily prevent misuse without deteriorating model performance—also known as the Misuse-Use Tradeoff.[28] Because AI models lack the context to understand malicious intent, guardrails that prevent them from, for example, writing phishing emails will likely stop them from writing other emails as well. The same holds true for writing code: Guardrails to prevent malware may reduce the quality of innocuous computer programs. At the current state of the art, building an AI system that can never be misused often means building a system that is barely useful for non-malicious purposes. While there are other steps AI developers and deployers can take to prevent model misuse, technical fixes alone will not eliminate misuse risks.

Attacks on AI Systems

Harm can also result from cyberattacks on AI systems.* As with the misuse mechanism, the AI system developers and deployers do not intend harm here. Instead, the harmful intentions lie with the attackers. The cybersecurity community categorizes attacks into three groups: confidentiality, integrity, and availability attacks.[29] Confidentiality attacks aim to extract sensitive information, integrity attacks aim to compromise the model's

* Adversarial attacks on AI models can occur at every stage of the AI lifecycle. Since relevant incidents in the AIID result from attacks on deployed systems, this section only covers post-deployment attacks.

performance, and availability attacks aim to halt the overall functioning of the model. Lately, an emerging category of attacks aims to circumvent generative models' safeguards.[30] Such exploitations of AI systems' security vulnerabilities can potentially lead to harmful outcomes. Moreover, in addition to the standalone harms they cause, attacks on AI systems can enable model misuse.

While there is ample evidence of the security vulnerabilities of AI systems, most attacks recorded in the AIID still occur in experimental settings that do not lead to real-world harm. For example, security researchers have uncovered vulnerabilities in GitHub Copilot that would enable attackers to modify Copilot's responses or leak the developer's data (confidentiality and integrity attacks).[31] Experiments showed that flaws in Tesla's autopilot could be exploited to make the car accelerate and veer into the oncoming traffic lane (an integrity attack).[32] Finally, an investigation found that a divergence attack on ChatGPT could force the system to leak training data, including personal identifiable information such as phone numbers and email and physical addresses (a confidentiality attack).[33]

Harm incidents from the AIID show that in practice, attacks on AI systems are often carried out to evade generative AI model safeguards. This practice, called "jailbreaking," relies on prompt injection attacks in which users come up with text prompts that induce the AI model to behave in ways that violate its policies.[34] Prompt injection attacks enabled users to evade ChatGPT's guardrails shortly after its release in order to produce discriminatory and violent content, as well as offer instructions on how to carry out criminal activities.[35] Even after models have undergone extensive safety testing and red-teaming, prompt injection attacks remain a popular and effective technique to circumvent guardrails. Hackers using large language models for malware creation employ dozens of prompting strategies for various models such as OpenAI's GPT-4 and Anthropic's Claude in order to get the model to write phishing emails, build scam websites, or create malware.[36]

AI Security Is Not Like Traditional Software Security

Addressing vulnerabilities in AI and building more secure models is especially challenging since many established cybersecurity practices and techniques do not transfer seamlessly to machine learning models.[37] The probabilistic and data-dependent nature of machine learning models is inherently unsuitable for the standard patch-and-defend models of traditional cybersecurity. Effective patches to known vulnerabilities might not exist, or might be too costly to implement. Their implementation can also introduce other security issues and performance trade-offs that might be too severe to justify the fix. Defenses are also often not generalizable, meaning that measures against a specific attack do not protect the model against similar offensive operations of the same class.[38]

Given the wide range of attack options and the limited reliability of existing defenses, developers and deployers of AI models should take a systems-level approach to securing their AI models.[39] This means that they need to operate under the assumption that parts of the AI system can and will be compromised, and build in redundancies to raise system resilience. In other words, essential model outputs (decisions, predictions, actions) should not rely on single inputs, but instead on a combination of readings and assessments to prevent cascading failures and protect against attacks on individual model elements.

Unintentional Harm

AI Failures

The most prominent harm mechanism in the AIID is AI failure. AI failure is defined broadly as encompassing situations where the system makes an error, malfunctions, degrades in performance, or produces biased output. This behavior is generally unexpected by the developer, deployer, and user, and can have disastrous consequences when the AI system is used in high-risk contexts, as the following examples illustrate.

In what became a touchstone controversy in discussions of algorithmic bias, ProPublica evaluated a recidivism risk prediction algorithm called COMPAS in 2016 and found differential performance rates across racial groups.[40] The algorithm was used in sentencing and pretrial bond hearings across the U.S., influencing not only the severity of defendants' punishment but also bond amounts and pretrial detainment decisions.

ProPublica's investigation revealed that the algorithm incorrectly labeled black defendants as being at high risk of recidivism almost twice as often as white defendants, while white defendants were twice as likely to be incorrectly classified as low risk. Because the algorithm was used to guide judges' assessments of defendants, it had tangible impacts on their trial outcomes. In fact, after one defendant's sentencing was reduced on appeal, the judge admitted to being influenced towards a harsher sentence by the COMPAS-assigned risk score. ProPublica's analysis covered more than 18,000 people who were assigned a COMPAS risk score between 2013 and 2014. Over the course of those two years, the algorithm likely contributed to thousands of unjust decisions by setting excessive bond amounts and prolonged sentences for black defendants, and perpetuating and aggravating historical racial biases in law enforcement and the judicial system.[41]

Besides the justice system, flawed AI systems have also been deployed in healthcare. One example is the algorithm instituted as part of a Medicaid waiver program in Arkansas. The program in question provided home care for people with disabilities, and the algorithm relied on the responses to a health assessment questionnaire to determine the number of care hours that would be allocated to the beneficiary. Although many recipients' health status and care needs were unchanged from the previous year's assessment, the introduction of the algorithm led to drastic cuts in care hours. Only scrutiny of the algorithm in court revealed errors in how it handled major health issues such as diabetes and cerebral palsy, which caused incorrect calculations for more than 19% of program beneficiaries.[42] Despite ongoing complaints about the algorithm's decisions, neither the deploying agency nor the AI vendor detected the error until a lawsuit forced them to inspect the software, thus revealing the critical importance of opportunities to meaningfully challenge algorithmic decisions for those adversely affected.[43]

Other examples of AI failures in the AIID abound; generative AI systems have threatened users, created biased images, and spread misinformation.[44] LinkedIn's search algorithm at one point apparently favored men's profiles over women's.[45] AI-powered driver assistance technology has led to numerous accidents, some of them fatal.[46] And there are at least eleven incidents in the database for wrongful arrests caused by faulty facial recognition technology.[47]

The occurrence of these AI failure incidents indicates that at least one of two things is true (assuming use according to developers' instructions): Either existing evaluation, risk management, and assurance practices are not yet effective enough to ensure these AI systems can be deployed safely, or these practices are insufficiently implemented by

AI developers and deployers. Mitigation strategies need to address both kinds of shortcomings to reduce the chance of AI failures.

