Towards Responsible AI: Rethinking Incident Management and Accountability in Artificial Intelligence

Abstract

As artificial intelligence (AI) systems become prevalent across various domains, the rise of AI-related incidents raises significant concerns regarding ethical implications, privacy threats, and organizational responses. This paper examines the landscape of AI incidents through the lens of the Artificial Intelligence Incident Database (AIID) and analyzes the accountability framework surrounding these incidents. The distributed responsibility inherent in the development of sociotechnical systems challenges traditional notions of assigning blame to individuals or organizations. The paper proposes a comprehensive approach to incident management, emphasizing learning, collaboration, and restorative justice. Additionally, it highlights the crucial role of ethics and privacy specialists within organizations and explores the possibility of AI product recalls in extreme cases.

Positionality and Reflexivity Statement

I consider myself a cynical person concerning technology; however, I do not perceive this as a negative personality trait. Aside from LinkedIn, I do not maintain any social media profiles and identify as a private person. Despite this, as a MIDS student and a user of AI tools, such as ChatGPT, I advocate for the advancement of technology and the positive impact that it can have on society. While I regard myself as a private individual, I recognize that others engage in social media and online communication, potentially exposing themselves to ethical or privacy harms that I avoid by abstaining from certain platforms. Additionally, due to my identity as a white, middle-class, cisgender, heterosexual, non-disabled male, I acknowledge that I cannot fully relate to certain outcomes of the incidents researched. In an effort to address this limitation, I draw on literature around intersectionality, among other perspectives, to better understand and approach diverse minority groups. In this report, my cynicism guides me toward proposing solutions that challenge the notion that technology companies solely prioritize their own interests. I aspire to envision a world where all individuals involved in the advancement of sociotechnical systems prioritize fair and equitable evolution, with financial motives taking a secondary role.

Introduction

The proliferation of artificial intelligence (AI) systems has brought about a corresponding increase in AI-related incidents, prompting a critical examination of the ethical, privacy, and accountability dimensions associated with these incidents. The United States and European Union have recently passed legislation regulating AI, which is a notable path forward towards minimizing the negative impacts that AI may have, but it is not sufficient to cover the broad scope of organizations, people, and systems involved. In an effort to further the responsible advancement of AI, this report delves into the Artificial Intelligence Incident Database (AIID) to explore the patterns, implications, and organizational responses surrounding AI incidents. Focusing on the distributed responsibility inherent in the development of sociotechnical systems, the report challenges conventional perspectives on accountability and proposes a multifaceted approach to incident management. From fostering collaboration to embracing restorative justice, the report aims to provide a holistic framework for addressing and learning from AI incidents.

Setting

As AI systems continue to proliferate across diverse domains, the occurrence of AI-related incidents has raised concerns regarding ethical implications, privacy threats, and organizational responses. According to the AIID, a public online resource that crowdsources AI incident information from its users, incidents have been on the rise in recent years, as seen in Figure 1 below.

