RESEARCH-ARTICLE

From Incidents to Insights: Patterns of Responsibility following AI Harms

ISABEL RICHARDS, University of Cambridge, Cambridge, Cambridgeshire, U.K.

CLAIRE BENN, University of Cambridge, Cambridge, Cambridgeshire, U.K.

MIRI ZILKA, University of Cambridge, Cambridge, Cambridgeshire, U.K.

Open Access Support provided by:

University of Cambridge

Published: 05 November 2025

Citation in BibTeX format

EAAMO '25: Equity and Access in Algorithms, Mechanisms, and Optimization
November 5 - 7, 2025
Pittsburgh, USA

Conference Sponsors:
SIGECOM
SIGAI

https://doi.org/10.1145/3757887.3763018
ISBN: 9798400721403

From Incidents to Insights: Patterns of Responsibility following AI Harms

Isabel Richards
University of Cambridge
Cambridge, United Kingdom
irichards0742@gmail.com

Claire Benn
University of Cambridge
Cambridge, United Kingdom
cmab3@cam.ac.uk

Miri Zilka
University of Cambridge
Cambridge, United Kingdom
mz477@cam.ac.uk

Abstract

The AI Incident Database (AIID) was inspired by aviation safety databases, which enable collective learning from failures to prevent future incidents. The database documents hundreds of AI Harm events, collected manually from the news and media. However, recent criticism highlights that the AIID's reliance on media reporting limits its utility for learning about implementation failures. In this paper, we argue that the AIID's value goes beyond technically-focused learning, since the dataset provides highly valuable insights into the reactions of developers, deployers, victims, wider society, and law-makers *after* AI failures. Through a three-tier mixed-methods analysis of 962 incidents and 4,743 related reports from the AIID, we examine patterns across incidents, focusing on cases with subsequent public responses tagged in the database. We identify common 'typical' incidents found in the AIID, from Tesla crashes to deepfake scams. Focusing on this interplay between relevant parties, we uncover patterns in accountability and social expectations of responsibility. We find that the presence of identifiable responsible parties does not necessarily lead to increased accountability. The likelihood of a response and what it amounts to depends highly on context, including who built the technology, who was harmed, and to what extent. Controversy-rich incidents provide valuable data about societal reactions, including insights around social expectations regarding responses. Equally informative are cases where expected controversy is notably absent. This work shows that the AIID's value lies not just in preventing technical failures, but in documenting and making visible patterns of harms and of institutional response and social learning around AI incidents. These patterns offer crucial insights for understanding how society adapts to and governs emerging AI technologies.

ACM Reference Format:
Isabel Richards, Claire Benn, and Miri Zilka. 2025. From Incidents to Insights: Patterns of Responsibility following AI Harms. In *Equity and Access in Algorithms, Mechanisms, and Optimization (EAAMO '25), November 05–07, 2025, Pittsburgh, PA, USA.* ACM, New York, NY, USA, 19 pages. https://doi.org/10.1145/3757887.3763018

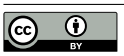
1 Introduction

In an era of unprecedented AI deployment, failure is inevitable – the key is to learn from it. AI systems, from machine learning classifiers to large language models (LLMs), are being rapidly deployed across many industries to tackle a range of perceived problems from elderly care [57] to traffic optimisation [9]. This accelerating adoption is powered by a wealth of investment [36] and an international strategic focus [4]. However, the potential for the high-stakes disruption of societies is vast, including impact on education [20], social cohesion [23], and even on concepts such as mourning [26].

The question of *how to learn from failure* is important yet complex. The aviation industry takes an approach of collective learning from failure. Aviation incident databases, such as those hosted by the US National Transportation Safety Board, provide a detailed public record of incidents through extensive investigations carried out by specialist industry teams. Commercial aviation has seen a significant increase in safety over decades, with much of this progress attributed to the use of these shared databases [34]. The aviation industry's approach to improving safety through public databases suggests a straightforward path to improving safety: document incidents so the entire industry can learn from others' mistakes and implement preventive solutions.

Inspired by the historical success of aviation safety, incident databases for *AI* systems have emerged as community-led efforts to document and learn from failures [34, 41]. These efforts share the same vision – failure must be recorded if it is to be learnt from [34]. The approach also echoes aviation in focusing on the sphere of influence of industry practitioners: to enable developers to learn from others' mistakes and implement preventive solutions. While highly valuable, this vision is complex to realise in practice. AI incident databases, in contrast to aviation, bridge incidents across many industries, and lack the investigatory infrastructure that aviation enjoys. Another challenge is that many AI-related harms are not solely technical in nature, and the diversity of harms themselves is broader. These complexities shape the consistently evolving documentation practices of what we can consider evidence of harm.

To create a public record of AI incidents, the databases rely on submissions of existing publicly available information about incidents, principally in the form of trade and news media (media reports). Recent research [15, 28, 47, 54] highlights that media reports lack sufficient information for developers to avoid the same implementation mistakes. The AI Incident Database (AIID) – one of the largest of these efforts – has accumulated extensive documentation of AI failures, from patrol robots warding off homeless people (Incident 261, AIID) to chatbot helplines advising those with eating disorders to diet (Incident 545, AIID). In addition to collecting information on AI harm incidents, the AIID began tagging responses to incidents. This is a welcome effort to supplement details on incidents with disclosures from developers and deployers – a model adopted in cybersecurity. Interestingly, the responses tagged recorded only a few official responses, instead surfacing a richer set of social responses.



EAAMO '25, Pittsburgh, PA, USA

ACM ISBN 979-8-4007-2140-3/25/11
https://doi.org/10.1145/3757887.3763018

In this paper, we highlight alternative forms of learning from AI incident databases that take advantage of this resource. Instead of focusing on the media reports as a limited source of information about the details of an incident, we demonstrate their value as a rich source of information about how different societal actors respond. We applied a three-tier approach to learn from the societal responses to incidents embedded in database entries: First, we quantitatively analysed all 962 incidents and their 4,743 associated reports based on the actors involved – developers, deployers and harmed groups. Next, we used quantitative and qualitative methods to examine the 48 incidents and 163 reports tagged as responses, i.e., with acknowledgement from developers, deployers, or other societal actors. Finally, we combined these perspectives to systematically study sub-categories of incidents, revealing patterns and key factors driving substantive responses – and lack thereof – in prominent recurring incidents.

While the statistics of the database are heavily impacted by sampling bias, the methodology developed allows us to meaningfully learn from the incidents that are documented, by comparing similar incidents that did or did not receive responses, and the contextual factors that may have contributed. We find distinctive patterns in how AI incidents unfold and are addressed. While these patterns are not necessarily representative of all AI incidents due to the limitations of the database, they provide valuable insights into understanding effective societal and legal mechanisms for accountability in varying contexts:

(1) **The identifiability of responsible parties.** Contrary to what might be expected, we show that having identifiable responsible parties does not correlate with better accountability. Rather, many incidents involving major technology companies as both developers and deployers generated few-to-no formal responses. In contrast, incidents with no known developers and deployers, mostly involving deepfakes, show evidence of stimulating the greatest demand for response and legislative action.
(2) **The status of harmed parties.** When an acknowledgement of harms does occur, tentative evidence suggests that in the absence of legislative support, organisational victims tend to receive more substantive responses (e.g., formal investigation vs. blog post) than individual users, suggesting institutional accountability is shaped more by social and economic factors than technical ones.
(3) **The level of public outcry in response to an incident.** We demonstrate that the AIID provides two types of valuable insights about how society negotiates acceptable boundaries for AI deployment: first, from incidents that sparked highly publicised controversies, and second, from incidents that, unexpectedly, gathered no responses.

Our analysis reveals that the AIID has evolved beyond its original, aviation-inspired vision. While it may not sufficiently facilitate direct technical learning for developers and deployers, the database serves as a highly valuable resource – by documenting the AI harms landscape and how different actors respond when things go wrong, we can start to map where socio-legal accountability measures are effective and where there is much left to be desired.

2 Background and Related Work

2.1 AI Incident Databases

Since 2018, several initiatives have emerged to document AI-related incidents [3, 41, 54]. These initiatives catalogue cases where AI systems have caused harm or created tangible risk of harm in real-world applications. Among these, two databases have gained prominence and are actively maintained: the AI Incident Database (AIID) and the AI, Algorithmic, and Automation Incidents and Controversies (AIAAIC). Both databases contain hundreds of incidents from publicly available reports, primarily from news and trade media. This format enables public submission of incidents for editorial review while maintaining free data accessibility. Other incident databases efforts include MIT's *AI Risk Repository*[1], WIRED's *Artificial Intelligence Database*[2], and Algorithm Watch's *Tracing the Tracers*[3]. In this work we focus exclusively on the AIID because the data is most suitably structured for analysis, but is otherwise representative of the core data in the databases.

The AI Incident Database, initiated by Sean McGregor with sponsorship from *Partnership on AI* aims to "discover and learn from the mistakes of the past" [34]. The database targets corporate product managers, risk officers, and engineers as well as serving a wider ecology of users (e.g., researchers, policymakers, public interest advocates, academics). It provides consistently catalogued incidents, accessible through a web-based interface [1]. The aviation industry explicitly inspired the goal of the database: to achieve a significant increase in safety, facilitated by incident reporting (McGregor, 2021). However, while it inherits some of the focus on avoiding failure – both due to technical issues and human decision-making as the cause of failure – it has evolved to include wider socio-technical information.

2.1.1 Learning from implementation failure and its limitations. Several studies have demonstrated how subsets of incidents from the database can be used to avoid implementation failure in specific technical domains. For instance, researchers have examined failures due to software bugs [24], deployed model monitoring [45], and threat assessment [52]. These studies typically use the database to develop or validate taxonomies for categorising system failures which can be applied to new systems to pre-empt issues. This learning is narrow, however, relating to specific highly technical fields and a small range of available incidents where there is sufficient technical information available.

One 2022 study [30] – analysing a range of reports, including media reports, about an autonomous vehicle crash in which the first pedestrian was killed – provides an example of broad and substantive learning from publicly available reporting following an AI incident (although it does not specifically use the database as the source of these reports). Macrae [30] distilled twenty-four detailed sources of risk associated with 'structural, organisational, technological, epistemic and cultural' considerations that can inform future implementations. This incident was unusual, however. As a transport accident, it benefited from publicly available investigatory reports from regulatory bodies that uncovered significant implementation details. These reports are absent for most incidents in the database.