In the Absence of Standards and Regulation, Redress Procedures Can Mitigate Harm Impacts From AI Failures

Enhancing AI accuracy and reliability is a field of ongoing research, as is the development of testing and evaluation protocols that can reliably detect model performance issues (and assess performance, for that matter).[48] The AI standards landscape is also still in its infancy, meaning there are not yet many agreed-upon techniques and practices that developers can rely on to ensure their AI systems are safe, accurate, and reliable. Even if there were, without regulatory requirements developers and deployers are free to implement or ignore AI safety practices at their will. Costly interventions, such as pre-deployment AI testing in real world conditions or post-deployment audits, are much less likely to be implemented unless they are mandated.

Even with strong regulation and rigorous standards, some degree of AI failure is likely inevitable. It is therefore worth considering what deployers can do to reduce the extent of harm when those failures occur. The incidents highlighted above, for example, reveal the importance of opportunities to meaningfully challenge AI systems' decisions when errors, biases, or other failures occur. The justice system can provide an avenue for redress for affected persons, but only for those with sufficient resources to leverage it. Complaint and redress procedures that include human review of AI-supported decisions should be implemented at the level of the deployer to enhance access to remedy for algorithmic harm and prevent overburdening the court systems as AI systems become deployed more widely.

Failures of Human Oversight

In sensitive contexts, AI systems are often deployed as part of human-machine teams. Algorithmic decision-support systems require a human operator to remain in charge of the final decision while taking the AI system's recommendation into account. Ideally, human oversight over the AI system's output should detect abnormal behaviors, biases, and performance issues, and ensure that any that emerge are disregarded in the decision-making process. In practice, however, human supervisors often fail to both detect issues and overrule the AI system's recommendations when they should.

One striking example of this is the cancellation of thousands of immigrant visas by the UK government based on the results of a flawed voice recognition system.[49] Immigrants to the United Kingdom must demonstrate English-language competence by passing a test called the Test of English for International Communication (TOEIC), which is administered by the international testing organization ETS. In order to detect cheating through proxy test-takers, ETS deployed voice recognition software to determine if the same voice turned up on multiple test recordings. If a test was flagged, two ETS staff had to agree for the test results to be classified as invalid. During the three years of its use, 97% of TOEIC test recordings were flagged as suspicious by the voice recognition AI. Despite this obvious abnormality, reviewers proceeded to classify more than half of these tests as invalid (and all others as questionable), passing the list of invalid tests on to the UK government. Based on these results, UK policymakers canceled accused test-takers' visas and began deportations.[50]

The failure of oversight in this case likely stemmed from psychological propensities in how humans interact with AI systems. While it is unclear whether human reviewers were subject-matter experts or received training to assess voice similarities, it is likely that their decisions were influenced by two cognitive biases. The first, automation bias, is the tendency to accept machine-produced output as more objective and truthful than it is. The second, anchoring bias, is the tendency for a human to use an AI system's prediction as an "anchor" from which they are unlikely to stray far in their own response.[51] Biases can be hard to detect from individual decisions, but become apparent when a larger set of decisions is analyzed. Such a post hoc, large-scale assessment was clearly not performed in this case.

In addition to cognitive biases, there can also be structural forces that prevent human decision-makers from overruling an algorithm's recommendation, as the NarxCare case demonstrates. NarxCare is an algorithmic tool built on top of a national prescription drug database that provides doctors and pharmacists with data analytics and healthcare management functions. Among them is an automatically generated patient risk score for opioid misuse and overdose.[52] While the company that provides the tool emphasizes that the score should be used to inform rather than automate decisions concerning patient care, practitioners face major obstacles to overriding the predicted scores. In many states, physicians and pharmacists are legally required to consult the program before prescribing controlled substances, and in some states law enforcement can access the data without a warrant to prosecute both doctors and patients. As a result, doctors have become increasingly anxious about appearing to over-prescribe medications. Fearful of losing their license, prescribers err on the side of extreme caution and rely heavily on the risk scores when making medication decisions, which has led to an overall reduction in prescriptions even for those in legitimate need.[53] In effect, the combination of surveillance and mandatory use of the algorithmic tool undermines doctors' agency over patient care. For chronic pain patients and those with multiple and severe illnesses, this has resulted in difficulties obtaining their needed medication, more pain, and reduced quality of life.

Another common failure of human oversight occurs during the operation of vehicles equipped with autopilot or driver-assist technology. When it comes to partly automated driving, humans are "on the loop" rather than "in the loop." This distinction, while subtle, encompasses a substantial shift in how humans are involved in the decision-action process, moving from an active role to a passive one. Being "on-the-loop" in a partly self-driving car means that the system operates autonomously under the supervision of the driver, who in theory is ready to take control at any point. Humans, however, are not well-equipped to vigilantly but passively monitor a system for long periods of time.[54] As a result, they can miss or react too slowly in situations where they should intervene, leading to accidents. This failure of human-machine teaming can be observed in the dozens of incidents involving AI-enabled cars in the AIID.[55] The National Highway Traffic Safety Administration (NHTSA), a U.S. government agency charged with investigating major self-driving car manufacturers, has termed the problem "automation complacency," a combination of excessive trust in the system's capabilities and the human susceptibility to disengage from monitoring tasks.[56] According to the agency's findings, automation complacency is at least partly driven by inappropriate design choices by the manufacturers. NHTSA's investigation of traffic accidents involving AI-enabled driver assistance technologies revealed many instances in which the vehicle's autopilot passed control of the vehicle to the human operator less than one second before impact, giving the operator insufficient time to react and take full control of the vehicle to prevent the accident.[57] Improving human-machine interfaces to account for human capabilities and limitations, such as attention spans and reaction times, is critical to enable humans-on-the-loop to take control of an AI system when the situation requires it.

The Challenges of Mitigating Cognitive Biases in Human-AI Teams

Cognitive biases, the institutional environment, and AI system design are important factors determining the success of human oversight of AI systems.[58] These aspects need to be carefully considered and addressed in order to make human oversight a meaningful mitigator of AI harm.