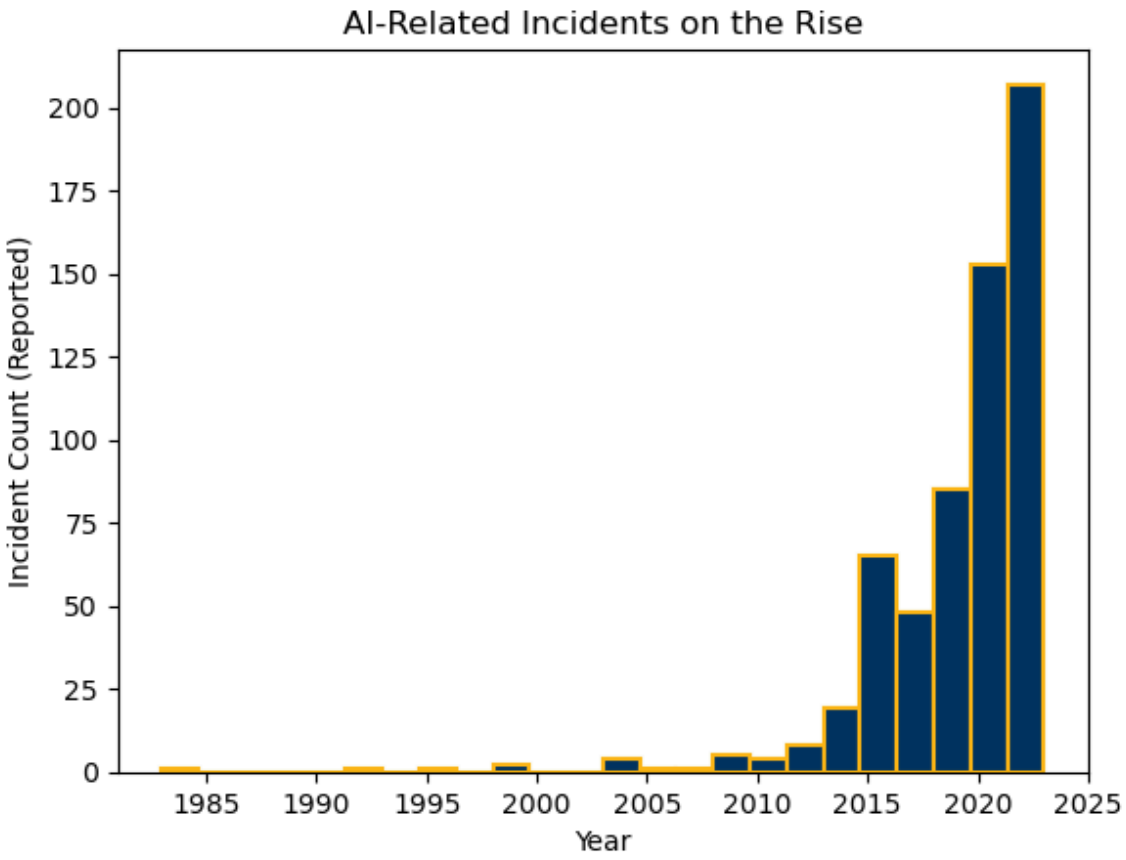

Figure 1: AI-Related Incidents on the Rise

AIID defines an incident as “an alleged harm or near harm event to people, property, or the environment where an AI system is implicated” (AIID 2023). AIID further defines “harm” to consist of multiple types, including by not limited to, physical harm, psychological harm, and harm to civil liberties. While AIID further defines “near harm” as well, some examples for both may be more difficult to label as others. However, the definition of what constitutes an incident from AIID is sufficient, and will be used throughout this report.

Incidents

Incidents reported to AIID include the following two categories, along with other detailed information:

- **Developer of AI System**: the organizations or individuals responsible for producing either the parts or the whole intelligent system implicated in the incident.
- **Harmed/Nearly Harmed Parties**: impacted classes (e.g., teachers, black people, women) or individuals.

Regarding the developer of the AI system, there are a total of 310 unique organizations, governments, universities, etc. across the 605 incidents reported to AIID to date. However, only 10 organizations account for roughly 43% of the incidents in the database.

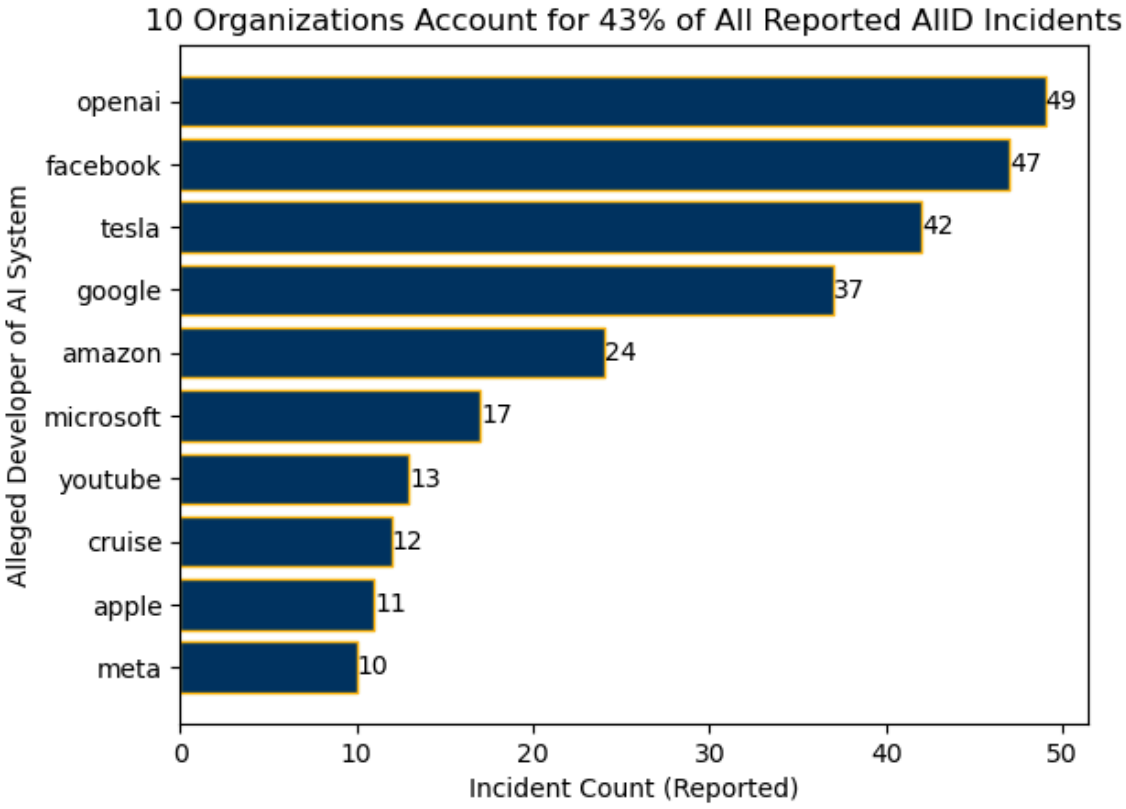

Figure 2: Top 10 Developers of AI System in AIID

While these organizations may not be surprising to see due to their prevalence in AI, these organizations do not necessarily hold themselves accountable or provide public information about any incident response or resolution to these incidents. In addition to crowdsourcing reports of incident occurrences, AIID also attempts to crowdsource the submission of incident responses. This report will use the AIID definition of incident response, where it is defined as “a public official response to an incident … from an entity (i.e., company, organization, individual) allegedly responsible for developing … the AI or AI system involved in said incident'' (AIID 2023). Unfortunately, although 605 incidents have been reported to AIID, zero incident responses have been published (two are pending review but are currently unavailable).

