Theme Article: Special Issue on AI Failures: Causes, Implications, and Prevention

Decoding Real-World AI Incidents

Julia De Miguel Velázquez, *King's College London, London, WC2B 4BG, United Kingdom*
Sanja Šćepanović, *Nokia Bell Labs, Cambridge, CB3 0FA, United Kingdom*
Andrés Gvirtz, *King's College London, London, WC2B 4BG, United Kingdom*
Daniele Quercia, *Nokia Bell Labs, Cambridge, CB3 0FA, United Kingdom*

Abstract—Initial efforts to report AI incidents aimed to improve transparency, yet systematic studies have been rare. To address this, we analyzed all 639 real-world incidents in the emerging AI Incidents Database, devising an ethical framework focused on evaluating incidents along what, where, who, and how dimensions. For each incident, we categorized 'what' type of harm occurred, finding malicious intent is uncommon. We identified 'where' harm originated, discovering that most harms occurred during human interactions rather than at the model development stage, emphasizing the need for sociotechnical considerations in development. We discerned 'who' was harmed, finding that incidents impacting corporations were likely under-reported, stressing the need for better reporting methods within companies. Lastly, we assessed 'how' harm could be morally judged: some incidents were considered low risk under current regulations but were still found to be morally wrong.

Artificial intelligence (AI) systems are increasingly integrated into our daily lives. Unfortunately, not all AI systems operate as intended, and some can have harmful consequences. For example, biased hiring algorithms can enable gender discrimination, and self-driving cars can cause injuries [1], [2].

The relatively recent but fast-paced AI industry poses challenges for developers to anticipate how their systems may produce harm. Therefore, assessing AI systems 'in the wild' is crucial to tackling this challenge, with documenting AI incidents as an initial step [3]. There are various initiatives that aim to document AI incidents, such as the AI Incident Database (AIID), the OECD AI Incidents Monitor and the AIAAIC. Initial attempts to report these incidents have aimed at improving bias auditing, risk assessments, and public awareness [4], [5]. Yet, these databases have not been systematically studied, which limits our understanding of how systems fail. This, consequently, impedes the development of tangible recommendations to prevent such incidents in the future.

To tackle this, we developed a comprehensive ethical AI framework by integrating various taxonomies of AI harms, enabling us to thoroughly examine the incident landscape. Subsequently, we applied this framework to map all 639 real-world incidents in the AIID. Our findings serve two main purposes: first, we analyzed and extracted insights from the existing landscape of AI incidents; second, we identified current gaps and deficiencies in incident documentation.

We found that AI systems often cause problems even though their developers have good intentions. This happens because the developers do not always consider all the possible ways their systems can cause harm once deployed in the real world. Most problems occur not because of how the AI is built, but because of unexpected situations when people use it. It is hard to predict these issues, but to solve this, we need people from different fields to work together and learn about both the social and technical aspects of AI.

We also found that most AI problems are reported by regular users, not by companies or institutions, which means many problems might not be reported at all. To fix this, we suggest new ways to report incidents, like anonymous reporting, to avoid confidentiality issues. We also found that some AI systems follow the

Digital Object Identifier 10.1109/XXX.0000.0000000
Project website: https://social-dynamics.net/ai-incidents

law but still seem wrong to people. We therefore think that developers should pay more attention to what people think about AI.

RELATED WORK

We categorized the literature on taxonomies of AI-based harms into four dimensions: 'what' type of harm it is, 'where' the harm originated, 'who' was harmed, and 'how' the harm could be judged.

What. AI harm taxonomies typically focus on distinguishing between different types of harm. While there are numerous AI domain-specific taxonomies (e.g., privacy), few encompass all AI domains [6]. The AIID uses various taxonomies to report its incidents, but they are more descriptive than truly taxonomical, relying on free text rather than predefined categories [7]. A recent study examining the AIID identified nine significant types of harm, although its taxonomy is not exhaustive, as some categories overlap [8]. For example, 'racial bias' and 'gender bias' are distinct categories but could potentially be aggregated into a higher-level category such as 'demographic biases'. In contrast, two separate teams at DeepMind and one at The Alan Turing Institute conducted scoping reviews of computer research, each offering comprehensive classifications of harms [1], [2], [9]. However, a common limitation across all these taxonomies is the lack of a clear definition of 'harm'.