However, reducing cognitive biases in human-AI interactions remains an unsolved challenge, because many measures that are expected to alleviate biases can also aggravate them. AI explainability features are often believed to mitigate overreliance tendencies and help build confidence in the system's output, but have been found to worsen cognitive biases in some cases.[59] Similar effects were found for AI literacy: While lower AI literacy is indeed associated with higher reliance on AI recommendations, higher AI literacy was associated with algorithmic aversion—meaning an excessive distrust of the AI's recommendation—which led to lower accuracy in decision outcomes.[60]

More experimental research on human factor engineering and human computer interaction is needed to uncover and evaluate techniques that mitigate cognitive biases, safeguard decision-makers' agency, and build trust in human-AI teams.

Integration Harm

Finally, AI systems used in inappropriate contexts can cause harm even when they function as intended, are not meant to be harmful, and are neither being attacked nor misused. In these cases, harm arises as an unintended consequence of the integration of AI in its deployment setting. Compared to the other harm mechanisms, the trigger in these cases is not a malfunction or a malicious user, but rather the deployment of an otherwise functional system in a mismatched context. Similar to when AI systems are designed to be harmful, integration harm emerges as a direct consequence of deployment of the system, but as an unintended side effect rather than a stated goal.

Viewed independently, an AI tool may serve a seemingly harmless function, such as ranking web pages to facilitate online search or predicting foot traffic to a store. It may even do so very successfully, but its integration into a deployment environment can lead to unintended side effects. This is because all AI tools are integrated into existing systems. A store traffic prediction tool becomes part of a workplace and is integrated into existing workflows. An online search engine is situated within the context of our online information environment. Incorporating AI systems into these contexts generates changes, and potentially harm, in these environments. The following examples illustrate instances where failure to account for context has led to harm.

Over the past two decades, online search has become the predominant way people access health-related information.[61] Search engines, powered by page-ranking algorithms, directly influence what information is found and consumed by internet users via the order in which search results are presented. During the coronavirus pandemic, anti-vaccine misinformation was boosted by those algorithms powering online search engines.[62] An audit of Amazon's algorithms revealed that the platform not only hosts a wide range of products that promote misinformation, such as books containing conspiracy theories about vaccines, but its search engine also exhibits ranking bias in favor of those products. When users search for popular vaccine-related terms, it displays products that misinform about vaccines above those that provide accurate and debunking information.[63] The audit also demonstrated the filter bubble effect of the platform's recommender algorithm: Users that interacted with misinforming products were more likely to encounter misinformation on their Amazon homepage and among their recommended products. In this way, the regular functioning of the search and recommendation algorithms, which intend to promote popular and relevant products to customers, had the unintended side effect of promoting misinformation to Amazon's customers.

Unintended consequences can also emerge as a result of one-sided consideration of stakeholders' needs in algorithmic design. This was the case when Starbucks deployed a scheduling algorithm across its stores. The AI system was created to optimize employees' shift allocation based on predicted store traffic. But its deployment resulted in constantly-changing shift schedules—often delivered to workers on short notice—and dramatically varying weekly hours.[64] The tool prioritized cost savings for the retail chain while disregarding the needs of the workforce to have predictable schedules and incomes. While not intentionally harmful—Starbucks failed to take the needs of its staff into account, but it is unlikely it deployed the system in order to cause them hardship—the use of the system severely affected workers' financial stability and their ability to plan and manage their non-work life.[65]

Finally, the deployment of AI systems can introduce opportunity costs by disrupting workflows and diverting resources. One example of this is the deployment of ShotSpotter, an acoustic gunfire detection system that was rolled out by the Chicago police department in 2016. An evaluation from 2024 found that in the six years that followed deployment, the AI system led to approximately 70 dispatches per day, a two-fold increase compared to pre-deployment. This increased demand for officer resources affected response times to 911 calls. Officers were dispatched to 911 calls more slowly, arrived at the scene later, and were less likely to arrest the perpetrator. The deployment of the gunfire detection system therefore reduced the effectiveness of the police force in responding to citizens' emergencies in Chicago.[66]

Algorithmic Impact Assessments Can Mitigate Risks of Integration Harm

Reducing the risk of integration harm requires predicting potential impacts of AI systems before their deployment, and adjusting algorithmic design accordingly. Ex ante AI impact assessments, such as the fundamental rights impact assessment required by the EU AI Act, are one example of a measure to achieve this. The goal of AI impact assessments is to surface and assess the broad range of harms potentially resulting from an AI system's use through analysis of the AI use case and environment, and through consultations with people affected by its deployment. Algorithmic impact assessments give the deployer and affected stakeholders an opportunity to evaluate the effects of an AI system before committing to its adoption.[67] Doing so effectively requires both knowledge of how the AI system works and a deep understanding of the context of use. Often, this entails engaging diverse stakeholders to survey the concerns and perspectives from affected groups. Impact assessments should be conducted on a use case-by-use case basis, and tailored to each deployment context and purpose. Their findings should inform algorithmic design choices, risk mitigation strategies, and ultimately deployment decisions.

Discussion

Investigating and categorizing AI harm mechanisms offers valuable insights. First, it surfaces the diverse ways in which the use of AI can cause harm, intentionally or unintentionally. Recognizing that AI systems do not have to malfunction or be misused for their use to be harmful is an important first step in broadening approaches to mitigating risks. Integration harm in particular is an often-overlooked mechanism due to the difficulty of anticipating the multifaceted impact of AI deployment. By identifying and explaining the variety of AI harm mechanisms, this report aims to ensure that risk mitigation efforts address all potential pathways to harm.

Second, considering the breadth of AI harm mechanisms demonstrates the need for equally diverse approaches to mitigation. A one-size-fits-all approach will not work. Instead, a whole variety of actors in the AI value chain (developers, deployers, policymakers, etc.) can take a wide-ranging set of sociotechnical mitigations to reduce the risk of different harm mechanisms occurring. As argued in this report, technical fixes to AI models alone will not prevent AI harm, and actions must range from implementing testing and evaluation protocols to participatory processes for AI adoption.

Third, dissecting the various harm mechanisms reveals that an AI system's propensity to harm is not always tied to its capabilities. In fact, many of the examples in the text illustrate cases where harm emerged from specialized, single-purpose AI, demonstrating that harm can and does result from all kinds of AI systems, irrespective of their size, power, and sophistication (often measured by proxy through the compute required to train them). A stronger determinant of harm risks is the context, manner, and governance of AI design, testing, and deployment.