Trends

As mentioned previously, AIID collects information about “Harmed/Nearly Harmed Parties,” which can vary based on how the incident was submitted to the database. Below I provide some definitions on the groupings that I conducted (e.g., “platform_users_customers”) based on the data available in AIID, and Figure 3 displays the top six groups impacted by incidents in AIID.

- **platform_users_customers**: users or customers of the platform identified as the “Developer of AI System” (e.g., Facebook users, Twitter users, Amazon customers, etc.)
- **minority_groups:** keywords include, but are not limited to, women, black people, transgender, Jewish people, minorities, BIPOC, communities of color (the full list is available in the repository for this project)
- **traffic_participants:** vehicle operators (e.g., uber drivers, tesla drivers, motorists, etc.), pedestrians
- **general_public**: AIID definition used
- **minors:** children, teenagers

- **developer_of_ai_system:** developer of the platform identified as the “Developer of AI System” (e.g., Microsoft, OpenAI, Tesla, etc.)

This grouping attempts to account for intersectionality between groups (Crenshaw 1989), albeit imperfectly. For example, a “platform_user_customer” may potentially identify as a black teenager, which would also make them a member of the “minority_groups” and “minors” categories, respectively. In this case, this person would be counted three times in the data behind Figure 3. The purpose of this summary level data is to attempt to assess impacts to groups at a higher level, but individual subgroups may have different experiences and impacts that are not necessarily covered here. There are also individual names included in AIID when listing “Harmed/Nearly Harmed Parties,” but those individuals are excluded to protect their privacy, and to avoid any improper classification based on external information of the individual.

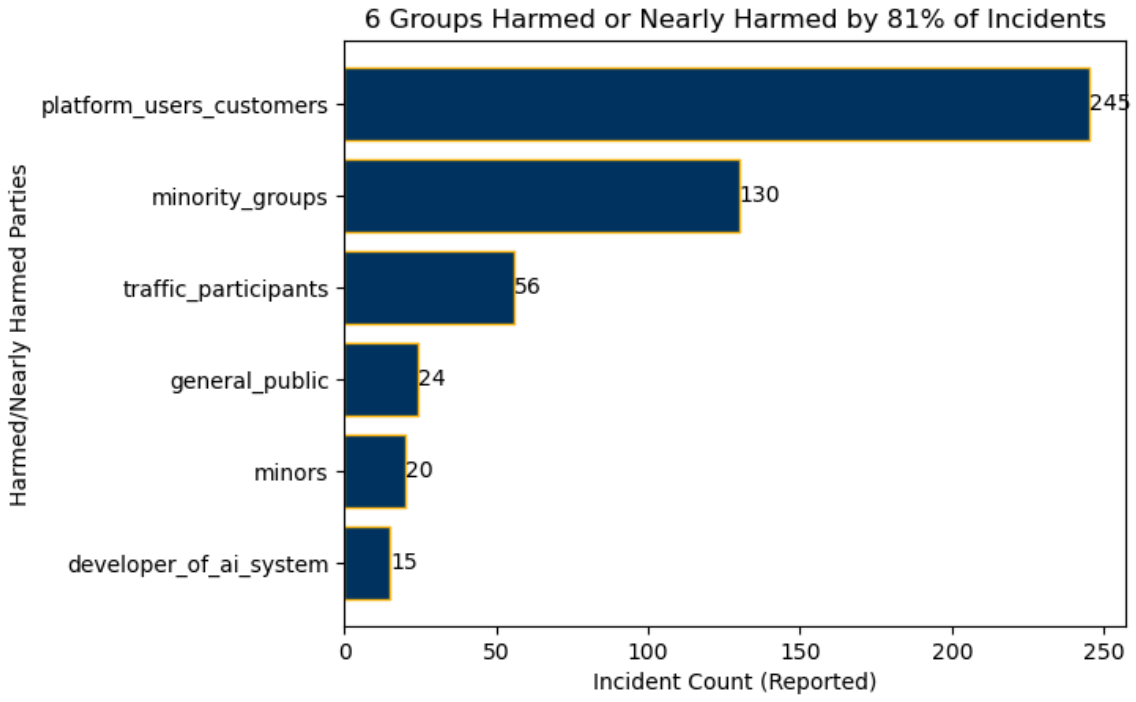

Figure 3: Top 6 Groups Harmed or Nearly Harmed in AIID

Ethical Considerations

The Belmont Report outlines three principles: respect for persons, beneficence, and justice (Department of Health, Education, and Welfare 1979). The top group harmed or nearly harmed by incidents reported to AIID, “platform_users_customers”, accounts for 40% of incidents. When compared to Figure 2, which shows that 43% of incidents are attributed to technology (“tech”) companies, this brings to mind the justice principle, which asks, “who ought to receive the benefits of research and bear its burdens?” (Department of Health, Education, and Welfare 1979). While being responsible for 43% of incidents, tech companies only account for roughly 2% of harmed or nearly harmed parties, as seen in Figure 3 (“developer_of_ai_system”). This lack of balance brings to light the beneficence principle as well, along with the following questions:

- Are tech companies being held accountable for the incidents that their systems create? It does not seem so as it relates to justice, since users are the ones bearing the brunt of the burdens.
- If tech companies are not being held accountable by direct repercussions, what other options might there be?