[1] https://airisk.mit.edu/ai-incident-tracker
[2] https://www.wired.com/category/artificial-intelligence/
[3] https://algorithmwatch.org/en/tracing-the-tracers/

Recent literature has critiqued the database's effectiveness in providing implementation failure insights more broadly, arguing that media reports generally lack desirable technical detail [15, 28, 47, 54]. The AIID team appears to be trying to address this issue, firstly, through efforts to engage developers and deployers of implicated AI systems to submit *response reports* which would lend detailed implementation insight [46]; and secondly, Pittaras and McGregor [39] suggests that they are developing a methodology for leveraging expert labelling from the AI safety community to provide speculative intuition into what might have gone wrong technically. However, to date, these have had limited impact: since the 'response' initiative began, only 5.8% of new incidents have responses, and there are no incidents with expert speculative labelling. However, this is, at least in part, due to limited editorial bandwidth and available resources. This paper focuses on what can be learnt from the AIID as-is, without additional technical detail.

2.1.2 Learning from the database in a way that goes beyond the implementation failure. In a handful of cases, incident data has been used for learning beyond implementation failure. Scholars have used accumulated incidents to learn about which risks need to be prioritised for investigation, without focussing on detailed technical information. Hoffmann and Frase [22] propose a standardized framework for capturing information about AI harm that facilitates comparability across cases, including incidents recorded in the AIID. Two further studies use manual or machine-learning-based content analysis to establish emergent categories in the incident data: one focusing on AI ethics issues in specific application areas [55] and another on which 'AI tasks' are vulnerable to which 'failure modes' [58]. Feffer et al. [18] explores the role of AI incident databases in raising awareness of AI harms. Specifically, this study adopted the AIID as an educational tool in a ML undergraduate course and found that it positively influenced students' understanding of, and sense of urgency around, AI harms.

2.2 What could be learnt from media reports?

Media and official reporting of autonomous vehicle crashes have prompted another sort of investigation: sociological studies. A prominent example is Stilgoe [48] – an analysis of a 2016 Tesla Model S crash which resulted in the first recorded fatality involving a self-driving car. The car failed to see a truck crossing its path while in the 'Autopilot' mode, and the car's owner died in the ensuing crash (Incident 52, AIID). This time, the investigation did not focus on technical details; instead, Stilgoe [48] interrogates the language of public reporting following the Tesla crash to identify what has been learnt from the incident, and what has been ignored. Observing the different ways actors responded to the incident provides a critical window into *haphazard social learning* – the way in which "society and its institutions make sense of novelty". With new technologies, the work highlights social reactions as an important primer for regulators to decide the distribution of liability and the thresholds of acceptable safety standards.

While media reports are not always an objective or complete way to understand what happened in the incident, they are an important source of information in revealing attitudes and actions that can be learnt from in their own right. Science and Technology Studies (STS) scholars consider public controversy around the use of sociotechnical systems – often documented in media – as *empirical occasions* –researchable events demonstrating relations between a whole variety of actors from science, politics and industry [33].

2.3 Responding to failure

Since the beginning of 2023 the AIID team have been recording responses to AI harms incidents. The official definition relates to one particular type of societal reaction to incidents – formal public acknowledgement from developers and deployers – but in practice, a wider set of responses have been collected. The importance of response is a key aspect of interpersonal ethics – understanding that responsibility lies not only in the prevention of failure and wrongdoing, but in acknowledging their inevitability and examining the subsequent patterns of objection, protest, and response [35, 43, 44, 50]. When we accept that mistakes and wrongdoing are unavoidable parts of human interaction, our focus shifts to how we respond to the harms we cause. This responsibility extends beyond prevention and includes providing appropriate responses when failures occur. The nature of an adequate response [40] is heavily contextual, and depends on the relationship between the parties involved [56], on who can be held accountable [11, 16], the degree to which the harm was foreseeable or preventable [10, 42], the involvement of multiple actors in causing the harm [59], and so on. Responses can manifest in various forms, from apologies to compensation of the victims and preventive measures for the future [51].

3 Methodology

In this section we describe the dataset, followed by our approach to analysing the data.

3.1 The AI Incident Dataset

The AIID is a repository of reported real-world *harm events* involving AI systems. The formal definition of incident is "an alleged harm or near harm event to people, property, or the environment where an AI system is implicated". Harm is not strictly defined, and contributors are advised to document harm if a "plausible argument" can be made that a harm occurred. The editors provide guidelines[4] to support consistency while leaving space and flexibility for editorial judgment. Each incident in the AIID is captured as a collection of publicly available media reports about the same event. The dataset, as downloaded on the 10th of March 2025, contains 962 incidents and 4743 related reports, submitted since 2019. The number of reports per incident varies significantly, between 1-58 reports per incident (see Appendix B). The dataset consists of two tables, one with basic data about each incident, and another with the corresponding reports and their submission details. The metadata covers the incident, the implicated parties (victims, developers, deployers), reporting and submission (see Appendix A).

Media reports and the relevant metadata are collected via a public webform and are reviewed by editors before being incorporated into the dataset. A handful of submitters have submitted hundreds of reports, however, most submit only one (see Appendix C). The top submitters are, or have been, formally associated with the AIID as editors. We note that this is likely due to capacity and resourcing

[4]The AIID editors provide guidance on recognising harms in the Editors Guide (https://incidentdatabase.ai/editors-guide/).

Table 1: Harmed group categories.

Harmed Groups	Notes	Examples (where needed)
User	Harmed as a direct user of AI-enabled software or hardware	
Participant	Forced to engage in AI-enabled software as a result of engaging with a distinct process	Applying for a job; being a customer in a shop; using social media but being harmed by some other AI incident via the content being shown
Public	Harmed while engaging in ordinary civic processes that cannot be opted out of	Being harmed while walking down the street; engaging in a government application; visiting a hospital; being scammed
Prominent	Harmed as a result of being an individual well known to the public	Being impersonated using deepfakes
Vulnerable individual	Harmed as a result of a protected characteristic	Disproportionately effected due to being poor; elderly targeted by scammers
Employee	Harmed in the course of employment	
Organisation	Where an organisation suffers significant harm	Reputational or financial damage
Nature	An explicit harm to nature as opposed to the general harm to the environment implied in all AI use	AI monitoring missed rhino poachers; AI monitoring failed to identify an increase in river pollution

and not any editorial gatekeeping. As significant engagement with the projects inevitably means time and institutional support.

3.2 Analytical approach

This research aims to uncover insights within the AIID beyond technical learning by paying close attention to the responses of different societal actors captured within media reports. We employed a three-tier exploratory approach where: 1) we analysed all recorded incidents with reference to the actors involved – developers, deployers and harmed groups; 2) we focused on a subset of incidents that received formal acknowledgement from developers and deployers ('responses'); 3) we combined these perspectives to systematically study sub-categories of incidents to reveal insightful patterns in the contextualised responses of different actors. Below are the methodology and limitations of each of these steps.

3.2.1 All incidents. To enable quantitative cross-comparison of all incidents with a focus on actors, we used existing data fields for developers and deployers and inductively developed categories:

Developer and deployer. 'Developer' and 'Deployer' are data fields in the AIID. The definitions from the database editor guide are:

- **Developer:** the organizations or individuals responsible for producing either the parts or the whole intelligent system implicated in the incident.
- **Deployer:** the organizations or individuals responsible for the intelligent system when it is deployed in the real world.

The 'Developer' and 'Deployer' fields were used to categorise incidents based on whether the responsible parties were known and, if known, whether they were the same party. This required limited data-cleaning; for example if 'unknown-hacker' was a recorded deployer, we categorised this as unknown.

Harmed groups. For each incident, the following information was reviewed: the 'alleged harmed or nearly harmed groups' tags (e.g., 'the-guardian', 'family-of-lilie-james'), and the description of the incident. While reviewing, appropriate harmed group categories were developed inductively. All incidents were then tagged accordingly. Harmed groups are not mutually exclusive. The labelling was done by two authors of the paper and checked for consistency by cross-labelling. Table 1 describes all of the developed categories.

3.2.2 Incidents with responses. An *AI Incident Response* refers to "a public official response to an incident in the AI Incident Database from an entity (i.e. company, organization, individual) allegedly responsible for developing or deploying the AI or AI system involved in said incident." (Def.1) [46]. Beginning in 2023, editors have sought to tag whether submitted reports constitute a response. Responses are expected to vary in completeness, but at a minimum, responses involve a proactive and direct acknowledgement of the incident. Responses may also include what happened, why it happened and what is being done to prevent something similar from happening again [46]. An incident with a response has at least one report tagged as a **response**. All incidents with responses were analysed qualitatively. This included reviewing and analysing 638 reports associated with 48 incidents.

3.2.3 Establishing patterns. Combining these perspectives, we systematically studied sub-categories of incidents and inductively grouped incidents by common characteristics, identifying 'typical incidents' and factors that impact whether a formal response was received.

3.2.4 Limitations of the dataset. The incidents captured in the AIID are neither a full, nor a representative set of the type and frequency of AI-related harms worldwide. One main reason for this sampling bias is the reliance on media reporting. Incidents published in the media are likely to be those that are surprising or controversial [13, 48], and may reflect some harmed groups over others [21, 31]. Harms that have yet to be well-articulated or investigated [32, 38] are likely to be excluded. In addition, only a very small fraction of the incidents reported in the media will be submitted to the database, and these are almost exclusively from media reports in English. As such the AIID cannot be viewed as a straightforward 'map' of underlying incidents. This is a common challenge with

incident reporting where accumulated data typically reflects reporting behaviour more than occurrence [29]. In recognition of these limitations, the analytical approach above avoids making claims about all incidents. Instead, we focus on learning from the documented incidents by comparing similar incidents that did or did not receive responses to understand relevant contextual factors.