Finally, incident reports of real-world harm are invaluable resources for learning about AI risks, offering essential insights for designing better mitigation strategies and targeted policy interventions. Many efforts, including CSET's AI harm framework and taxonomy, and the MIT AI incident tracker, are underway to make existing incident data more accessible and analyzable in order to better derive policy insights.[68] However, data limitations—some of which are described in the limitations section of this report—remain an important obstacle to rigorous analysis. A formal incident reporting regime could overcome the shortcomings of existing tracking efforts by gathering detailed information about the AI system involved. Instituting such a mandatory incident reporting regime should be a priority for policymakers interested in promoting safe and responsible AI deployment.

Conclusion

As AI capabilities advance and the variety of potential use cases grows, public and private sector voices alike are pushing for a more forceful deployment of AI to realize productivity and other benefits. Before incorporating AI into every aspect of our lives, however, it is critical to better understand the risk of harm from doing so. Widespread AI deployment and use has already led to hundreds of recorded harm incidents. Better preventing harms in the future requires an improved understanding of how harm happens in practice. To this end, this analysis leverages incident reports from the AI incident database to identify six key mechanisms of harm that describe how the use of AI can lead to harmful outcomes (Table 3).

Table 3: The Six AI Harm Mechanisms

Intentional Harm	Unintentional Harm
• Harm by design • AI misuse • Attacks on AI systems	• AI failures • Failures of human oversight • Integration harm

A review of incidents associated with these mechanisms revealed several key takeaways that should guide policy efforts to reduce harm occurrences in the future:

1. **A one-size-fits-all approach to harm mitigation will not work.** The pathways to harm are diverse, as this report illustrates, and require equally diverse mitigation strategies. Purely technical approaches will fall short, especially in addressing integration harms and failures of human oversight.

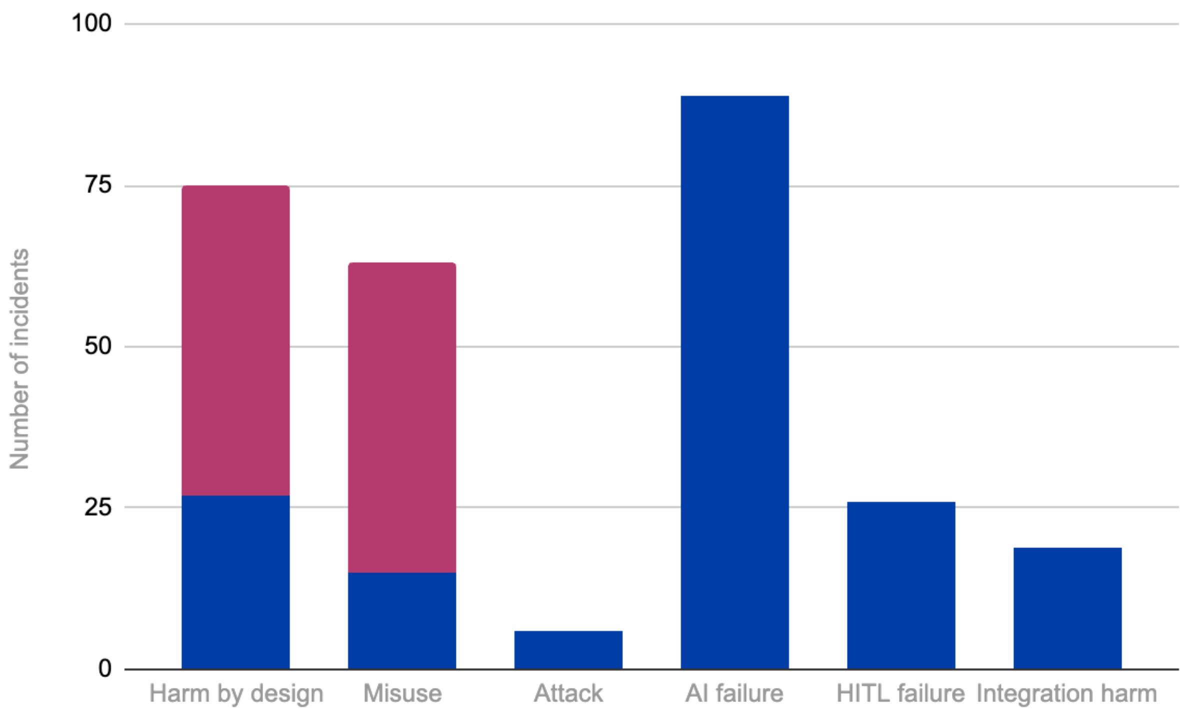
2. **Model capabilities, as proxied by computing power, are an inadequate predictor for the propensity to do harm.** This report showcases many examples of single-purpose AI systems being implicated in harm. Concentrating risk mitigation efforts on advanced AI systems would fail to address the very real risks stemming from the irresponsible design, deployment, and use of specialized AI systems.

3. **Comprehensive incident tracking is necessary to enhance our capacity to identify and respond to risks posed by AI.** While implementing broad, sociotechnical mitigation strategies can significantly reduce the occurrence of harm from AI, it will not prevent incidents entirely. As AI innovation reveals new capabilities with new failure modes, deployers design new use cases, and

nefarious actors find new ways to attack and misuse AI systems, new harms will emerge. Agile responses and rapid adaptation of mitigating approaches, enabled by effective learning from incident reporting, are necessary to keep pace with technological innovation.

Appendix

Figure 1: Distribution of AI Harm Mechanisms



Note: Annotation of harm mechanisms for a random sample of 200 AI incidents. Numbers do not add up to 200 because multiple mechanisms can contribute to a single incident. Harm by design and misuse are represented by two bars. The blue bars reflect confirmed cases, where information about the AI system used was available and sufficient to determine whether a model was misused or developed for the purpose of harm. The red bars represent the 48 incidents where this determination was not possible, i.e., that could be cases of harm by design or misuse.

Source: Author's analysis based on data from the AI Incident Database.

The figure shows the distribution of harm mechanisms across a random sample of 200 incidents from the AIID. AI failures are by far the dominant cause of harm, with 89 incidents of the sample linked to this mechanism. This suggests that AI systems continue to be deployed for purposes that they cannot (yet) reliably perform, resulting in significant harm.