The second most prevalent group harmed or nearly harmed by incidents reported to AIID, “minority_groups”, account for 22% of incidents. This does not necessarily imply that non-minority groups account for the remaining 78% of incidents, but rather highlights the impact that AI incidents have had on minority groups. Similarly to “platform_users_customers”, concerns arise around beneficence and the harm that these incidents have resulted in for historically marginalized communities.

Privacy Considerations

While only accounting for being harmed or nearly harmed in 3% of incidents, privacy around “minors” is

a concern with the incidents reported to AIID. Additionally, the example incident below impacts both minors and minority groups.

- **Title**: Skating Rink's Facial Recognition Cameras Misidentified Black Teenager as Banned Troublemaker
- **Description:** A Black teenager living in Livonia, Michigan, was incorrectly stopped from entering a roller skating rink after its facial-recognition cameras misidentified her as another person who had been previously banned for starting a skirmish with other skaters (AIID 2023).

The use of facial recognition at a skating rink to ban anyone, let alone teenagers, seems unreasonable, and brings to light concerns around contextual integrity (Nissenbaum 2011). If a skating rink in Livonia, Michigan is using facial recognition, are customers aware? Are there other similar establishments (bowling alleys, arcades, etc.) using the same technology? If so, are customers aware?

Solove's Taxonomy takes a holistic approach in providing an end-to-end privacy framework that can be applied universally. The taxonomy is broken into four activities that affect privacy; Information Collection, Information Processing, Information Dissemination, and Invasions (Solove 2006). In this example, particularly, information about individuals, including minors (i.e., their facial profile) is being surveilled, likely without informed consent. This also raises explicit concerns around the information collection and invasion activities described in Solove's Taxonomy.

Accountability

From large tech companies, to local skating rinks, AI is being used by organizations to run their business. As mentioned previously, of the 605 incidents in AIID, zero of them have published incident responses in

AIID. However, when incidents arise, who is held accountable? What does accountability look like in the age of AI?

Organizational Responses

Of the 605 incidents in AIID, roughly 5% of them resulted in the death of one or more individuals, either directly or indirectly. For example, one incident describes how “a Tesla Model S driver on Autopilot mode reportedly went through a red light and crashed into a Honda Civic, killing two people” (AIID 2023). This is an example of a direct result of AI. An indirect example in AIID is when “TikTok’s recommendation algorithm was alleged in a lawsuit to have intentionally and repeatedly pushed videos of the “blackout” challenge onto children’s feeds, incentivizing their participation which ultimately resulted in the death of two young girls” (AIID 2023). Internet searches for both of these incidents did not result in finding an incident report from Tesla or TikTok, and neither exist in AIID.

Perhaps the most notable incident in AIID, however, is the Boeing 737 MAX incident that “crashed into the sea, killing 189 people, after faulty sensor data caused an automated maneuvering system to repeatedly push the plane's nose downward” (AIID 2023). This was the second crash related to this airplane model, and both incidents received a lot of publicity (Langweiesche 2019, Hawkins 2021, Kesslen 2020), and resulted in the 737 MAX being grounded for over two years (German 2021).

The airline industry, however, is no stranger to incidents and incident management. Aero Inside, for instance, has over 18,000 articles available that cover “aviation incidents, accidents and plane crashes as well as news and reports” (Aero Inside 2023), compared to the 605 incidents related to AI in AIID. Therefore, it is not necessarily surprising, especially given the relatively high publicity due to the amount of fatalities associated with it, that a lengthy incident response was released regarding the 737 MAX

incidents. Unfortunately, Boeing did not release this report — the House Committee on Transportation and Infrastructure ("the committee") did (Majority Staff of the Committee on Transportation and Infrastructure 2020). In this report, the committee presents the following conclusions:

> *"The report reveals several unmistakable facts. The MAX crashes were not the result of a singular failure, technical mistake, or mismanaged event. They were the horrific culmination of a series of faulty technical assumptions by Boeing's engineers, a lack of transparency on the part of Boeing's management, and grossly insufficient oversight by the FAA—the pernicious result of regulatory capture on the part of the FAA with respect to its responsibilities to perform robust oversight of Boeing and to ensure the safety of the flying public. The facts laid out in this report document a disturbing pattern of technical miscalculations and troubling management misjudgments made by Boeing. It also illuminates numerous oversight lapses and accountability gaps by the FAA that played a significant role in the 737 MAX crashes."*

This report puts two agencies at fault: Boeing and the FAA. However, what did Boeing find? What did the FAA find? The lack of insight from these organizations is concerning, and if catastrophic incidents like these do not result in public incident reports from the organizations involved, what precedent is being set moving forward? Particularly, in the airline industry, an industry with a long history of safety-related incidents.

The committee's conclusion also draws parallels to the concept of *sociotechnical systems*, in which both humans and machines are necessary to make any technology work as intended (Selbst et al. 2019). Blame in the committee's report is placed across "faulty technical assumptions," a "lack of transparency [by] management," FAA's "insufficient oversight," and "technical miscalculations," among other details.