Where. Previous research has highlighted an excessive focus on addressing the technical aspects of AI systems [2]. However, AI risks, influenced by both technological and social factors, cannot be adequately addressed through technical solutions alone. Recognizing this, DeepMind introduced a sociotechnical framework that incorporates social context to identify *where* the harm originates [1]. This framework expands the scope of harms beyond those occurring during model development ('capability' in Figure 1) to include downstream harms during human interactions ('human interaction'), or broader societal and environmental impacts ('systemic impact'). This conceptual distinction is crucial for developing more effective and nuanced mitigation strategies.

Who. Often overlooked in taxonomies, certain studies have explored the aspect of *who* was harmed [2]. While some investigations distinguish between individual, collective, societal or biospheric harms [9], [10], others suggest standardized criteria in line with the EU AI Act, distinguishing between 'AI Subject' and 'AI User' [11]. Additionally, some studies categorize the social groups most affected, such as children [7].

How. Although significant progress has been made in defining and categorizing AI harms, it is still unclear how individuals perceive AI incidents. This is particularly relevant as there are increasing efforts to involve the public in AI design [12]. Developers should consider how people evaluate their systems as well-designed AI systems tend to gain higher acceptance from people. While some studies have used Moral Foundation Theory (MFT) to explore people's judgments of AI systems in their lives [13], no studies have investigated moral judgments towards real-world AI incidents.

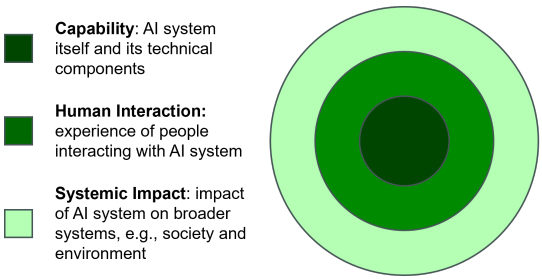

FIGURE 1. DeepMind's sociotechnical framework offers three layers to assess *where* harm can originate [1]. Starting from 'capability' to 'human interaction', and up to 'systemic impact'.

CONCEPTUAL FRAMEWORK

'What' type of harm is it?

To characterize the 'what' dimension, we chose DeepMind's taxonomy of AI harms [1] (summarized in Table 1 with actual AI incidents), and did so because of three main reasons. First, its framework includes a harm typology that proposes a solution to the previous over-focus on model-related harms. Second, it is widely endorsed by the technology industry and researchers, and is often included in AI risk assessments and auditing processes [10], [14], [15]. Lastly, its ability to capture complex concepts in a minimalistic way makes it easy to visualize and analyze.

'Where' did the harm originate?

To characterize the 'where' dimension, we again opted for DeepMind's classification [1], which distinguishes three sociotechnical layers of 'where':

> *Capability layer* refers to the AI systems themselves, focusing on their technical aspect and how they are created, including the data used for training and model refinement.
>
> *Human interaction layer* refers to the actual experience of individuals interacting with the

TABLE 1. Taxonomy reflecting 'what' type of harm, which was taken from DeepMind's classification [1]. The examples were taken from our analysis of the AIID, except those marked with '*' which were taken from the original paper.