Aside from AI failures, the prevalence of cases of harm by design and misuse is noteworthy. Understanding of the underlying problem, however, is impeded by the lack of insight into the AI system used in a given incident. Is it the case that a growing number of actors have the resources to build AI systems for their needs, resulting in greater availability of harmful models such as AI-powered "nudify" apps or malware-producing generative AI? Or are users abusing AI systems against the developers' intentions at significant rates? To better answer these questions, and to allocate appropriate resources to the mitigation of these risks, it is imperative to invest in incident tracking and data collection.

Overall, most incidents emerge from unintentional harm mechanisms. This makes mitigation more complex, because it is impossible to pinpoint the cause of harm as precisely as for intentional mechanisms. The ability to identify a motivated actor such as an attacker or a nefarious user as a source of harm allows the creation of specific and targeted mitigation strategies. But for unintentional harm mechanisms, and especially integration harms and human oversight failures, a more nuanced, multipronged approach is necessary—often one that involves more stakeholders.

Author

Mia Hoffmann is a research fellow at CSET where her work focuses on AI governance.

Acknowledgments

The author would like to thank Josh Goldstein, Colin Shea-Blymyer, Mina Narayanan, Madhu Srikumar, and Danny Atherton for comments on earlier drafts of this report, and Owen Daniels for his feedback and support.

Document Identifier: doi: 10.51593/20240052

Endnotes

[1] Sean McGregor, "Preventing Repeated Real World AI Failures by Cataloging Incidents: The AI Incident Database," in *Proceedings of the Thirty-Third Annual Conference on Innovative Applications of Artificial Intelligence (IAAI-21)* 35, no. 17 (2021).

[2] Office of Management and Budget, *Accelerating Federal Use of AI through Innovation, Governance, and Public Trust* (Washington, D.C.: Executive Office of the President, April 3, 2025), https://www.whitehouse.gov/wp-content/uploads/2025/02/M-25-21-Accelerating-Federal-Use-of-AI-through-Innovation-Governance-and-Public-Trust.pdf; Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonized rules on artificial intelligence (Artificial Intelligence Act), OJ L, 2024/1689, https://eur-lex.europa.eu/eli/reg/2024/1689/oj; "Anthropic's Responsible Scaling Policy," Anthropic, May 14, 2025, https://www.anthropic.com/rsp-updates.

[3] Heather Frase and Mia Hoffmann, "Adding Structure to AI Harm" (Center for Security and Emerging Technology, July 2023), https://cset.georgetown.edu/publication/adding-structure-to-ai-harm/.

[4] "CSETv1 Charts," AI Incident Database, https://incidentdatabase.ai/taxonomy/csetv1/.

[5] European Commission, "The General-Purpose AI Code of Practice," https://digital-strategy.ec.europa.eu/en/policies/contents-code-gpai.

[6] Frase and Hoffmann, "Adding Structure to AI Harm."

[7] "CSETv1 Charts," AI Incident Database, https://incidentdatabase.ai/taxonomy/csetv1/.

[8] Incidents 544, 565, and 673 from the AI Incident Database illustrate this well.

[9] Kateryna Stepanenko, "The Battlefield AI Revolution Is Not Here Yet: The Status of Current Russian and Ukrainian AI Drone Efforts," (Institute for the Study of War, June 2, 2025), https://understandingwar.org/research/russia-ukraine/the-battlefield-ai-revolution-is-not-here-yet-the-status-of-current-russian-and-ukrainian-ai-drone-efforts/; Sheera Frenkel and Natan Odenheimer, "Israel's A.I. Experiments in Gaza War Raise Ethical Concern," *The New York Times*, April 25, 2025, https://www.nytimes.com/2025/04/25/technology/israel-gaza-ai.html; "Incident Number 672: Lavender AI System Reportedly Directs Gaza Strikes with High Civilian Casualty Rate," Artificial Intelligence Incident Database, Responsible AI Collaborative, April 3, 2024, https://incidentdatabase.ai/cite/672/.

[10] See e.g. incidents 480, 530, 597, 765, 769, 771, 772, 777, 874, and 903 from the AI Incident Database.

[11] Megan Hughes, "How Technology Can enable Violence Against Women and Girls" (Center for Emerging Technology and Security, August 2024), https://cetas.turing.ac.uk/publications/how-technology-can-enable-violence-against-women-and-girls.

[12] Matt Burgess, "Deepfake Porn Is Out of Control," *Wired*, October 16, 2023, https://www.wired.com/story/deepfake-porn-is-out-of-control/?_sp=7405a578-d28d-4413-9f8a-8fbbb4387704.1740429425353.

[13] "Incident Number 191: Korean Internet Portal Giant Naver Manipulated Shopping and Video Search Algorithms to Favor In-House Services," Artificial Intelligence Incident Database, Responsible AI Collaborative, October 6, 2020, https://incidentdatabase.ai/cite/191/; Sonali Pednekar, "Incident Number 435: Coupang Allegedly Tweaked Search Algorithms to Boost Own Product," Artificial Intelligence Incident Database, Responsible AI Collaborative, July 4, 2025, https://incidentdatabase.ai/cite/435/; Sonali Pednekar, "Incident Number 437: Amazon India Allegedly Rigged Search Results to Promote Own Products," Artificial Intelligence Incident Database, Responsible AI Collaborative, December 31, 2016, https://incidentdatabase.ai/cite/437/.

[14] Khoa Lam, "Incident Number 339: Open-Source Generative Models Abused by Students to Cheat on Assignments and Exams," Artificial Intelligence Incident Database, Responsible AI Collaborative, September 15, 2022), https://incidentdatabase.ai/cite/339/.

[15] Natalya Kosmyma, Eugene Hauptmann, Ye Tong Yuan, Jessica Situ, Xian-Hao Liao, Ashly Vivian Beresnitzky, Iris Braunstein and Pattie Maes, "Your Brain on ChatGPT: Accumulation of Cognitive Debt When Using an AI Assistant for Essay Writing Task," arXiv preprint arXiv:2506.08872, https://arxiv.org/pdf/2506.08872v1.

[16] Janet Schwartz, "Incident Number 508: Celebrities' Deepfake Voices Abused with Malicious Intent," Artificial Intelligence Incident Database, Responsible AI Collaborative, January 30, 2023, https://incidentdatabase.ai/cite/508/.

[17] Daniel Atherton, "Incident Number 644: Alleged State-Sponsored Hackers Escalate Purported Phishing Attacks Using Artificial Intelligence," Artificial Intelligence Incident Database, Responsible AI Collaborative, February 18, 2024, https://incidentdatabase.ai/cite/644/.