While the committee reached a conclusion on who to blame (i.e., Boeing and FAA), recommendations failed to address responsibility within each organization. As an outside organization, the committee lacks the day-to-day experience of engineers, managers, regulators, etc. that may lead to incidents. However, touching on the concept of sociotechnical systems, the committee indirectly points to the problem of distributed responsibility that exists within organizations producing these systems.

Distributed Responsibility

While the Boeing 737 MAX incident served as an example of dangerous outcomes of AI and the lack of organizational responses, the following will discuss responsibility in more general terms as it relates to individuals and technology that contribute to sociotechnical systems.

Individual Contributors

In any organization that develops AI, the individual contributors consist of anyone on the development team that directly contributes to the construction and deployment of the solution. This can be Data Scientists, Machine Learning Engineers, Test Engineers, Developers, etc., all of which are not individually responsible for the success of the product. In most cases, organizations structure their development teams to limit the cognitive load of the individual (Skelton and Pais 2019), so it may not be reasonable to ask a computer scientist to also be an expert in ethics, for example. At a higher-level, it may not be reasonable to ask a team of engineers to also consider the broad scope of ethics in their solution development. Additionally, due to bias in tech, these teams of individual contributors "lack racial, geographic, class, and gender diversity, with Black and Latinx technologists especially poorly represented" (Moss and Metcalf 2020).

Ethics Owners

Ethics owners within organizations can take many forms, as outlined by Moss and Metcalf (2020). They may be explicit roles (e.g., Ethics Officer, Privacy Officer) within an organization, or they may be individual contributors within teams who deploy AI systems. As it relates to AI incidents, ethics can serve fundamentally as a risk management process (Moss and Metcalf 2020). However, asking an ethics owner, often with conflicting priorities between organizational duties and ethical responsibilities, to be the

leading voice for ethics within an organization, is not a reasonable expectation due to their limited direct influence on a product.

Mid-Level Management/Leadership

One level below the executives lies the mid-level management and leaders within an organization. Typically, these individuals have responsibilities for the day-to-day operations of their teams, and the insight into organizational strategy from the executives. In theory, these individuals can serve as a catalyst for communication between levels above them and below them. As it relates to ethics, they face similar challenges of ethics owners. On one hand, they are indirectly responsible for the performance of their direct reports, and the products that they deliver. On the other hand, they can face pressure from their superiors on delivering products quickly to meet financial incentives for both themselves and for the organization at large. While important pieces of the organization, this role typically lacks the ability to implement ethical or privacy practices in their AI system, as well as the higher-level authority to dictate procedures that may ultimately impact the bottom line of the organization.

Senior Leadership/Executives

At the highest hierarchical level of the typical organizational structure lies the senior leadership and executive teams. While these people may be the most equipped to handle wide-reaching changes across an organization as it relates to ethical or privacy policies, they likely lack the subject matter expertise to do so individually. For instance, of the 52 “most influential tech founders and CEOs” in 2020, only one of them has a background in Philosophy (Leskin and Vega 2020). The majority of the remaining individuals have math, computer science, or other technical degrees. While we may expect some self-study in the field of ethics, these individuals have been trained elsewhere, so laying the expectation for ethical and privacy practices strictly with them may not be the optimal solution either. Additionally, in a capitalist

society, these individuals are often more rewarded for their financial performance than the societal or ethical impact their products make.

The AI System

Traditionally, the view of responsibility for incidents lies with the human pieces of an organization. However, as AI continues towards Artificial General Intelligence (AGI), one must consider the role advanced computers may have within the outcomes of an organization, incidents included. Anna Strasser argues that it "seems plausible that moral responsibility can be distributed between artificial and human agents" (Strasser 2021).

Regulators

Safety-critical industries (e.g., airline, automobile, etc.) have a long history of established regulatory bodies providing oversight for the benefit of the safety of the general population. As concerns of AI's impact on society have increased in recent years, the United States (US) and European Union (EU) recently passed legislation in an attempt to proactively address the ethical and privacy concerns presented by AI. The Executive Order on Safe, Secure, and Trustworthy Artificial Intelligence (Biden 2023) and the EU Artificial Intelligence Act (Council of the EU 2023) have varying degrees of requirements, but are positioned to play a pivotal role in the future of AI ethics, privacy, and safety, and any subsequent incidents that will arise. However, the ultimate decision for the release of AI systems does not entirely lay with regulators, and likely will continue to be owned by the organizations building these systems, so relying solely on regulators is an incomplete approach.

Users

As seen above in the Livonia, Michigan roller skating rink incident, users of AI systems can contribute to incidents as well. This applies more broadly as well, including the over 180 million users of ChatGPT (Nerdynav 2023) and other Generative AI products. In a case regarding Large Language Models (LLMs) specifically, a US lawyer faced sanctions when ChatGPT provided incorrect information that was subsequently used in court (Weiser and Schweber 2023). Both of these examples show how users can be held responsible for the output of the AI that they use, and the general public should learn from these instances to take responsibility.