Having acknowledged the aim of learning from a biased subset of incidents, further limitations restrict what can be concluded from the reports themselves. Media reports are not objective accounts of the incidents. They lack detailed information about the implemented systems [54] and are embedded with assumptions rooted in geography [13] and industry [8] (see Section 2.1). The number of reports attached to each incident is also not a true measure of the media attention the incident received. While the number of reports submitted per incident was envisaged as a proxy for interest in an incident, in practice this is not the case. Instead reports being submitted by many submitters, it is often the case that a single, prolific submitter, submits many reports for a given incident. The quantity of reports depends on their efforts as well as on the number of available reports. In an informal conversation with one of these submitters, they described how they would stop searching for more reports for an incident when the new articles ceased to provide new or interesting information, or when they ran out of time. Due to these reasons, we choose not to consider the number of reports as a key factor in the analysis, choosing a mixed-method approach to allow a more meaningful comparison between incidents.

Additional limitations arise from potential subjectivity within tagging within the database. Even with clear guidelines, there is still ambiguity and inconsistencies in data filling and tagging. Similarly, the "response" tag cannot be considered a standardised signal as it is evolving and intentively shaped by evolving editorial choices and practical constraints. Decisions around whether to tag a company statement buried in a news article as a response or only stand-alone press releases can depend on context, bandwidth, and the nature of the report itself. Given that, the responses captured in the AIID are neither a full, nor a representative set and should not be interpreted as such. Nonetheless, our analysis assumes that a significant official response to a recorded incident is likely to be captured. Lastly, we note the database is continuously growing and evolving, with the number of incidents and responses growing rapidly, as well as new fields added to the dataset. This analysis, therefore, is based on a snapshot in time of the database, and the work is intended to encourage continuous analysis of the patterns emerging from the dataset.

4 Results

We start by presenting the quantitative analysis of all incidents in the AIID (Section 4.1), identifying patterns across the database with respect to who is harming and who is harmed. We then focus on incidents with responses (Section 4.2), as these contain richer information on what occurred after the incident. The final section presents the 'typical incidents' found in the databases under inductive categories (Section 4.3), giving the reader a concrete sense of the harms and their aftermath. We emphasize that the quantitative results presented in this paper reflect patterns within the AIID, and should not be interpreted as representative of the full reality of AI-related harms worldwide.

4.1 AI harm incidents

We begin by analysing all incidents across two principal dimensions: groups harmed, and developers and deployers (see Table 1).

4.1.1 Which groups were harmed? As shown in Figure 2 (left), within incidents recorded in the AIID, the public, users and vulnerable individuals were most frequently harmed, with 39%, 35% and 34% of incidents, respectively (see Table 1 for definitions). The public – defined as individuals harmed while engaging in normal civic activities for which there is no alternative (e.g., walking down the street, crossing border control, or participating in election, see Table 1) – were harmed the most, in 39% of incidents. Harmed groups are not mutually exclusive. Figure 2 (right) shows the proportion of the different harmed groups and how often they co-occurred in the same incident. We can see that it is not uncommon for vulnerable individuals to be harmed in the same incidents as users or the public. In incidents where the 'vulnerable' and 'user' tags are both used it is often not only vulnerable users who are harmed. For instance, consider when Yandex, a Russian technology company, released a chatbot which responded to questions with racist and pro-violence replies (Incident 58, AIID). In such cases, harm is not limited to vulnerable users (like marginalised groups) but also impacts non-vulnerable users encountering harmful content, as well as vulnerable individuals who do not use the chatbot but are still impacted by its influence. With incidents where both 'vulnerable' and 'public' tags are used however, it is common for the person harmed to be in the intersection of these categories, for example where they are harmed as a student: they are classed as 'vulnerable' as a child and as 'public' because they cannot avoid going to school.

4.1.2 What can be learnt about the developers and deployers? Figure 1 (left) shows whether the developer and the deployer are known (see Section 3.2.1 for definitions) and whether they are the same organisation if they are both known. In almost 45% of incidents the developer is the same as the deployer. Cross-referencing these categories with harmed groups (Figure 1 right) shows that when users are harmed, it is often when the developer and the deployer are the same. This naturally invites the question as to whether these organisations are what is commonly referred to as *Big Tech*[5]. Indeed, Big Tech companies are quite prominent, with 35% of incidents having *Big Tech* companies as developers or deployers, and 68% when the developer and deployer are the same organisation.

Figure 1 (right) also raises concerns about third-party developers. In the case where someone is harmed as a *Participant* – someone who is harmed by an AI system which they did not directly sign up to use – the developer is not the deployer, or is unknown. We can see that in these cases the developer is often a third-party (25%) or cannot be identified (32%). An example of this is incident 673 in the AIID, where participants are harmed using Adobe stock imagery because, unbeknownst to them, they are using AI-generated images created by third parties which misrepresent reality.

The category of incidents where neither the developer nor the deployer are known is particularly puzzling (9.4%). A qualitative anal-

[5]'Big Tech' was categorised manually from a list of 16 prominent global technology companies. The list was developed based on iterative Google searches with terms such as 'big tech', and includes: ['alphabet','google','youtube','amazon','apple', 'meta', 'facebook', 'microsoft', 'baidu', 'alibaba', 'tencent', 'openai', 'twitter', 'tesla', 'tiktok', 'bytedance'].

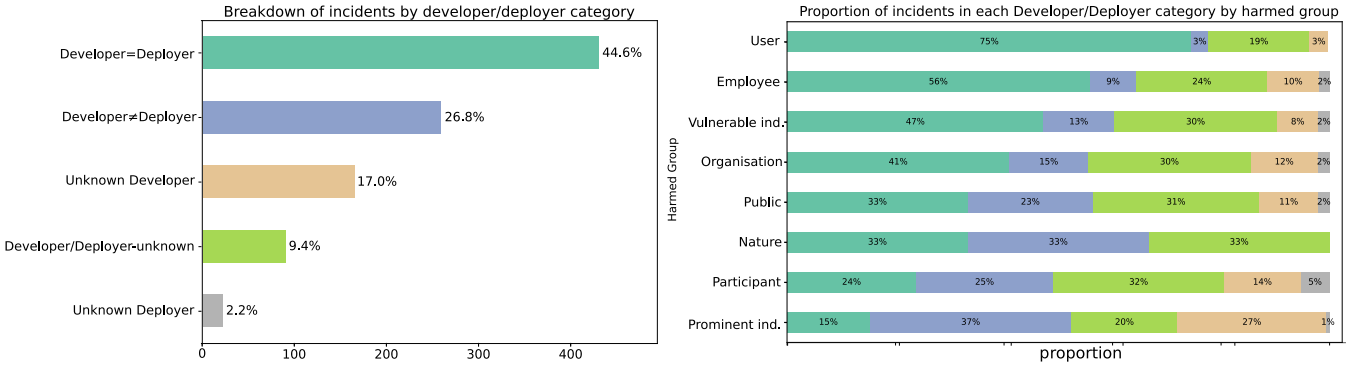


Figure 1: Left: A breakdown of incidents based on whether the developer and the deployer are known and whether they are the same organisation if they are both known. Right: Shown separately for each harmed group.

ysis of these incidents reveals an unusually high similarity between incidents in this category – 80% involve deepfakes. For example, two Canadian residents were scammed by an anonymous caller who used AI voice synthesis to replicate their son's voice asking them for legal fees, disguising himself as his lawyer (Incident 446, AIID).

4.2 Responses

Since the initiative to tag official responses from developers and deployers began, 3.4% of the submitted reports have been tagged as responses. This corresponds to 163 reports, unevenly distributed among 48 incidents, where one incident has 28 responses, and 25 incidents have a single one (see Appendix D). Seven of these incidents have five or more responses – as an official response from the developer or deployer one would naively expect one, or perhaps, two. However, a qualitative analysis of the reports confirmed that 97% of the tagged responses do not meet the definition of a response (see Section 3.2.2). Instead, the responses fit into three categories: (a) response of societal actors; (b) indirect acknowledgement from known developers or deployers; and (c) official response from developers or deployers, in line with the definition.

Response from societal actors. Of the incidents with responses, 29% have no known developer or deployer – 3x over-representation compared the their occurrence in the database. Instead of the responses of developers and deployers, what was being recorded in these cases was the reaction and activity of different societal actors (e.g., victims, concerned parties, legislators), impacted, but not responsible for, the incident. Responses of wider societal actors were not confined to incidents with unknown developers and deployers. All incidents with more than one response (23/48) had responses which recorded the reaction of other parties. Based on the definition of responses (see Def.1 Section 3.2.2), this was not originally anticipated by the AIID – editors set out to tag official responses from developers and deployers but began tagging the responses of wider societal actors. Three incidents (Incidents 597, 616, 645; AIID) had significantly more responses than others. These also had considerably more reports (see Appendix D). Upon inspection, this is because these incidents generated public controversy and thus a greater response of societal actors had been captured.

Indirect acknowledgement from developers or deployers. Indirect response reports refer to a statement or quote from the responsible party gained via private requests, such as from journalists or regulators, instead of as part of an official statement. These often involved blaming a third party or the user. For example, Sports Illustrated magazine was accused of using AI to generate fake authors and their articles (Incident 616, AIID). The response documents a spokesperson from the magazine's publisher saying that the particular articles under scrutiny were licensed from a named third party which had "assured us that all of the articles in question were written and edited by humans" [19]. Another example of blame shifting occurred when a sepsis prediction algorithm generated significantly higher error rates than advertised when used in a hospital setting (Incident 123, AIID). A developer spokesperson was quoted saying "their customers . . . had the ability to tailor the sepsis model to their specific practice".

Official responses. Only 5 out of the 64 reports are official and proactive responses from the developer or deployer. The first incident in this category concerned Microsoft's chatbot, 'Tay', which began tweeting racist, sexist, and anti-semitic comments during the first 24 hours of deployment (Incident 6, AIID). In response, the Corporate Vice President published a blog titled *Learning from Tay's introduction.* A second incident involved an "autonomous security robot colliding with a 16-month-old boy while patrolling" a shopping centre (Incident 51, AIID). The response was the organisation's first ever *Field Incident Report* shared in the trade media. A more recent incident concerned racially and politically biased outputs from Google's Gemini chatbot. Possibly from a result of "overcorrection" of biased outputs, Gemini generated racially diverse Nazis, which lead to an statement from Google acknowledging that "Gemini is offering inaccuracies in some historical image generation depictions", published on X (previously Twitter), and saying they are working to fix it. However, the statement did not include an apology per se, just stated that the wide representation was generally good, but here it was "missing the mark". In all incidents, there was little scope for deniability, and the responses expressed the intention to learn and iteratively improve. However, they communicated nothing about what was being learnt or what would be done differently, engaging (at least publicly) only superficially in learning from the incident.