Risk area	Examples
Representational and Toxicity: AI systems where data misrepresents certain social groups or performs differently, and generating toxic, offensive, abusive, or hateful content.	
Unfair representation	Researchers from Boston University and Microsoft Research demonstrated gender bias in the most common techniques used to embed words for natural language processing (NLP).
Unfair capability distribution	New Zealand passport robot reader performs worse for Asian people, and once rejected the application of an applicant with Asian descent, claiming his eyes were closed.
Toxic content	MIT Media Lab researchers created AI-powered "psychopath" named Norman by training a model on the "dark corners" of Reddit.
Misinformation Harms: AI systems generating and facilitating the spread of inaccurate or misleading information that causes people to develop false beliefs.	
Propagating false beliefs and misconceptions	Google's AI chatbot Bard provided false information in a promotional video on the first satellite to photograph a planet outside the solar system, causing shares to temporarily fall.
Erosion in trust in public information	Michael Cohen, former lawyer for Donald Trump, used Google's AI chatbot Bard to generate legal case citations, which were unknowingly included in a court motion.
Pollution of information ecosystem	Wikipedia bots meant to remove vandalism were clashing with each other and form feedback loops of repetitive undoing of the other bot's edits.
Information and Safety Harms: AI systems leaking, reproducing, generating or inferring sensitive, private, or hazardous information.	
Privacy infringement	Australian government reviewers of grant applications input applicants' work to systems such as ChatGPT to generate assessment reports, posing confidentiality and security issues.
Dissemination of dangerous information	Amazon was reported to have shown chemical combinations for producing explosives and incendiary devices as frequently bought together items via automated recommendation.
Malicious Use: AI systems reducing the costs and facilitating activities of actors trying to cause harm (e.g. fraud or weapons).	
Influence operations	A deepfake video claimed France 24, a French media outlet, reported a Kyiv plot to assassinate French President Macron, which was later debunked by France 24.
Fraud	A mother in Arizona received a ransom call from an anonymous scammer who created her daughter's voice allegedly using AI voice synthesis.
Defamation	Voices of celebrities and public figures were deepfaked for impersonation and defamation and were shared on social platforms such as 4chan and Reddit
Human Autonomy and Integrity Harms: AI systems compromising human agency, or circumventing meaningful human control.	
Violation of personal integrity	Instagram allegedly contributed to the death of a teenage girl in the UK through exposure and recommendation of suicide and self-harm content.
Persuasion and manipulation	A Black man was wrongfully detained by the Detroit Police Department due to a false facial recognition result.
Overreliance	Major Australian retailers reportedly analysed in-store footage to capture facial features of their customers without consent.
Misappropriation and exploitation	Text-to-image model Stable Diffusion was reportedly using artists' original works without permission for its AI training.
Socioeconomic and Environmental Harms: AI systems amplifying existing inequalities or creating negative impacts on employment, innovation, and the environment.	
Unfair benefits distribution from access to models	Better hiring and promotion pathways for people with access to generative AI models in a way creating digital divide.*
Environmental damage	Increase in net carbon emissions from widespread model use.*
Inequality and precarity	Zillow's AI predictive pricing tool wrongly forecasted housing prices due to rapid market changes, prompting division shutdown and layoff of a few thousand employees.
Undermine creative economies	Substituting original works with synthetic ones, hindering human innovation and creativity.*
Exploitative data sourcing and mining	Content moderators at Facebook demand better working conditions, as automated content moderation system failed and exposed them to psychologically toxic content.

AI systems such as the effects, usability, and quality of the system.

Systemic impact layer refers the wider societal impacts of AI systems such as the economy and the environment.

The three layers do not necessarily follow a specific order in time or sequence. They are simply meant to show where harm can start. For example, systemic impact harms can happen not only after the AI system is used (e.g., a political deepfake in South Korea allegedly led to public mistrust) but also during model development (e.g., Amazon and Uber allegedly deployed AI systems that offered gig workers lower wages than expected.) [10].

'Who' was harmed?

Often overlooked in AI harm taxonomies is the question of *who* was harmed in the incident, which is crucial for understanding the landscape of harms [2]. To align with efforts to standardize reporting for the EU AI Act and unify incident descriptions, we distinguish between two key stakeholders: the *AI User*, who deploys and manages the AI system, and the *AI Subject*, who uses the AI system. Additionally, we include a third class of stakeholders: *Institutions, General Public, and Environment*, in line with other research [9], [10]. This class encompasses wider societal actors and environmental elements affected by AI system use.

In our analysis, we focused on the primary stakeholder harmed in each incident, excluding potential negative side effects. For instance, if YouTube's recommendation algorithm failed to filter inappropriate content for children, we identified children (AI Subject) as the harmed stakeholder, rather than YouTube (AI User), even if YouTube's reputation suffered as a side effect. While every AI system may have side effects, we concentrated on direct harms (or wrongs) resulting from the AI failures that led to specific incidents.

'How' is it morally judged?