Related Work on Accountability

Due to the many parties involved in the development and use of AI, responsibility cannot be attributed to a single person or party. As described above, multiple touchpoints result in a distributed responsibility framework. However, when incidents arise, traditional human tendency is the need to be able to point the finger at one person or a few people, in order to provide a sense of relief knowing that bad outcomes are otherwise unavoidable if we are simply able to address single failure points. However, in complex, distributed, sociotechnical systems, we should look for complex explanations, not simple ones, such as single failure points.

The Field Guide to Understanding “Human Error”

In his landmark book in the field of human factors engineering, Sidney Dekker outlines two opposing viewpoints of incident responses, the “Old View” and the “New View” (Dekker 2014). Table 1 below shows a brief overview of these opposing viewpoints.

Table 1: Old View vs. New View (Dekker 2014)

Old View	New View
"Human error" is the *cause* of trouble	What we call "human error" is a symptom of deeper trouble
"Human error" is a separate category of behavior, to be feared and fought	"Human error" is an attribution, a judgment that *we* make after the fact
"Human error" is the target; people's behavior is the problem we need to control	Behavior is systematically connected to features of people's tools, tasks and operating environment
"Human error" is something to declare war on. People need to practice perfection	"Human error" is information about how people have learned to cope (successfully or not) with complexities and contradictions of real work
"Human error" is a simple problem. Once all systems are in place, just get people to pay attention and comply	A "human error" problem is at least as complex as the organization that helps create it
With tighter procedures, compliance, technology, and supervision, we can reduce the "human error" problem	With better understanding of the messy details of people's daily work, we can find ways to make it better for them
We can, and must, achieve zero errors, zero injuries, zero accidents	We can, and must, enhance the resilience of our people and organization

The Old View relies on the benefit of hindsight, where people can see the outcome of decisions after the fact, unlike the people or systems involved that made decisions prior to knowing the outcome. When we look from the outside after lengthy analysis, most errors may seem preventable, which might prompt organizations or governments to do the following in response:

- Fire employees;
- Implement more rules, governance, regulations;
- Tell people to "be more careful";
- Replace "unreliable people" with technology (Dekker 2014)

The Boeing 737 MAX case provides an example of the Old View in action. After 18 months of review and lengthy documentation, one of the committee's conclusions was that the incident occurred, at least partially, due to "faulty technical assumptions by Boeing's engineers, a lack of transparency on the part of Boeing's management, and grossly insufficient oversight by the FAA" (Majority Staff of the Committee on Transportation and Infrastructure 2020). "Human Error", disguised in the committee's report as bad engineers, lazy managers, and negligent regulators, is an insufficient response to a complex system of distributed responsibility that ultimately led to the death of hundreds of people.

The committee's report does not specifically address the automated system. However, what happens when the technology is at least partially responsible for an incident? How can we hold the system accountable? Is the system full of "bad data", or, perhaps, is it a "bad algorithm"?

Where Fairness Fails

Similar to Dekker, Anna Lauren Hoffman points out that "the first tendency centers on the law's concern with neutralizing inappropriate conduct on the part of individual perpetrators", what she calls "the 'bad actor' frame" (Hoffman 2019). Applying a similar mindset of Dekker's to AI and documented social harms, Hoffman points out that the traditional view of racism, for example, is that it is an individual and personal trait, one that must be addressed at an individual level. In turn, Hoffman notes, this often results in reducing a system's shortcomings to the biases of its imperfect human designers.

When individual bias is determined to be a cause of an unethical outcome, unconscious bias training programs are often introduced, drawing parallels to implementing regulations, or simply telling individuals to "stop being racist." However, algorithms that produce these unethical or privacy-related incidents are shaped by the cultural context in which they reside. Therefore, while data and algorithms

(the AI system) contribute to outcomes, they are also intimately influenced by the particular meanings of societal definitions, and often simply reinforce certain biases rather than inventing new ones.

Similar to Dekker's New View, Hoffman points out that "efforts to isolate 'bad data', 'bad actors', or localized biases of designers and engineers are limited in their ability to address broad social and systemic problems" (Hoffman 2019). The problems that exist in unethical and privacy pervasive outcomes and incidents must instead be addressed with a new view, with sustained and iterative attention to system failures.

Proposed Path Forward

Due to the distributed responsibility that exists in the development of sociotechnical systems, attributing AI incidents simply to "human error", "bad data", or "bad algorithms" risk conclusions similar to Dekker's "Old View", which may result with reactionary solutions that fail to analyze the underlying logic and processes that humans and systems use to produce unethical or privacy-pervasive outcomes. Similarly, relying exclusively on government oversight, such as the newly established Executive Order on Safe, Secure, and Trustworthy Artificial Intelligence (Biden 2023) and the EU Artificial Intelligence Act (Council of the EU 2023), eliminates the involvement of all other actors in the development of these systems.

To supplement traditional views of accountability and the involvement of regulators, I propose the following to aid in managing incidents and incident response in the context of AI moving forward, in an effort to improve learning across the industry, limit bias, and create a modern accountability framework.

Incident Management Processes

The AIID has built a strong foundation to expand upon, with publicly available data around AI incidents, as well as the intention of creating a knowledge repository of the outcome of incident responses. The data is limited, however, to crowdsourced reporting from third parties.