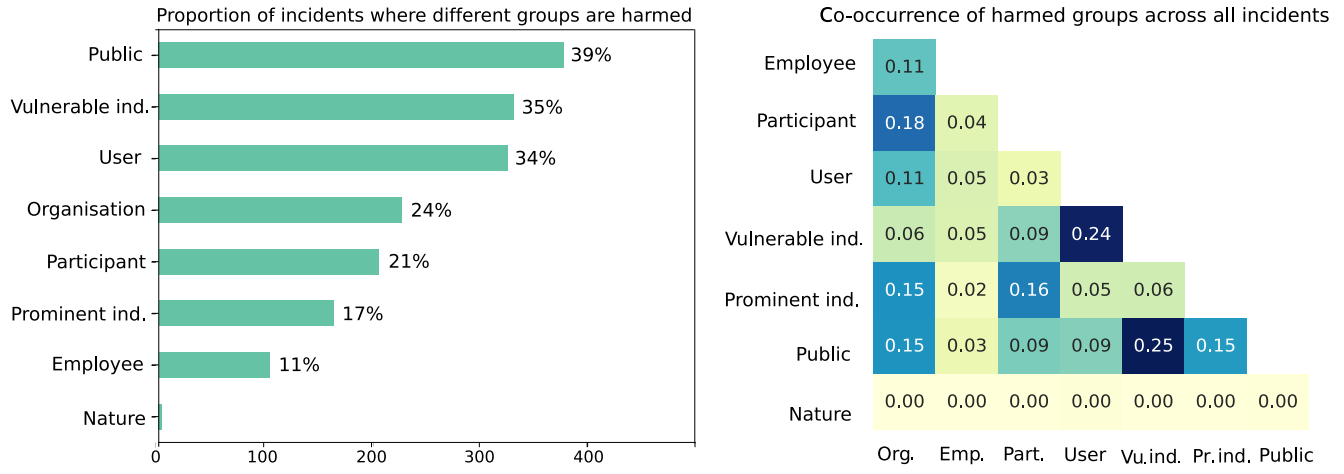


Figure 2: Left: proportion of all incidents by harmed group (not mutually exclusive; see Table 1 for full definitions of all groups). Right: co-occurrence of different harmed groups using the Jaccard Similarity Index (Higher values indicate groups frequently appear together).

4.3 'Typical' incident categories

The different analytical framings introduced above allowed us to study subsets of the incidents and identify incident patterns which were likely to prompt a response.

4.3.1 Developer=deployer 'typical' incident: Big tech incidents where users are harmed. Users were harmed in 76% of incidents where 'big tech' companies were both developers and deployers. 174 incidents fit this category and cover a wide range of harms. Broadly, these can be divided into two subcategories: (1) incidents relating to social media companies, and (2) Tesla incidents.

Typical incident 1: Social media incidents. These incidents tend to be related to how social media companies present and share information with users. For example, four incidents in this subset relate specifically to TikTok's *For You* algorithm:

> An investigation by NewsGuard into TikTok's handling of content related to the Russia-Ukraine war showed its "For You" algorithm pushing new users towards false and misleading content about the war within less than an hour of signing up. (Incident 185, AIID)

Typical incident 2: Tesla crashes. 41 incidents within this category relate to Tesla crashes. In several cases, emergency service vehicles are harmed. For example:

> A Tesla on Autopilot mode failed to see a parked fire truck and crashed into its rear on an interstate in Indiana, causing the death of an Arizona woman. (Incident 319, AIID)

While the responsible parties are known, harmed users seem relatively powerless in calling for meaningful accountability. Proportionately, there are fewer incidents with responses when the Developer is also the Deployer compared to other categories (see Appendix E) – with only 33% of the incidents with responses belonging to this category compared to 45% of the incidents as a whole. Analysing existing responses provides further nuance: there are seven incidents with responses where users are harmed in the 'Developer = Deployer' category. In three of these, the organisations were held accountable by legislators, including by invoking existing trading laws (Incident 435, AIID) and data legislation (Incident 513, AIID). In the remaining cases, the response was mostly insubstantial, from statements of acknowledgement (Incident 642, AIID) to temporarily disabling features (Incident 861, AIID). In the case of the chatbot Tay (Incident 6, AIID; see Section 4.2), where no regulation was applicable, and the only response from Microsoft was a blog post in which it promised to 'iterate' on the technology. In contrast, when Microsoft harmed an organisation (Incident 612, AIID), it launched an official probe into the incident within a month.

4.3.2 Developer ≠ deployer.

Typical incident 3: Large language models deployed by third-party businesses. 'Big tech' developers are much less common when they are not also the deployers; however, this is the case for a significant minority. Typically, these incidents are those where LLMs, developed by 'Big tech', are deployed by third parties such as businesses or individuals.

> The AI-produced, procedural-generated sitcom broadcasted as a Twitch livestream "Nothing, Forever" received a temporary ban for featuring a transphobic and homophobic dialogue segment intended as comedy. (Incident 462, AIID)

Typical incident 4: Government applications where the developer is known. This category consists of incidents where the public is harmed and the deployer is a government institution (37% of all incidents where developer and deployer are known but not the same). One third of these relate to applications of AI systems used by police forces, mostly in similar contexts. For example, 7 incidents relate to police use of *ShotSpotter* – an AI-based product used to

triangulate the location of gunfire, based on sound recording. We note this focus could be due to sampling bias, as many of these reports were submitted by a small team of database editors.

> ShotSpotter audios were previously admitted to convict an innocent Black man in a murder case in Chicago, resulting in his nearly-one-year-long arrest before being dismissed by prosecutors as insufficient evidence. (Incident 255, AIID)

Despite clear accountable parties, there were no responses in the cases of LLMs deployed by third parties, and only responses in the case of government applications. This response is attached to two incidents (Incidents 74 & 529 AIID) delating changes to how Detroit police uses facial recognition following 3 false arrests and one successful lawsuit. The response suggests more safeguarding measurers but double down on the benefits of using the technology. Similarly, most responses obtained in this category relied upon existing accountability mechanisms, including the courts (Incident 608, AIID) and public-sector independent research (Incident 123, AIID). (See Appendix F for example responses in this category.)

4.3.3 Both Developer and Deployer are unknown. Incidents where both the developer and the deployer were unknown make up only 9.4% of incidents in the database. Despite this small proportion, this category provides valuable insight.

Typical incident 5: Deepfakes shared on social media. 80% of incidents in this category involved deepfakes. Of these, the majority shared anonymously on social media. For example:

> In Spain, an AI app was used to digitally alter photos of young girls to appear naked. This manipulation sparked an investigation after these images were circulated in Almendralejo, a town in the Extremadura region, raising serious concerns about digital privacy violations and the potential spread of these images on pornographic sites. (Incident 610, AIID)

The prominence of this type of incident suggests deepfakes shared on social media are particularly liable to generate incidents with no identifiable accountable parties. Moreover, social media companies play a role in obscuring developers and deployers in these cases: deepfakes can be shared without the source being declared and under an unverifiable pseudonym.[6] The responses in this category captures the reaction of wider societal actors rather than responsible parties. The responses suggest that the powerlessness created by a lack of recourse contributed to the mobilisation of different societal actors. For instance, when Tom Hank's likeness was used in deepfaked advertisements (Incident 606, AIID) he is reported to have said: "[T]here [are] discussions going on in … to come up with the legal ramifications of my face and my voice — and everybody else's — being our intellectual property".

Incident 597 (see the case study below) is amongst the most prominent examples of strong societal response. This incident – in which nude deepfakes of female students were circulated in a New Jersey high school – has 21 responses (see Appendix G). Importantly, there is sufficient granularity such that you could begin to map out the actions of various actors (victims, communities, the school, the police, the county, legislators) following the incident and the subsequent impact on policy.

Typical incident 6: Scams using Deepfakes. There are 44 incidents in this category reporting on attempted or successful scams, and in almost all of these, scammers used deepfakes. Many of the deepfakes are of prominent individuals, including romantic scams, in which the victim believed they were in a romantic relationship with a famous individual. For example:

> Scammers allegedly used AI-generated image manipulation tools, along with fake social media and WhatsApp accounts, to reportedly impersonate actor Brad Pitt and convince a French interior decorator that she was in a romantic relationship with him. Over 18 months, they allegedly fabricated selfies and messages, reportedly leading Anne to divorce her husband and transfer $850,000 under the false pretence that Pitt needed money for kidney treatment while his accounts were frozen. (Incident 901, AIID)

5 Discussion

The original goal of the database, inspired by aviation, was to use it to avoid recurring implementation mistakes. Encouraging developers and deployers to submit 'official responses' is one of two ways the AIID is trying to augment the database with implementation details to enable technical collective learning. However, in contrast to the aviation case – in which investigatory details are provided by dedicated industry teams – the success of this endeavour relies on the cooperation and commitment of the industry to collective learning. So far, our findings suggest that the approach of seeking proactive, substantive responses from developers and deployers is not fruitful: none of the responses submitted appeared to be submitted directly – as hoped, by developers or deployers – and even the few responses which met the core definition (5/163) lacked the substantive technical detail required for effective learning.

We claim that currently, the primary value of the database is not learning to avoid implementation failure, but rather learning about the state of both incidents and the responses from different actors in the wake of AI harm. Our results demonstrate the value of the AIID as a means of documenting how accountable and non-accountable parties are responding to incidents, and how wider societal actors call for action. We take inspiration from interpersonal ethics in focusing not on the avoidance of wrongdoing, but on its inevitability. Once we accept that it is impossible to live a life completely devoid of mistakes and wronging others, what becomes central to responsibility is our responsiveness to the resulting harms we do [35, 43, 44, 50]. All responsible actors are subject to requirements not only to avoid wronging others where possible but to provide an adequate response when they inevitably fail to do so.

What constitutes an adequate response, and whether it is achieved depends on contextual factors: the nature of the relationship between the wrongdoer and wronged party [56]; standing to blame and hold accountable [11, 16]; the avoidability and foreseeability of the harm and additional failures of responsibility that lead to

[6] Some social media platforms (e.g., TikTok) have policies requiring "disclosure of synthetic or manipulated media" [27] others (e.g., Meta) have banned them [53] although this does not appear to prevent sharing in practice.