[18] "Incident Number 736: Underground Market for LLMs Powers Malware and Phishing Scams," Artificial Intelligence Incident Database, Responsible AI Collaborative, December 1, 2023, https://incidentdatabase.ai/cite/736/; Daniel Atherton, "Incident Number 443: ChatGPT Abused to Develop Malicious Softwares," Artificial Intelligence Incident Database, Responsible AI Collaborative, December 21, 2022, https://incidentdatabase.ai/cite/443/.

[19] Devon Colmer, "Incident Number 1057: 'Jewish Baby Strollers' Provided Anti-Semitic Google Images, Allegedly Resulting from Hate Speech Campaign," Artificial Intelligence Incident Database, Responsible AI Collaborative, August 15, 2017, https://incidentdatabase.ai/cite/1057/.

[20] Michael Golebiewski and Danah Boyd, "Data Voids: Where Missing Data Can Easily Be Exploited" (Data & Society, May 11, 2018), https://datasociety.net/library/data-voids-where-missing-data-can-easily-be-exploited/.

[21] For another type of adversarial reporting, see e.g. Shelby Grossman, Christopher Giles, Cynthia N.M., R. Miles McCain, Blair Read, "The New Copyright Trolls: How a Twitter Network Used Copyright

Complaints to Harass Tanzanian Activitists" (Stanford Internet Observatory, December 2, 2021), https://cyber.fsi.stanford.edu/io/publication/new-copyright-trolls.

[22] Patrick Hall, "Incident Number 133: Online Trolls Allegedly Abused TikTok's Automated Content Reporting System to Discriminate Against Marginalized Creators," Artificial Intelligence Incident Database, Responsible AI Collaborative, December 15, 2020, https://incidentdatabase.ai/cite/133.

[23] Jessica Ji, Josh A. Goldstein, and Andrew Lohn, "Controlling Large Language Model Outputs: A Primer" (Center for Security and Emerging Technology, December 2023), https://cset.georgetown.edu/publication/controlling-large-language-models-a-primer/.

[24] Jon Bateman, Dan Baer, Stephanie A. Bell, Glenn O. Brown et al., "Beyond Open vs. Closed: Emerging Consensus and Key Questions for Foundation AI Model Governance" (Carnegie Endowment for International Peace, July 23, 2024), https://carnegieendowment.org/research/2024/07/beyond-open-vs-closed-emerging-consensus-and-key-questions-for-foundation-ai-model-governance?lang=en.

[25] Josh A. Goldstein and Girish Sastry, "The PPOu Framework: A Structured Approach for Assessing the Likelihood of Malicious Use of Advanced AI Systems," *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* 7, no. 1 (October 2024): 503–518, https://doi.org/10.1609/aies.v7i1.31653.

[26] Jessica Ji, "What Does AI Red-Teaming Actually Mean?," CSET blog, October 24, 2023, https://cset.georgetown.edu/article/what-does-ai-red-teaming-actually-mean/.

[27] Thomas Woodside and Helen Toner, "Evaluating Large Language Models," CSET blog, July 17, 2024, https://cset.georgetown.edu/article/evaluating-large-language-models/; Goldstein and Sastry, "The PPOu Framework."

[28] Markus Anderljung, Julian Hazell, and Moritz von Knebel, "Protecting Society From AI Misuse: When Are Restrictions on Capabilities Warranted?", AI & Society, 40: 3841–3857 (2025) https://doi.org/10.1007/s00146-024-02130-8; Arvind Narayanan and Sayash Kapoor, "AI Safety Is Not a Model Property," AI Snake Oil, March 12, 2024, https://www.aisnakeoil.com/p/ai-safety-is-not-a-model-property.

[29] Andrew Lohn, "Hacking AI: A Primer for Policymakers on Machine Learning Cybersecurity" (Center for Security and Emerging Technology, December 2020), https://cset.georgetown.edu/publication/hacking-ai/.

[30] Apostol Vassilev, Alina Oprea, Alie Fordyce, and Hyrum Anderson, *Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations* (Washington, D.C.: National Institute of Standards and Technology, January 2024), https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-2e2023.pdf.

[31] Marlon Fabiano, "Zeroday on Github [sic] Copilot," gccybermonks, July 27, 2024, https://gccybermonks.com/posts/github/, retrieved from https://incidentdatabase.ai/reports/3996/.

[32] Patrick Howell O'Neill, "Hackers Can Trick a Tesla Into Accelerating by 50 Miles Per Hour", *MIT Technology Review*, February 19, 2020, retrieved from https://incidentdatabase.ai/reports/2316/; Evan

Ackerman, “Three Small Stickers in Intersection Can Cause Tesla Autopilot to Swerve Into Wrong Lane,” IEEE Spectrum, April 1, 2019, retrieved from https://incidentdatabase.ai/reports/1518/.

[33] Milad Nasr, Nicholas Carlini, Jonathan Hayase et al., “Scalable Extraction of Training Data from (Production) Language Models”, arXiv preprint arXiv: 2311.17035 (2023), https://arxiv.org/abs/2311.17035.

[34] Daniel Atherton, “Incident Number 473: Bing Chat’s Initial Prompts Revealed by Early Testers Through Prompt Injection,” Artificial Intelligence Incident Database, Responsible AI Collaborative, February 8, 2023, https://incidentdatabase.ai/cite/473/; Collin Starkweather, “Incident Number 677: ChatGPT and Perplexity Reportedly Manipulated into Breaking Content Policies in AI Boyfriend Scenarios,” Artificial Intelligence Incident Database, Responsible AI Collaborative, April 29, 2024, https://incidentdatabase.ai/cite/677.

[35] Khoa Lam, “Incident Number 420: Users Bypassed ChatGPT's Content Filters with Ease,” in Artificial Intelligence Incident Database, Responsible AI Collaborative, November 30, 2022, https://incidentdatabase.ai/cite/420.

[36] “Incident Number 736: Underground Market for LLMs Powers Malware and Phishing Scams,” Artificial Intelligence Incident Database, Responsible AI Collaborative, December 1, 2023, https://incidentdatabase.ai/cite/736/; David Winder, “New AI Attack Compromises Google Chrome’s Password Manager,” *Forbes*, March 23, 2025, https://incidentdatabase.ai/reports/5011/.