Using the AIID as a guide, I propose an industry-wide AI incident learning repository, where organizations submit incident reports to address what was found in an effort to spread knowledge so that others can learn from incidents. This would, of course, include technical remedies, but also must include ethical, privacy, and bias findings for others to leverage when building new systems. While this approach may seem implausible, it is not dissimilar from the Executive Order on Safe, Secure, and Trustworthy Artificial Intelligence, where safety test results are required to be shared with the US government (Biden 2023). Instead, this expands the sharing of this information to the general public, which provides a mechanism for all members of the sociotechnical system to learn and use the system responsibly.

One potential way to contribute to this as well is for organizations to make their systems open-sourced so that more technical individuals can implement solutions that may have already been made to address biased, unethical, or privacy-pervasive outcomes of other systems. The EU Artificial Intelligence Act also incentivizes open-sourced solutions, as “providers of free and open-source models are exempted from most of these obligations” (European Commission 2023).

Fostering collaboration among organizations, research institutions, and regulatory bodies could potentially shift the focus from retributive justice to restorative justice and forward-looking accountability.

Restorative Justice and Forward-looking Accountability

An effective incident management framework must result in outcomes that encourage individuals and organizations to participate. Regulations, sanctions, fines, etc. risk organizations pulling back from the public sphere and retreating to private development to avoid facing perceived interference from regulators or elsewhere. Thus, a focus on restorative justice and forward-looking accountability could aid in adoption of knowledge sharing if the ramifications of incidents focus on the benefit of society as opposed to the punishment of those deemed responsible.

Retributive justice suggests that if the outcome of an incident hurts an individual or a group of individuals, then the response should hurt those found responsible. Restorative justice, on the other hand, suggests that if the outcome of an incident hurts an individual or a group of individuals, then the response should heal those impacted. This should benefit both the organizations that developed the AI system as well as those impacted, as organizations could replace the payment of fines with the investment in impacted communities and additional incident management processes.

Related to retributive justice, backward-looking accountability means blaming organizations for past incidents, and "holding them accountable" for events that already happened. Dekker describes that "this does not work … and it only motivates others to be more careful in reporting and disclosure" (Dekker 2014). If this is the case, backward-looking accountability would only dissuade organizations and individuals from sharing incident reports. Conversely, forward-looking accountability views people as a solution to harness, rather than a problem to control (Dekker 2014). Similar to restorative justice, forward-looking accountability can help people and organizations focus on improvements to AI systems and communities served and impacted.

Ethics and Privacy Owners

While an incident management process could prove successful in the public sharing of incidents and AI learning across industries, individuals within organizations that specialize in ethics and privacy are also essential moving forward. As discussed previously, relying on individuals in organizations who specialize elsewhere (e.g., engineering, management, etc.) only impacts the cognitive load of these individuals, whose training, background, and focus is often elsewhere.

As Moss and Metcalf describe in their work, “ethics can’t live on paper … they live inside people” (Moss and Metcalf 2020). Ignoring the impact of individuals who specialize in this field risks resulting in incident reports concluding in governing procedures for individual contributors to follow, for example. While ethics owners described red teams as useful for understanding unanticipated harms that might occur when releasing a product, ethics owners still rely largely on individual users bringing harms to their attention through customer service portals (Moss and Metcalf 2020). These findings of Moss and Metcalf further solidify the need for effective incident management tools, since no one involved in the sociotechnical system can predict every harmful outcome ahead of time.

Recalls

Lastly, a recommendation that should only be saved for extreme measures is to recall AI products until certain safety criteria have been met. While this may not be unfamiliar to AI deployed to heavily regulated industries like airline and automobile (Langweiesche 2019, Hawkins 2021, Kesslen 2020, Zalubowski 2023), tech companies could also benefit from recalls when products are deemed harmful. When AI systems could directly impact the safety of human lives, such as autonomous vehicles (Zalubowski 2023), the concept of recalls is fairly straightforward. However, when the impact of AI does

not result in physical harm, but rather the asymmetrical harm to different populations of people, this is less apparent. This would require the decision of an organization to voluntarily recall systems, or regulators to enforce such recalls. However, further work is needed in this space to fully understand how this could be leveraged as a solution to address AI biases and safety concerns.

Conclusion

The complex landscape of AI incidents necessitates a paradigm shift in how we perceive and respond to challenges in the development and deployment of AI systems. The proposed approach, emphasizing incident management processes, restorative justice, and forward-looking accountability, aims to foster a culture of learning and collaboration across industries. The crucial role of ethics and privacy specialists within organizations cannot be overstated, as they play a pivotal role in ensuring responsible AI development. Moreover, the contemplation of AI product recalls, while a drastic measure, underscores the importance of prioritizing safety, privacy, and ethical considerations. As the AI field continues to evolve, a proactive and adaptive approach to incident management will be indispensable in shaping a responsible and accountable AI ecosystem.