Case Study: AI-Generated Fake Nudes Circulated at a New Jersey High School (Incident 597, AIID)

This incident involves the image-based sexual abuse of female students from a high school in Westfield, New Jersey. AI-generated fake nude images of thesen students were circulated amongst the student body. The incident, occurring during the summer and brought to the school's attention on October 2023, has sparked widespread concern among parents, students, educators, lawmakers, and the public. Despite a lack of concrete evidence of the images' existence, the incident triggered a police investigation, calls for legislative action, and a broader societal debate about the dangers of generative AI misuse and the gaps in legal recourse for victims.

After the incident came to light, the first reported response was a letter from one of the parents to other parents calling them to connect and demand a better response. The second is a report, written by a group of parents, of the interaction between the parents and the school following the incident. Parents organized meetings and voiced demands for preventative measures, such as AI misuse awareness training for students. Despite the incident occurring outside school and during the summer, beyond the school's direct jurisdiction, it was expected that the school will respond strongly to the incident. The school acknowledged the incident and took some steps to address it. The school's actions included addressing the deletion of the alleged images and educating students about the consequences of AI misuse. The School Principal's "pledge to raise awareness marks the beginning of what could be a widespread educational effort to prevent such abuses of technology."

The police launched an investigation into the incident, but their efforts were hampered by the absence of concrete evidence, as no images had been recovered. Reports indicated that some students and staff had seen the images in private group chats, but the lack of access to the images complicated efforts to trace their origins or identify the perpetrators. The local council also publicly acknowledged the incident, which marked the beginning of a broader public discussion. Notably, reports from the UK started to pick up the story after the council's acknowledgment, suggesting a growing international interest in the issue.

The media played a critical role in amplifying the incident and framing the public debate. Articles ranged from basic reporting on the police investigation to in-depth op-eds exploring the implications of generative AI misuse. Opinion pieces called for stronger legal protections and societal awareness about the dangers of deepfakes ("State lawmakers should be making the creation and distribution of fake pornography illegal").

A Westfield High School student, 14, who said she was among more than 30 female students whose photos were manipulated and possibly shared publicly, and her mother, have expressed frustration over what they say is a "lack of legal recourse in place to protect victims of AI-generated pornography". The student launched a website and charity focused on supporting victims of AI misuse and raising awareness about its risks. She and her mother also met with lawmakers and participated in advocacy efforts in Washington, D.C., helping to push for new legislation. While the U.S. Justice Department claims this kind of content would be prosecutable under existing federal child pornography laws that cover drawings and cartoons depicting minors engaged in explicit sex, it can't point to a single prosecution for AI child porn under this legislation. Following the incident, New Jersey State committed to drafting new laws to criminalize the creation and sharing of AI-generated fake nudes. Lawmakers recognized the incident as a catalyst for accelerating conversations about regulating AI and protecting victims. At the federal level, U.S. Representatives Joe Morelle (D-NY) and Tom Kean Jr. (R-NJ) reintroduced the "Preventing Deepfakes of Intimate Images Act," which would criminalize the non-consensual sharing of AI-generated intimate images and require AI tools to include clear disclosures indicating generated content. The bipartisan nature of this effort underscored the urgency of addressing the issue.

the wrong taking place [10, 42]; the role of other actors in causing the harm [59] and so on. The response itself can take many forms: contrition, apologies, recompense, and attempts to avoid future incidents, amongst others [51].

The database proves as a vital record of not just *who* is accountable but also successful and unsuccessful attempts in the practice of *holding* them to account. Our approach demonstrates the database as a valuable resource to interrogate these responses and the lack thereof, reflecting on accountability in practice. This answer calls from the algorithmic accountability literature: worries that the opacity of algorithms [7] and increasingly complex barriers to accountability [12] are "provid[ing] an easy excuse for irresponsibility" [48]. Our work adds to scholarship seeking to understand the practical limitations of accountability in different circumstances by scrutinising and studying real-word cases [5, 40, 48] as a means to open up practical, social or legal, ways forward [2].

Our findings reveal patterns not only in the responses of developers and deployers, but also in the response of wider societal actors in demanding an adequate response. These patterns highlight factors that appear to impact the likelihood of a substantive response in prominent incidents.

The existence of an known accountable party does not increase the likelihood of a response to an incident. In fact, there is some evidence to the contrary. In 71.5% of incidents in the database the developer and the deployer are known. Intuitively, and perhaps naively, these are incidents where we would expect a substantive response to be more likely – the responsible organisations or individuals can, in principle, provide more context, technical detail and an indication of what can be done to avoid the issue in the future. The results show that practice this is not the case in: the 'Developer = Deployer' category has a large proportion of all incidents but a smaller proportion of those with responses (45% vs. 33%, respectively). It is the

opposite when both the Developer and Deployer are unknown – despite there being no known accountable party, and perhaps even because of it, the demand for a response from wider societal actors is more significant.

Even when an accountable party responds it does not always reflect accountability. For example, Microsoft's response to the Tay incident (Incident 6, AIID; see Section 4.2) illustrates how responses can be used opportunistically, by embedding implicit opinions about the terms of trial-and-error that the public ought to accept [32]. Explaining how the product was tested in the Tay case, Lee says: "once we got comfortable with how Tay was interacting with users [through user studies], we wanted to invite a broader group of people to engage with [it]. It's through increased interaction where we expected to learn more... the logical place for us to engage with a massive group of users was Twitter". The widely accepted view that it is important to learn incrementally through seeking a broader group of users, is used to smuggle in the much more contentious view that a "massive" group of users is the appropriate next step. By claiming this step is logical, the implication is that it ought to be accepted. Similar patterns of language have been found elsewhere, with organisations framing new experimentation as "mere baby steps" [6, 48]. These are ways in which the narrative following incidents is co-opted by implicated parties as an opportunity to prepare society for the innovation they intend to carry out [32], rather than reflecting genuine accountability.

The extent of response from accountable parties depends on who is harmed and whether deployment is direct or third-party. Consider the case in which 'Big Tech' is both developer and deployer and users are harmed. When the harms impacting users were under the jurisdiction of a regulator, and top-down pressure was applied, Big Tech companies were forced to engage with their roles as accountable actors (see Section 4.2). When the harms impacting users were not covered by regulation, engagement was either not forthcoming or it was superficial (see Section 4.2). This highlights how power can impact accountability: circumstances can arise in which accountable actors are sufficiently powerful such that merely exposing a shameful incident is not enough to compel them to a substantive response [2]. Explicit regulator pressure does not seem to be required, however, when the harmed party were organisations. For example, when Microsoft harmed users it produced a blog post, but when it harmed an organisation it launched an internal inquiry (see Section 4.2). This suggests that when incidents are exposed, organisations might be in a better position of power, which might compel a more substantive response from Big Tech than users.

The 'Developer ≠ Deployer' category reveals further patterns in the lack of responses. *Typical incident 3* involved LLMs being deployed by third-party businesses. Despite its prominence, no responses from Big Tech were found. In contrast, there were several responses recorded when the incident was similar but Big Tech was both developer and deployer of the LLM (Incident 6 & 645, AIID; see Section 4.2). This suggests that the shift to third party deployment of LLMs might have decreased the degree to which Big Tech are held accountable as deployers. There are examples in which third-party deployers have been forced to respond substantively. These cases involved vulnerable individuals and existing bottom-up accountability mechanisms such as via the courts (see Section 4.3). These mechanisms are unlikely to be useful for holding Big Tech accountable in response to the growing set of harms caused by third party deployment of generative AI [25], where harmed users are often neither vulnerable nor protected in the relevant ways by regulation. In Europe, the EU AI Act specifically designates responsibilities to developers and deployers. While it is too soon to measure its impact on the actions of different actors, this could be an area of valuable future study using the database.

The data is rich where there is controversy, but the lack of controversy where one might be expected is also a valuable source of learning. Counterintuitively, incidents where there is no known developer and deployer provide the richest insight into contexts with a substantive response following AI harm. This richness stems from the public controversy stimulated in a number of these cases. As shown in the case study (see the blue box above), when deepfakes were anonymously shared on social media, the incident reports captured significant response from a wide variety of societal actors – from victims to local organisations to legislators. The demands, actions, and promises of different societal actors provide a window onto a process of *haphazard social learning* whereby society works out the "terms" of responsibility with regards to new technology [48]. While doubtless fuelled by an awareness of deepfakes as a new and severe threat [14], it was the lack of recourse that seemed to motivate the societal response, and shape the legislative demands that result.

The debate around deepfake incidents did not identify the role that social media companies play in obscuring the developers and deployers responsible. Public discussion centred on the lack of recourse, with the legislative effort focused almost exclusively on holding the perpetrator responsible. However, social media companies play a critical role in obscuring perpetrators. Our analysis revealed how sharing practices obscured responsibility in most incidents involving deepfakes shared on social media. This contributes to the 'many hands' discussion within the algorithmic accountability literature – the common use of open-source libraries and pre-trained models complicates the questions of responsibility [12, 37]. Here we identify an additional problem: how some 'hands' can obscure others. In this case, the private mode of sharing has obscured the developer, deployer and the human perpetuator. This highlights that in the case of AI incidents, the question of who to demand a response from, and in what ways, is nuanced, and depends on the specific context and technology involved.

A shift in controversy around similar incidents is also enlightening. *Typical incident 2* (Section 4.3) are Tesla crashes in which a Big Tech company were implicated. In the literature, the response of societal actors to autonomous vehicle incidents has been criticised – a response is demanded from the wrong actor – the human operator is typically blamed, despite little control over the behaviour of the system [17]. However, we see an interesting pattern where a prominent sub-group of autonomous vehicle incidents caused a different level of controversy and a call for accountability. These were incidents involving emergency vehicles. For example, when a Tesla on autopilot crashed into a parked police car in California, journalists blamed Tesla and not the driver:"This marks the third time a Tesla on Autopilot has been reported to have hit a parked emergency vehicle... The real question is why Tesla is allowed to provide a system it admits is a beta and not a fully-tested product."

(Incident 323, AIID).