[37] Andrew Lohn and Wyatt Hoffman, “Securing AI: How Traditional Vulnerability Disclosure Must Adapt” (Center for Security and Emerging Technology, March 2022), https://cset.georgetown.edu/publication/securing-ai-how-traditional-vulnerability-disclosure-must-adapt/.

[38] Tom Tseng, Euan McLean, Kellin Pelrine, Tony T. Wang, and Adam Gleave, “Can Go AIs Be Adversarially Robust?,” arXiv preprint arXiv:2406.12843 (2025), https://arxiv.org/pdf/2406.12843.

[39] Lohn, “Hacking AI.”

[40] Jeff Larson, Surya Mattu, Lauren Kirchner, and Julia Angwin, “How We Analyzed the COMPAS Recidivism Algorithm,” ProPublica, May 23, 2016, https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm; Roman Yampolskiy, “Incident Number 11: Northpointe Risk Models,” Artificial Intelligence Incident Database, Responsible AI Collaborative, May 23, 2016, https://incidentdatabase.ai/cite/11/.

[41] Larson et al., “How We Analyzed COMPAS”; Yampolskiy, “Incident Number 11.”

[42] Roman Lutz, “Incident Number 110: Arkansas’s Opaque Algorithm to Allocate Health Care Excessively Cut Down Hours for Beneficiaries,” Artificial Intelligence Incident Database, Responsible AI Collaborative, January 1, 2016, https://incidentdatabase.ai/cite/110.

[43] Lutz, Incident Number 110.

[44] Khoa Lam, “Incident Number 503: Bing AI Search Tool Reportedly Declared Threats Against Users,” Artificial Intelligence Incident Database, Responsible AI Collaborative, February 14, 2023, https://incidentdatabase.ai/cite/503/; Daniel Atherton, “Incident Number 645: Seeming Pattern of Gemini Bias and Sociotechnical Training Failures Harm Google’s Reputation,” Artificial Intelligence Incident Database, Responsible AI Collaborative, February 21, 2024, https://incidentdatabase.ai/cite/645/; Logan B, “Incident Number 763: Grok AI Chatbot Reportedly Spreads Unfounded Rumors About Trump’s Dentures,” Artificial Intelligence Incident Database, Responsible AI Collaborative, August 13, 2024, https://incidentdatabase.ai/cite/763/.

[45] Roman Yampolskiy, “Incident Number 47: LinkedIn Search Prefers Male Names,” Artificial Intelligence Incident Database, Responsible AI Collaborative, September 6, 2016, https://incidentdatabase.ai/cite/47/.

[46] “Incident Number 20: A Collection of Tesla Autopilot-Involved Crashes,” Artificial Intelligence Incident Database, Responsible AI Collaborative. June 30, 2016, https://incidentdatabase.ai/cite/20/.

[47] Thomas Giallella, “Incident Number 295: Wrongful Attempted Arrest for Apple Store Thefts Due to NYPD’s Facial Misidentification,” Artificial Intelligence Incident Database, Responsible AI Collaborative, November 8, 2018, https://incidentdatabase.ai/cite/295/; Daniel Atherton, “Incident Number 515: Facial Recognition Error Reportedly Leads to Wrongful Arrest of Georgia Man and $200K Settlement in Louisiana,” Artificial Intelligence Incident Database, Responsible AI Collaborative, November 25, 2022, https://incidentdatabase.ai/cite/515/; Khoa Lam, “Incident Number 517: Man Arrested For Sock Theft by False Facial Match Despite Alibi,” Artificial Intelligence Incident Database, Responsible AI Collaborative, February 15, 2018, https://incidentdatabase.ai/cite/517/; Khoa Lam, “Incident Number 288: New Jersey Police Wrongful Arrested Innocent Black Man via FRT,” Artificial Intelligence Incident Database, Responsible AI Collaborative, January 30, 2019, https://incidentdatabase.ai/cite/288/; Daniel Atherton, “Incident Number 630: Alleged Macy’s Facial Recognition Error Leads to Wrongful Arrest and Subsequent Sexual Assault in Jail,” Artificial Intelligence Incident Database, Responsible AI Collaborative, January 22, 2022, https://incidentdatabase.ai/cite/630/; Kate Perkins, “Incident Number 439: Detroit Police Wrongfully Arrested Black Man Due to Faulty Facial Recognition,” Artificial Intelligence Incident Database, Responsible AI Collaborative, July 31, 2019, https://incidentdatabase.ai/cite/439/; Daniel Atherton, “Incident Number 598: False Arrest of Georgia Man Due to Louisiana Police’s Faulty Facial Recognition Technology,” Artificial Intelligence Incident Database, Responsible AI Collaborative, November 25, 2022, https://incidentdatabase.ai/cite/598/; Nickie Demakos, “Incident Number 74: Detroit Police Wrongfully Arrested Black Man Due to Faulty FRT,” Artificial Intelligence Incident Database, Responsible AI Collaborative, January 30, 2020, https://incidentdatabase.ai/cite/74/; Khoa Lam, “Incident Number 440: Louisiana Police Wrongfully Arrested Black Man Using False Face Match,” Artificial Intelligence Incident Database, Responsible AI Collaborative, November 25, 2022, https://incidentdatabase.ai/cite/440/; Daniel Atherton, “Incident Number 692: London Metropolitan Police’s Facial Recognition Technology Reportedly Misidentified Shaun Thompson as Suspect Leading to Arrest,” Artificial Intelligence Incident Database, Responsible AI Collaborative, February 1, 2024, https://incidentdatabase.ai/cite/692; Daniel Atherton, “Incident Number 816: Cross-Jurisdictional Facial Recognition Misidentification by NYPD Leads to Wrongful Arrest and Four-Year Jail Time in New Jersey,”

Artificial Intelligence Incident Database, Responsible AI Collaborative, November 29, 2019, https://incidentdatabase.ai/cite/816.

[48] Anka Reuel, Amelia Hardy, Chandler Smith, Max Lamparth et al., "BetterBench: A Repository of AI Benchmark Assessments for Informed Benchmark Selection Through Quality Evaluation and Best Practice Analysis," arXiv preprint arXiv:2411.12990 (2024), https://arxiv.org/abs/2411.12990.

[49] "Incident Number 162: ETS Used Allegedly Flawed Voice Recognition Evidence to Accuse and Assess Scale of Cheating, Causing Thousands to be Deported from the UK," Artificial Intelligence Incident Database, Responsible AI Collaborative, January 1, 2014, http://www.incidentdatabase.ai/cite/162.