References

Aero Inside. (2023). *Aviation incidents, accidents and airplane crashes* [dataset]. https://www.aeroinside.com/

AIID. (2023). *Artificial Intelligence Incident Database (AIID)* [dataset]. https://incidentdatabase.ai/

Andrew D. Selbst, Danah Boyd, Sorelle A. Friedler, Suresh Venkatasubramanian, & Janet Vertesi. (2019). *Fairness and Abstraction in Sociotechnical Systems*. FAT* '19: Proceedings of the Conference on Fairness, Accountability, and Transparency. https://dl.acm.org/doi/10.1145/3287560.3287598

Andrew J. Hawkins. (2021, April 9). Boeing 737 Max airplane crashes: All of the news, updates, and analyses. *The Verge*. https://www.theverge.com/2019/3/21/18274868/boeing-737-max-airplane-crash-updates-highlights

Anna Lauren Hoffman. (2019). *Where fairness fails: Data, algorithms, and the limits of antidiscrimination discourse*. Information, Communication & Society. https://www.tandfonline.com/doi/full/10.1080/1369118X.2019.1573912

Anna Strasser. (2021). *Distributed responsibility in human–machine interactions*. https://link.springer.com/article/10.1007/s43681-021-00109-5

Ben Kesslen. (2020, September 16). 737 Max crashes that killed 346 were "horrific culmination" of failures by Boeing and FAA, House report says. *NBC News*. https://www.nbcnews.com/news/us-news/737-max-crashes-killed-346-were-horrific-culmination-failures-boeing-n1240192

Benjamin Weiser & Nate Schweber. (2023, June 8). The ChatGPT Lawyer Explains Himself. *New York Times*. https://www.nytimes.com/2023/06/08/nyregion/lawyer-chatgpt-sanctions.html

Artificial intelligence act: Council and Parliament strike a deal on the first rules for AI in the world, (2023). https://www.consilium.europa.eu/en/press/press-releases/2023/12/09/artificial-intelligence-act-council-and-parliament-strike-a-deal-on-the-first-worldwide-rules-for-ai/

Daniel Solove. (2006). The Taxonomy of Privacy. *University of Pennsylvania Law Review*, *153*(3). https://papers.ssrn.com/sol3/papers.cfm?abstract_id=667622

David Zalubowski. (2023, December 13). Tesla recalls over 2 million vehicles to fix defective Autopilot monitoring system. *NPR*. https://www.npr.org/2023/12/13/1219008292/tesla-recall-2-million-autopilot

Department of Health, Education, and Welfare. (1979). *The Belmont Report* (pp. 1–10). https://www.hhs.gov/ohrp/regulations-and-policy/belmont-report/read-the-belmont-report/index.html

Emanuel Moss & Jacob Metcalf. (2020). *Ethics owners: A new model of organizational responsibility in data-driven technology companies*. Data & Society. https://datasociety.net/library/ethics-owners/

European Commission. (2023). *Artificial Intelligence – Questions and Answers*. https://ec.europa.eu/commission/presscorner/detail/en/QANDA_21_1683?mkt_tok=MTM4LU

VaTS0wNDIAAAGQA3kdv-6WuPdUucj7CZZfaAP8eR-uVHSogP_l-G-oc0VGexVVNou8TdHBdymo7DjQbo5NGWsNLLcnyN-r3wZis5ydiglxNDBg92Ht5HGabe75

Helen Nissenbaum. (2011). *A Contextual Approach to Privacy Online*.

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

Kent German. (2021, June 19). 2 years after being grounded, the Boeing 737 Max is flying again. *CNET*. https://www.cnet.com/tech/tech-industry/boeing-737-max-8-all-about-the-aircraft-flight-ban-and-investigations/

Kimberle Crenshaw. (1989). *Demarginalizing the Intersection of Race and Sex: A Black Feminist Critique of Antidiscrimination Doctrine, Feminist Theory, and Antiracist Politics*. University of Chicago Legal Forum. https://chicagounbound.uchicago.edu/uclf/vol1989/iss1/8/?utm_source=chicagounbound.uchicago.edu%2Fuclf%2Fvol1989%2Fiss1%2F8&utm_medium=PDF&utm_campaign=PDFCoverPages

Majority Staff of the Committee on Transportation and Infrastructure. (2020). *THE DESIGN, DEVELOPMENT & CERTIFICATION OF THE BOEING 737 MAX*. https://democrats-transportation.house.gov/imo/media/doc/2020.09.15%20FINAL%20737%20MAX%20Report%20for%20Public%20Release.pdf

Matthew Skelton & Manuel Pais. (2019). *Team Topologies* (1–1). https://teamtopologies.com/

Michael Anderson. (2023, September 12). Restorative Justice Vs. Retributive Justice: What's The Difference? *Callforjustice.Org*. https://www.callforjustice.org/restorative-vs-retributive-justice/

Nerdynav. (2023, December 6). *107 Up-to-Date ChatGPT Statistics & User Numbers [Dec 2023]*. https://nerdynav.com/chatgpt-statistics/

Paige Leskin & Nick Vega. (2020, February 13). From Elon Musk to Tim Cook, here's where the world's most influential tech founders and CEOs went to college—And what they studied. *Business Insider*. https://www.businessinsider.com/college-degrees-and-majors-of-famous-tech-ceos?op=1

Sidney Dekker. (2014). *The Field Guide to Understanding "Human Error"* (Third). https://www.routledge.com/The-Field-Guide-to-Understanding-Human-Error/Dekker/p/book/9781472439055

William Langewiesche. (2019, September 18). What Really Brought Down the Boeing 737 Max? *New York Times*. https://www.nytimes.com/2019/09/18/magazine/boeing-737-max-crashes.html