Categories where there is an apparent lack of debate are also illuminating. For instance, worryingly, there was only one response found in *Typical incident 4*, where the public was harmed by governments, although harms were substantial (e.g., wrongful arrests). An analysis of the words and actions within these reports could be performed to understand why such incidents have not stimulated greater debate. A similar approach was used to discover that a public challenge relating to the implementation of AI had ceased in Canada due to a convergence of expectations around the technology [13]. Noticing a lack of debate is important: allowing harm to go unchallenged and risks them becoming 'just another fact of life' Stilgoe [49].

6 Conclusion

In this paper, we demonstrate the AIID's potential as a resource to understand where and by whom harms occur; where substantive responses were achieved and through which mechanisms; what responses look like in particular circumstances, and when they are missing. The patterns of response (or lack thereof) we unveil provide insights about how society negotiates responsibility for AI harms. Our contribution is two-fold. Firstly, we establish a revised approach to using the AIID as a rich resource for pragmatic socio-technical learning about effective accountability mechanisms. This approach avoids existing criticisms: while media reports lack detailed implementation information, they contain insights into how actors respond and when accountability is achieved in practice. The second contribution are the recurring patterns in the factors affecting the likelihood of achieving a substantive response from accountable actors following AI harm. We believe this approach will be valuable for scholars studying algorithmic accountability, technology governance, and institutional responses to technological harm. For policy makers and regulators, the findings highlight important gaps in current accountability frameworks, and how such datasets can be used to monitor how these evolve with new regulatory measures.

Acknowledgments

The authors thank Sean Mcgregor and Daniel Atherton for valuable discussions and detailed feedback on an earlier draft of this paper. MZ acknowledges support from the Leverhulme Trust grant ECF-2021-429.

References

[1] 2024. *AI Incident Database - About.* https://incidentdatabase.ai/about
[2] Mike Ananny and Kate Crawford. 2018. Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *new media & society* 20, 3 (2018), 973–989.
[3] Daniel Atherton. 2024. AI Incident Information Sharing Resources. https://github.com/jphall663/awesome-machine-learning-interpretability/blob/master/README.md#ai-incident-information-sharing-resources
[4] Jascha Bareis and Christian Katzenbach. 2022. Talking AI into being: The narratives and imaginaries of national AI strategies and their performative politics. *Science, Technology, & Human Values* 47, 5 (2022), 855–881.
[5] Solon Barocas, Sophie Hood, and Malte Ziewitz. 2013. Governing algorithms: A provocation piece. *Available at SSRN 2245322* (2013).
[6] Mads Borup, Nik Brown, Kornelia Konrad, and Harro Van Lente. 2006. The sociology of expectations in science and technology. *Technology analysis & strategic management* 18, 3-4 (2006), 285–298.
[7] Jenna Burrell. 2016. How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big data & society* 3, 1 (2016), 2053951715622512.
[8] Harry Camilleri, Carolyn Ashurst, Nithya Jaisankar, Adrian Weller, and Miri Zilka. 2023. Media coverage of predictive policing: Bias, police engagement, and the future of transparency. In *Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*. 1–19.
[9] Christopher Carey. 2023. Google rolls out AI to optimise traffic lights and cut emissions. *Cities Today* (October 2023). https://cities-today.com/google-rolls-out-ai-to-optimise-traffic-lights-and-cut-emissions/
[10] D Justin Coates. 2016. The epistemic norm of blame. *Ethical Theory and Moral Practice* 19, 2 (2016), 457–473.
[11] Gerald A Cohen. 2006. Casting the First Stone: Who Can, and Who Can't, Condemn the Terrorists? 1. *Royal Institute of Philosophy Supplements* 58 (2006), 113–136.
[12] A Feder Cooper, Emanuel Moss, Benjamin Laufer, and Helen Nissenbaum. 2022. Accountability in an algorithmic society: relationality, responsibility, and robustness in machine learning. In *Proceedings of the 2022 ACM conference on fairness, accountability, and transparency*. 864–876.
[13] Guillaume Dandurand, Fenwick McKelvey, and Jonathan Roberge. 2023. Freezing out: Legacy media's shaping of AI as a cold controversy. *Big Data & Society* 10, 2 (2023), 20539517231219242.
[14] Adrienne De Ruiter. 2021. The distinct wrong of deepfakes. *Philosophy & Technology* 34, 4 (2021), 1311–1332.
[15] Francis Durso, MS Raunak, Rick Kuhn, and Raghu Kacker. 2022. Analyzing failures in artificial intelligent learning systems (FAILS). In *2022 IEEE 29th Annual Software Technology Conference (STC)*. IEEE, 7–8.
[16] Gerald Dworkin. 2000. *Reasoning practically*. Oxford University Press, Chapter Morally Speaking.
[17] Madeleine Clare Elish. 2019. Moral crumple zones: Cautionary tales in human-robot interaction (pre-print). *Engaging Science, Technology, and Society (pre-print)* (2019).
[18] Michael Feffer, Nikolas Martelaro, and Hoda Heidari. 2023. The ai incident database as an educational tool to raise awareness of ai harms: A classroom exploration of efficacy, limitations, & future improvements. In *Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*. 1–11.
[19] Sarah Fortinsky. 2023. Sports illustrated responds to accusations it published AI-generated content. https://thehill.com/policy/technology/4330197-sports-illustrated-responds-to-accusations-it-published-ai-generated-content/. [Accessed 08-01-2025].
[20] Adrian Groza and Anca Marginean. 2023. Brave new world: Artificial Intelligence in teaching and learning. *arXiv preprint arXiv:2310.06856* (2023).
[21] Stephen Hilgartner. 2000. *Science on stage: Expert advice as public drama*. Stanford University Press.
[22] Mia Hoffmann and Heather Frase. 2023. Adding structure to AI harm. https://cset.georgetown.edu/publication/adding-structure-to-ai-harm/
[23] Julian Jacobs. 2024. The artificial intelligence shock and socio-political polarization. *Technological Forecasting and Social Change* 199 (2024), 123006.
[24] Mohamad Kassab, Joanna DeFranco, and Phillip Laplante. 2022. Investigating Bugs in AI-Infused Systems: Analysis and Proposed Taxonomy. In *2022 IEEE International Symposium on Software Reliability Engineering Workshops (ISSREW)*. IEEE, 365–370.
[25] Khoa Lam. 2023. ChatGPT Incidents and Issues. https://incidentdatabase.ai/blog/chatgpt-incidents-and-issues/.
[26] Nora Freya Lindemann. 2022. The ethical permissibility of chatting with the dead: Towards a normative framework for 'Deathbots'. *Publications of the Institute of Cognitive Science* 1 (2022).
[27] Ben Lovejoy. 2023. TikTok deepfakes: MrBeast, Tom Hanks, Gayle King; call for ban. https://9to5mac.com/2023/10/04/tiktok-deepfakes/. [Accessed 08-01-2025].
[28] Giampiero Lupo. 2023. Risky artificial intelligence: The role of incidents in the path to AI regulation. *Law, Technology and Humans* 5, 1 (2023), 133–152.
[29] Carl Macrae. 2016. The problem with incident reporting. *BMJ quality & safety* 25, 2 (2016), 71–75.
[30] Carl Macrae. 2022. Understanding Autonomous Vehicle Risk: A Case Study Analysis. *Safety Science* (2022).
[31] Noortje Marres. 2015. Why map issues? On controversy analysis as a digital method. *Science, Technology, & Human Values* 40, 5 (2015), 655–686.
[32] Noortje Marres. 2021. No issues without media: The changing politics of public controversy in digital societies. In *MEDIA*. Intellect, 228–243.
[33] Noortje Marres and David Moats. 2015. Mapping controversies with social media: The case for symmetry. *Social Media+ Society* 1, 2 (2015), 2056305115604176.
[34] Sean McGregor. 2021. Preventing repeated real world AI failures by cataloging incidents: The AI incident database. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 15458–15463.
[35] Michael McKenna. 2012. *Conversation & responsibility*. Oup Usa.
[36] L McWilliams. 2024. New EY research finds AI investment is surging, with senior leaders seeing more positive ROI as hype continues to become reality. https://www.ey.com/en_us/newsroom/2024/07/new-ey-research-finds-ai-investment-is-surging-with-senior-leaders-seeing-more-positive-roi-as-hype-continues-to-become-reality

[37] Helen Nissenbaum. 1996. Accountability in a computerized society. *Science and engineering ethics* 2, 1 (1996), 25–42.
[38] Rob Nixon. 2011. *Slow Violence and the Environmentalism of the Poor*. Harvard University Press.
[39] Nikiforos Pittaras and Sean McGregor. 2022. A taxonomic system for failure cause analysis of open source AI incidents. *arXiv preprint arXiv:2211.07280* (2022).
[40] Inioluwa Deborah Raji, Peggy Xu, Colleen Honigsberg, and Daniel Ho. 2022. Outsider oversight: Designing a third party audit ecosystem for ai governance. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*. 557–571.
[41] Rowena Rodrigues, Anais Resseguier, and Nicole Santiago. 2023. When artificial intelligence fails: The emerging role of incident databases. *Pub. Governance, Admin. & Fin. L. Rev.* 8 (2023), 17.
[42] Gideon Rosen. 2004. Skepticism about moral responsibility. *Philosophical perspectives* 18 (2004), 295–313.
[43] Thomas M Scanlon. 2000. *What we owe to each other*. Harvard University Press.
[44] Thomas M Scanlon. 2008. *Moral dimensions: Permissibility, meaning, blame*. Harvard University Press.
[45] Tim Schröder and Michael Schulz. 2022. Monitoring machine learning models: a categorization of challenges and methods. *Data Science and Management* 5, 3 (2022), 105–116.
[46] Janet Schwartz. 2022. Introducing AI Incident Responses. https://incidentdatabase.ai/blog/introducing-ai-incident-responses/.
[47] Kris Shrishak. 2023. How to deal with an AI near-miss: Look to the skies. *Bulletin of the Atomic Scientists* 79, 3 (2023), 166–169.
[48] Jack Stilgoe. 2018. Machine learning, social learning and the governance of self-driving cars. *Social studies of science* 48, 1 (2018), 25–56.
[49] Jack Stilgoe. 2020. Who's driving innovation. *New Technologies and the Collaborative State. Cham, Switzerland: Palgrave Macmillan* (2020).
[50] Peter F Strawson et al. 2003. Freedom and resentment. *Free will* 2 (2003), 72–93.
[51] Richard Swinburne. 1989. *Responsibility and atonement*. Oxford University Press.
[52] Lionel Nganyewou Tidjon and Foutse Khomh. 2022. Threat assessment in machine learning based systems. *arXiv preprint arXiv:2207.00091* (2022).
[53] The Straits Times. 2023. News anchors targeted by deepfake scammers on Facebook. https://www.straitstimes.com/world/news-anchors-targeted-by-deepfake-scammers-on-facebook. [Accessed 08-01-2025].
[54] Violet Turri and Rachel Dzombak. 2023. Why we need to know more: Exploring the state of AI incident documentation practices. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*. 576–583.
[55] Mengyi Wei and Zhixuan Zhou. 2022. AI ethics issues in real world: Evidence from AI incident database. *arXiv preprint arXiv:2206.07635* (2022).
[56] Roger Wertheimer. 1998. Constraining condemning. *Ethics* 108, 3 (1998), 489–501.
[57] Tammy Worth. 2024. Are robots the solution to the crisis in older-person care? *Nature* (2024).
[58] Xinhui Zhan, Heshan Sun, and Shaila M Miranda. 2023. How does AI fail us? A typological theorization of AI failures. *ICIS 2023 Proceedings* 25 (2023), 1–17.
[59] Sjoerd D Zwart. 2015. Responsibility and the problem of many hands in networks. In *Moral Responsibility and the Problem of Many Hands*. Routledge, 131–166.