When assessing moral judgements, we compared the social acceptance (or disapproval) of the incidents against the more official assessments of incidents. We opted for the widely used MFT to capture how individuals may perceive and judge the moral implications of AI incidents [16]. MFT has been utilized to comprehend public perceptions of AI systems [13]. It extracts five moral dimensions, each with negative (-) and positive (+) connotations:

Harm includes connotations such as harmful (-), violent (-), protective (+), and caring (+).

Fairness includes connotations of unjust (-), discriminatory (-), fair (+), and impartial (+).

Loyalty includes connotations such as disloyal (-), traitor (-), loyal (+), and devoted (+).

Authority includes connotations such as disobedient (-), defiant (-), lawful (+), and disrespectful (-).

Purity includes connotations such as indecent (-), obscene (-), decent (+) and virtuous (+).

We then compared these perceptions with the classifications outlined in the EU AI Act [11], which categorizes AI-related harms based on their level of risk. This allowed us to identify morally questionable incidents that are considered low risk under the EU AI Act.

METHODOLOGY

Collecting and exploring the incidents

AI incident databases. We chose the AIID for reporting AI incidents because it gives detailed information about carefully curated incidents. Other databases either have fewer – even if detailed – incidents (such as AIAAIC, from which the AIID gets some of its incidents) or have many incidents taken automatically from news sources without much human review (such as OECD AIM).

AIID. As an initiative that aims to standardize reporting of AI incidents, users can upload incidents which are reviewed by the main editors. Currently, there are 649 incidents, each derived from one or more reports, totalling 3412 reports. These reports are news articles. The incidents span from 2013 to 2024, with an increasing number uploaded each year. Additionally, there are 28 incidents scattered between 1983 and 2012. Geographically, most incidents happen in the US (82%), followed by the United Kingdom (6%), and China (4%).

Collecting the AI Incidents. We gathered the incidents from the AIID website, retrieving all the information available as of March 2024.

Exploring the AI incidents. The incidents consistently include a title, description, date, author, and occasionally other information [7]. We examined the variables present in all incidents ('title', 'description', 'AI

https://incidentdatabase.ai/

deployer', 'AI developer', 'AI harmed subject'), along with others that appeared in some (15%) but could provide valuable context ('AI system description' and 'AI harm distribution'). Duplicate incidents were identified and eliminated, as were minor issues that had been downgraded from incidents. This resulted in a total of 639 distinct incidents for analysis.

Classifying the incidents

Operationalizing the framework with Large Language Models (LLMs). To process all the incidents, we needed to choose and validate an automatic approach. We opted for LLMs to implement the framework and classify the incidents, leveraging their capability to process large text volumes [17]. In line with our research goals, LLMs have demonstrated utility as zero-shot data annotators [17]. We operationalized the framework using two prompts to guide the LLMs in the classification:

1. *Harm assessment prompt*. This prompt assessed the harm itself ('who', 'where', and 'what') from the perspective of an expert in responsible AI development and implementation. It used input fields from the AIID, including the incident 'description', 'AI deployer/developer', and 'AI harmed subject'. It also included definitions of key concepts proposed in previous work to ensure a thorough description of the AI use involved in each incident [11]. The prompt directed the LLM to describe each AI incident based on these concepts, and it provided definitions of different layers and types of harm.

2. *Risk assessment prompt*. This prompt assessed the moral and legal acceptability of each incident. We ran the prompt three times and selected connotations that appeared at least twice for consistency. It presented five axes with moral dimensions and their positive and negative connotations. Additionally, we included descriptions from the EU AI Act to evaluate whether the AI system in each incident can be categorized as 'Limited or Low Risk', 'High Risk', or 'Prohibited' under the Act.

Running the two prompts on the incidents and validating the output. To assess the LLMs, we engaged in an iterative process, refining the prompts as needed to address any observed discrepancies. For instance, we experimented with incorporating examples alongside the definitions of key concepts, leading to improvements in output quality. This iterative refinement continued until the quality of the prompts aligned with manual results, achieving an accuracy rate exceeding 86% on a previously unseen 10% sample. Subsequently, we applied these refined prompts to analyze the remaining incidents.

RESULTS

What: Most harms are non-malicious

Most incidents involve 'Human Autonomy and Integrity Harms' (37% in Table 2) and 'Representation and Toxicity' (33%). These are followed by 'Information and Safety Hazards' (23%) and 'Misinformation Harms' (21%). The least frequent types are 'Socioeconomic and Environmental' (12%) and 'Malicious Use' (11%).