[50] Ed Main and Richard Watson, "The English Test That Ruined Thousands of Lives," *BBC*, February 8, 2022, https://www.bbc.com/news/uk-60264106.

[51] Charvi Rastogi, Yunfeng Zhang, Dennis Wei et al., "Deciding Fast and Slow: The Role of Cognitive Biases in AI-assisted Decision-making," *Proceedings of the ACM on Human-Computer Interaction* 6, no. 83 (April 2022): 1-22, https://doi.org/10.1145/3512930.

[52] Sean McGregor, "Incident Number 172: NarxCare's Risk Score Model Allegedly Lacked Validation and Trained on Data with High Risk of Bias," Artificial Intelligence Incident Database, Responsible AI Collaborative, July 1, 2020, https://incidentdatabase.ai/cite/172/.

[53] Maya Szalavitz, "The Pain Was Unbearable. So Why Did Doctors Turn Her Away?" *Wired*, August 11, 2021, https://www.wired.com/story/opioid-drug-addiction-algorithm-chronic-pain/.

[54] N.H. Mackworth, "The Breakdown of Vigilance During Prolonged Visual Search," *Quarterly Journal of Experimental Psychology* 1, no. 1 (April 1948): 6–21, https://doi.org/10.1080/17470214808416738; Eric T. Greenlee, Patricia R. DeLucia, and David C. Newton, "Driver Vigilance in Automated Vehicles: Hazard Detection Failures Are a Matter of Time," *Human Factors* 60, no. 4 (March 2018): 465–476, https://doi.org/10.1177/0018720818761711.

[55] See e.g. Catherine Olsson, "Incident Number 4: Uber AV Killed Pedestrian in Arizona," Artificial Intelligence Incident Database, Responsible AI Collaborative, March 18, 2018, https://incidentdatabase.ai/cite/4; Khoa Lam, "Incident Number 306: Tesla on Autopilot TACC Crashed Into Van on European Highway," Artificial Intelligence Incident Database, Responsible AI Collaborative, May 26, 2016, https://incidentdatabase.ai/cite/306; "Incident Number 321: Tesla Model X on Autopilot Crashed Into California Highway Barrier, Killing Driver," Artificial Intelligence Incident Database, Responsible AI Collaborative, March 23, 2018, https://incidentdatabase.ai/cite/321; Khoa Lam, "Incident Number 353: Tesla on Autopilot Crashed Into Trailer Truck in Florida, Killing Driver," Artificial Intelligence Incident Database, Responsible AI Collaborative, March 1, 2019, https://incidentdatabase.ai/cite/353; and Daniel Atherton, "Incident Number 434: Sudden Braking by Tesla Allegedly on Self-Driving Mode Caused Multi-Car Pileup in Tunnel," Artificial Intelligence Incident Database, Responsible AI Collaborative, November 24, 2022, https://incidentdatabase.ai/cite/434.

[56] National Transportation Safety Board, "Investigative Outcomes and Recommendations," NTSB, December 4, 2024, https://www.ntsb.gov/Advocacy/SafetyIssues/Pages/Vehicle-Automations-Investigative-Outcomes.aspx.

[57] Christiaan Hetzner, "Elon Musk's Regulatory Woes Mount as U.S. Moves Closer to Recalling Tesla's Self-Driving Software," *Fortune*, June 10, 2022, https://fortune.com/2022/06/10/elon-musk-tesla-nhtsa-investigation-traffic-safety-autonomous-fsd-fatal-probe/.

[58] Lauren Kahn, Emelia S. Probasco, and Ronnie Kinoshita, "AI Safety and Automation Bias" (Center for Security and Emerging Technology, November 2024), https://cset.georgetown.edu/publication/ai-safety-and-automation-bias/.

[59] Astrid Betrand, Rafik Belloum, James R. Eagan, and Winston Maxwell, "How Cognitive Biases Affect XAI-Assisted Decision-Making: A Systematic Review," *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society (AIES '22)* (July 2022): 78:91, https://doi.org/10.1145/3514094.3534164.

[60] Maja Jacobs, Melanie F. Pradier, Thomas H. McCoy Jr. et al., "How Machine-Learning Recommendations Influence Clinician Treatment Selections: The Example of Antidepressant Selection," *Translational Psychiatry* 11, no. 108 (February 2021), https://www.nature.com/articles/s41398-021-01224-x.

[61] "Most internet users start at a search engine when looking for health information online" (Pew Research Center, October 29, 2006), https://www.pewresearch.org/internet/2006/10/29/most-internet-users-start-at-a-search-engine-when-looking-for-health-information-online/.

[62] Sean McGregor, "Incident Number 139: Amazon's Search and Recommendation Algorithms Found by Auditors to Have Boosted Products That Contained Vaccine Misinformation," Artificial Intelligence Incident Database, Responsible AI Collaborative, January 21, 2021, http://www.incidentdatabase.ai/cite/139/.

[63] Prerna Juneja and Tanushree Mitra, "Auditing E-Commerce Platforms for Algorithmically Curated Vaccine Misinformation," *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21)*, no. 186 (May 7, 2021): 1–27, https://doi.org/10.1145/3411764.3445250.

[64] Catherine Olsson, "Incident Number 10: Kronos Scheduling Algorithm Allegedly Caused Financial Issues for Starbucks Employees," Artificial Intelligence Incident Database, Responsible AI Collaborative, August 14, 2014, https://incidentdatabase.ai/cite/10/.

[65] Jodi Kantor, "Working Anything But 9 to 5," *The New York Times*, August 13, 2014, https://www.nytimes.com/interactive/2014/08/13/us/starbucks-workers-scheduling-hours.html.

[66] Michael Topper and Toshio Ferrazares, "The Unintended Consequences of Policing Technology: Evidence from ShotSpotter," Working Paper, October 29, 2024, https://michaeltopper.netlify.app/research/jmp_michael_topper.pdf.

[67] Dillon Reisman, Jason Schultz, Kate Crawford, and Meredith Whittaker, "Algorithmic Impact Assessments: A Practical Framework for Public Agency Accountability" (AI Now Institute, April 2018), https://ainowinstitute.org/publication/algorithmic-impact-assessments-report-2.

[68] Simon Mylius and Jamie Bernardi, "Tracking and Classifying Incidents of Harm from AI," MIT AI Risk Repository, https://airisk.mit.edu/ai-incident-tracker.