A Key data fields for each table in the AIID

Reports:

- Authors of the report
- Description and text of the (news) report
- Date published
- Report URL
- Source domain

Report submitters:

- Date submitted
- Submitter (can be anonymous)

Incidents:

- Date (of incident)
- Description of the incident
- Alleged developer(s) (of system)
- Alleged deployer(s) (of system)
- Alleged harmed or nearly harmed parties (organic tags, i.e., the submitter based it on reading of the articles. They do not meet a predefined taxonomy.)
- Reports relating to incident (list)

B Reports per incident

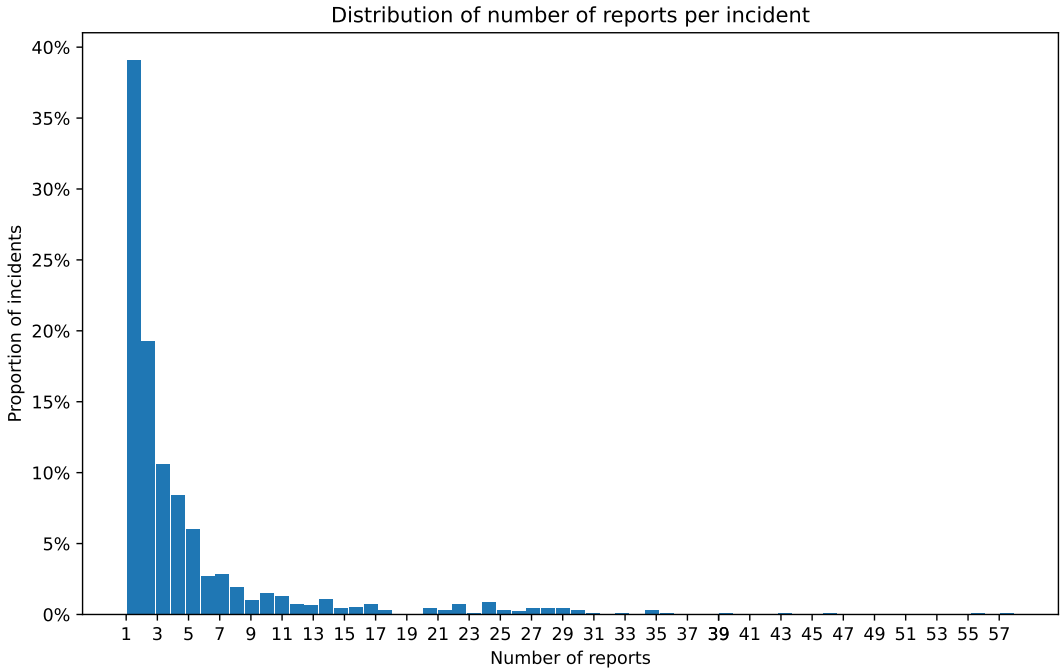


Figure 3: A distribution of number of reports per incident

C Reports by submitter

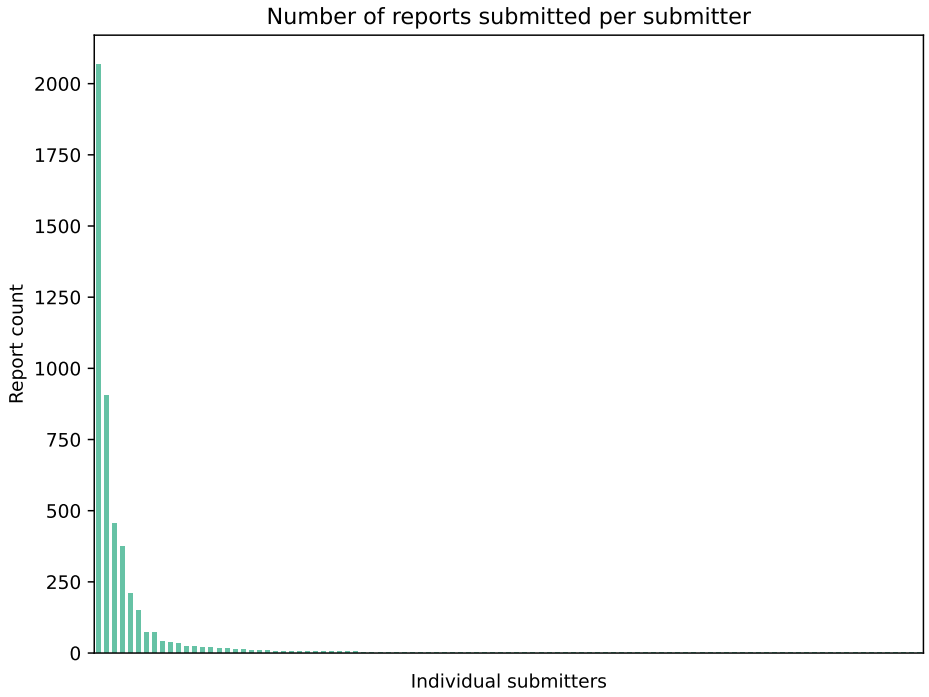


Figure 4: The number of reports submitted by individual submitters

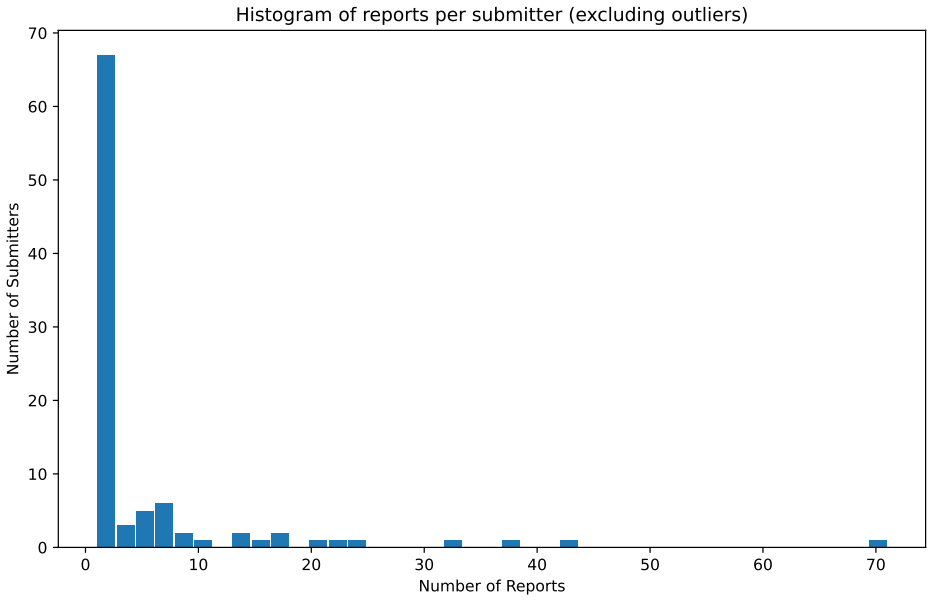


Figure 5: A histogram of reports per submitter. Outliers are excluded.