Where: Harm primarily occurs when AI interacts with humans

We then identified where harm originates by assigning each incident to one of the three layers in Figure 1. This approach expanded our understanding of incidents beyond AI's technical side, recognizing their broader societal impacts. Notably, despite the focus of much research and practitioner work, the capability layer accounts for only a minority of incidents (28% in Table 2). However, the majority of incidents occur at the human interaction layer (69%), representing instances where individuals directly interact with AI systems. Incidents at the systemic impact layer are relatively rare (3%), typically involving job loss or environmental harm.

Who: AI subjects were most frequently reported in incidents

We conducted an analysis to determine 'who' the primary harmed stakeholder was. For this analysis, we focused solely on the stakeholder primarily affected by each incident, without considering potential secondary effects. The overwhelming majority of reported incidents harmed the AI Subject (91.8% in Table 2), the individual directly interacting with the AI system. This was followed by the AI User (7.6%), the stakeholder responsible for deploying and managing the AI system. Incidents affecting Institutions, the General Public, and the Environment were rarely reported (0.6%).

How: AI incidents considered low-risk legally may still face social unacceptability

Using MFT, we found that most incidents were seen as unfair (89.2% in Table 2) and harmful (82.1%). We then grouped these judgments based on the EU AI Act's risk assessment. Prohibited incidents under the Act were negatively charged, with an average of 3.46 negative terms per incident. In contrast, incidents classified as limited or low risk had fewer negative terms on

TABLE 2. Summary of the results for ‘what’, ‘where’, ‘who’, and ‘how’. The results are ranked by the fraction of incidents related to each part, with the result related to main takeaway (reported at the bottom) marked in red. Note that for ‘what’ and ‘how’, more than one category can apply to each incident, but for ‘where’ and ’who’, only one category can apply.

What

Definition: Refers to the type of harm produced by the AI system involved in the incident.

Results		Examples
Human Autonomy and Integrity	38%	Instagram contributed to the death of a teenage girl in the UK allegedly through exposure and recommendation of suicide and self-harm content.
Representation and Toxicity	33%	A Black man was wrongfully detained by the Detroit Police Department due to a false facial recognition result.
Misinformation	23%	Major Australian retailers reportedly analysed in-store footage to capture facial features of their customers without consent.
Information and Safety Hazards	21%	Facebook was reported by users for blocking posts of legitimate news about covid, allegedly due to a bug in an anti-spam system.
Socioeconomic and Environmental	12%	Uber launched a new but opaque algorithm to determine drivers' pay in the US which allegedly caused drivers to experience lower fares.
Malicious Use	11%	Two Canadians were scammed by an anonymous caller who used AI voice synthesis to replicate their son's voice asking them for legal fees.

Takeaway: Harms tend to be non-malicious, suggesting that AI is often built with good intentions despite failing.
Recommendation: Educating developers on AI incidents to overcome 'failure of imagination’ of potential harms.

Where

Definition: Identifies the layer where the harm was produced.

Results		Examples
Human Interaction:	69%	Waze, a Google-owned directions app, led California drivers into the 2017 Skirball wildfires as they tried to evacuate the area.
Capability:	28%	Google’s Gemini AI image generator was overcorrecting racial diversity, offering distorted portrayals in historically white-dominated scenes.
Systemic Impact:	3%	A childcare benefits system (Netherlands) falsely accused thousands of families of fraud, due to an algorithm deeming dual nationalities as risky.

Takeaway: Harm mainly originates at interactions with humans, but mitigation strategies tend to focus on capabilities.
Recommendation: Promoting interdisciplinary collaboration during AI development to assess sociotechnical risks.

Who

Definition: Identifies the stakeholder that was directly harmed in the incident.

Results		Examples
AI Subject:	92%	Google Photos image processing software mistakenly labelled a black couple as "gorillas."
AI User:	7%	A Knightscope K5 security robot ran itself into a water fountain in Washington, DC.
Institutions, General Public, Environment:	1%	A South Korean political candidate created a deepfake avatar which political opponents alleged to be fraudulent and a threat to democracy.