D Incidents with Responses

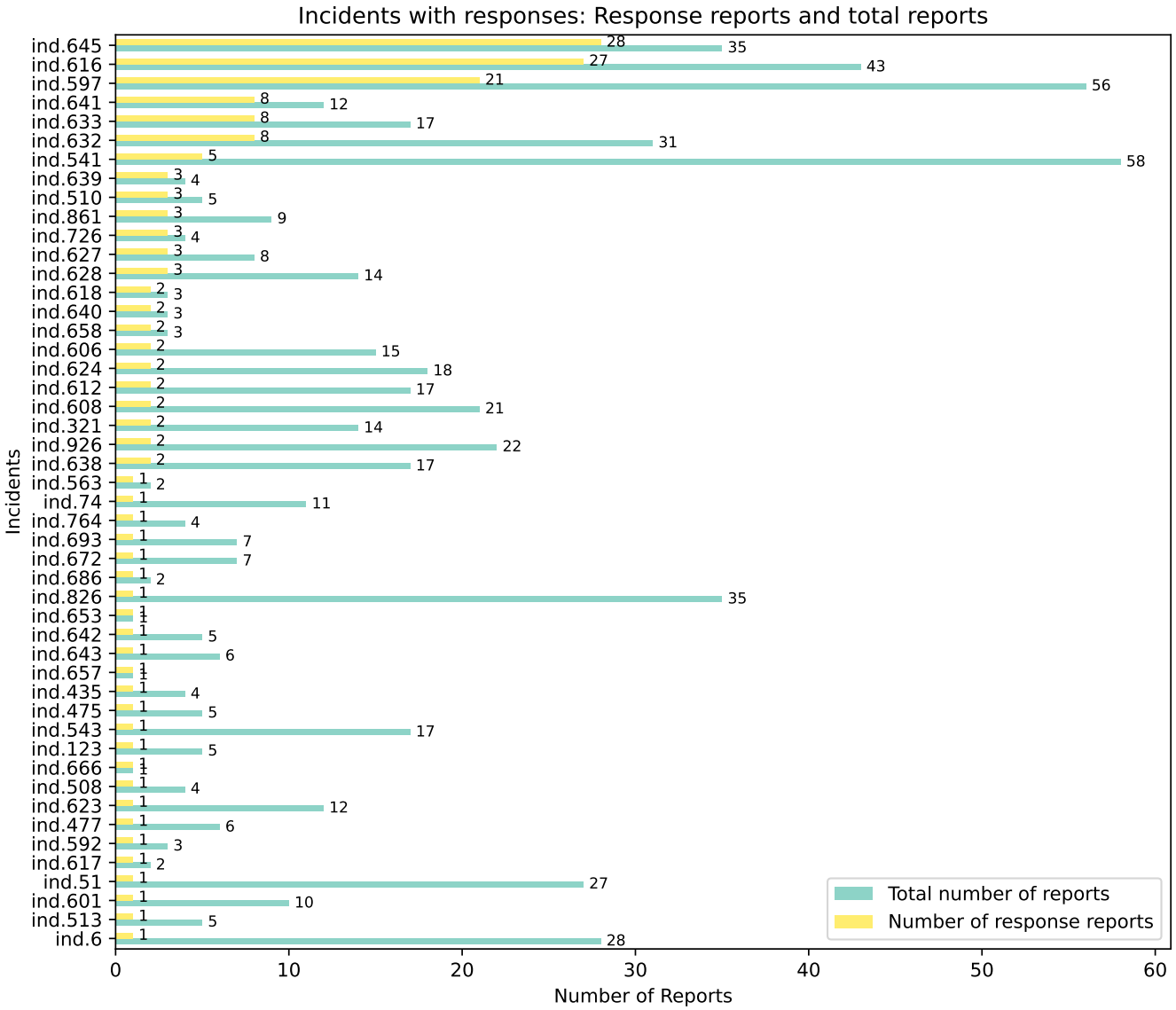


Figure 6: The number of reports and responses reports for each incidents with at least one response.

E Responses by Developer/Deployer category

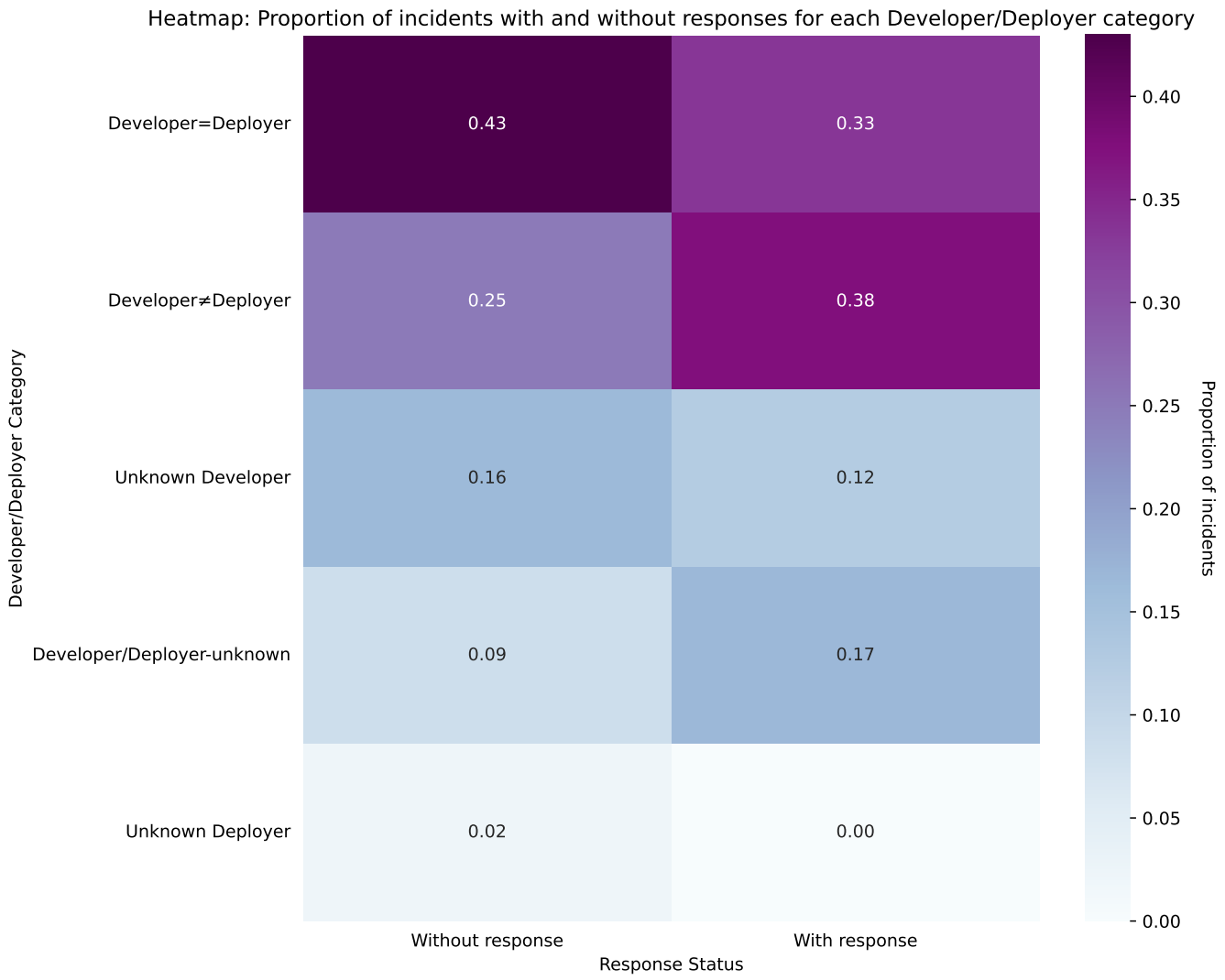


Figure 7: Proportion of incidents with and without responses by Developer/Deployer category.

F Responses when the developer and deployer are known but not the same

Title	Notes
Security Robot Rolls Over Child in Mall	Undeniable physical harm to child
*Epic Systems's Sepsis Prediction Algorithms Revealed to Have High Error Rates on Seriously Ill Patients	Revealed by researchers based at a hospital
Celebrities' Deepfake Voices Abused with Malicious Intent	Responsible start-up tweets with ideas to 'iterate' safeguarding features
Bing Chat Tentatively Hallucinated in Extended Conversations with Users	Bing blogs with iterations to safeguarding features
*UnitedHealth Accused of Deploying Allegedly Flawed AI to Deny Medical Coverage	Victims raise with courts
ChatGPT Reportedly Produced False Court Case Law Presented by Legal Counsel in Court	Discovered by courts

Table 2: Incidents with responses where the deployer and the developer are known but they are different organisations. Responses marked with a * where existing accountability mechanisms are employed in cases where vulnerable individuals were harmed.

G Incident 597 reports and response timeline

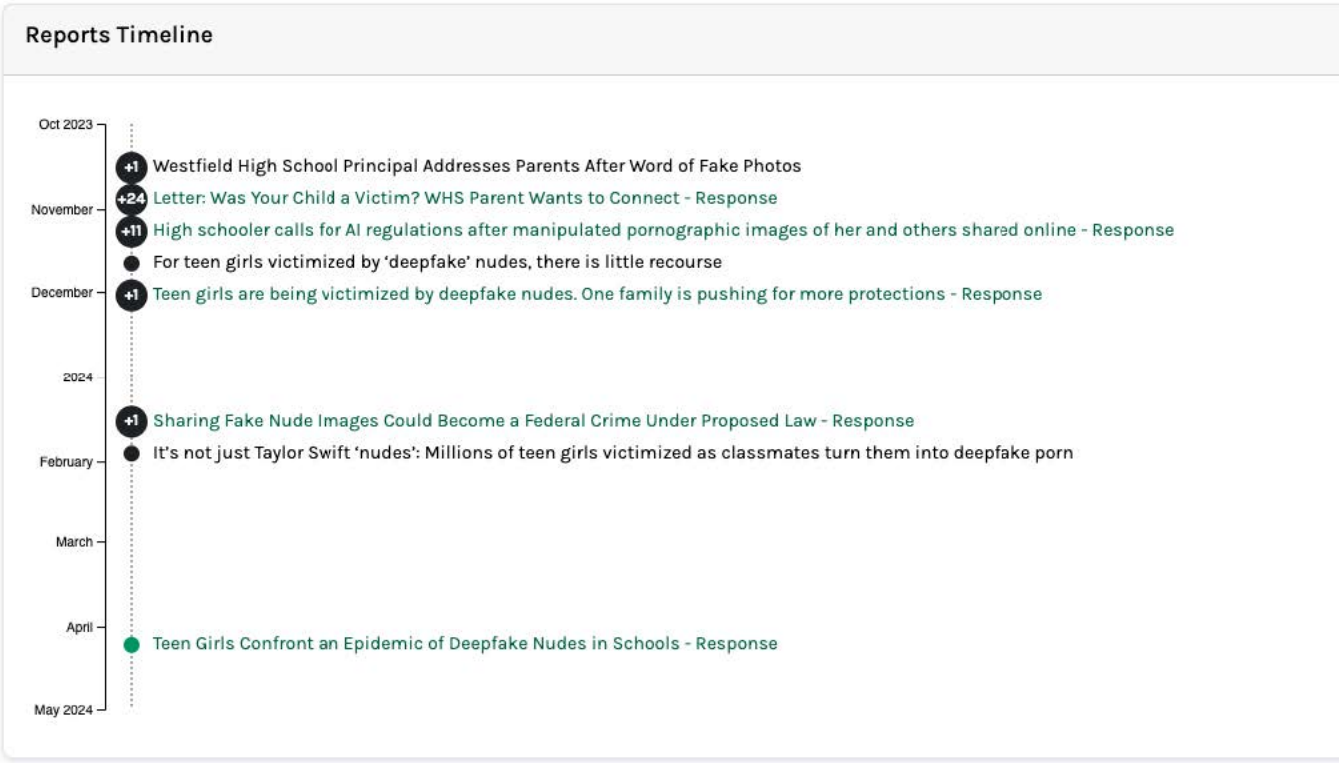


Figure 8: Incident 597 reports and response timeline. Copied from the AIID website