Takeaway: AI Subjects were mostly harmed, while incidents from corporations (AI User) may be underreported.
Recommendation: Calling for new ways (e.g., anonymous and confidential) of incident reporting to avoid conflicts.

How

Definition: Captures people’s moral perceptions of harm, which we compared with institutional assessments.

Results			Examples
Fairness	Unjust: Discriminatory:	89% 30%	The data used to train AI systems that classify chest X-rays led to gender, socioeconomic, and racial biased decisions.
Harm	Harmful: Violent:	82% 4%	An Uber autonomous vehicle (AV) in autonomous mode struck and killed a pedestrian in Tempe, Arizona.
Authority	Disobedient: Defiant:	44% 1%	New York Police targeted Black Lives Matter activists using facial recognition, limiting their right to protest
Loyalty	Disloyal:	22%	Starbucks managers used Kronos’ scheduling algorithm which negatively impacted their employees.
Purity	Indecent: Obscene:	14% 6%	YouTube’s content filtering and recommendation algorithms exposed children to disturbing and inappropriate videos.

Takeaway: Incidents may be perceived negative morally charged perceptions, even if legal.
Recommendation: Developers should include the public’s perception on AI, coupled with the institutional guidelines.

average (2.89%). We delved deeper into the incidents categorized as limited or low risk under the Act but perceived as morally questionable, defined as those containing four or more negative terms. This analysis aimed to uncover potential discrepancies between official risk assessments and moral perceptions, highlighting instances where low-risk incidents may face public resistance. Two main types emerged: incidents with inadequate content moderation (44.4%), such as instances where Google's search engine has displayed antisemitic content when prompted by the word 'Jewish', and incidents involving false content generation (22.2%), such as Australian academics using Google LLMs to generate case studies for a parliamentary inquiry.

DISCUSSION

Educating developers on AI incidents to overcome their 'failure-of-imagination' of risks

We found that harm rarely originates from malicious intent, despite non-maliciousness being one of the five key ethical AI concerns [18]. Malicious use includes instances like non-consensual sex-related deepfakes, political propaganda, and fraud. Their infrequency suggests that AI systems are often built with good intentions, yet they still result in significant incidents. This observation aligns with literature indicating that AI developers often fail to anticipate potential risks from their systems – a phenomenon termed 'failure of imagination' [3]. Sometimes, even with good intentions, developers' backgrounds may limit their awareness of risks, especially towards marginalized groups [4]. Understanding the landscape of AI incidents through harm taxonomies can help practitioners better anticipate risks in the systems they build. This educational approach should be integrated into graduate curricula and workplace training programs [4].

Promoting interdisciplinary collaboration during development to prevent downstream harms

We discovered that the majority of harms occur during interactions with humans, underscoring the crucial need for broader ethical considerations during development to anticipate and prevent downstream negative impacts. Concrete steps to better anticipate downstream harms include increasing research on the human interaction layer, such as user research and behavioral experiments, as well as on the systemic impact layer, such as impact assessments, forecasts, and simulations [1]. These efforts are vital as the contextual use of AI systems significantly shapes the resulting harms [2].

Advocating for innovative incident reporting methods

End users have been the most commonly affected group, while incidents negatively impacting corporations or institutions have been relatively low. However, these numbers may not accurately reflect the true extent of harm, as under-reporting is a significant concern. There are several reasons for potential under-reporting when incidents affect corporations or public institutions. Concerning corporations, low reporting rates may stem from confidentiality and reputation concerns. Despite the option for anonymous submission in the AIID, the lack of strict confidentiality measures may deter AI developers from reporting incidents. To address this issue, we advocate for the establishment of a dedicated, confidential incident reporting database [5]. Incidents are usually reported when they directly impact specific individuals or groups, often resulting in anecdotal reporting. As such, incidents involving public institutions, despite the increasing deployment of AI systems by them (e.g., governments using AI for visa application assessments), are not often reported.

Developers should take into account people's opinions regarding their AI systems even at design stage

As expected, negative moral perceptions were prevalent, with incidents often seen as unfair and harmful. While incidents categorized as 'low risk' under the EU AI Act were generally less morally charged than 'prohibited' ones, we found exceptions. Specifically, incidents involving content moderation and generation categorized as 'low risk' eventually resulted in significant stigma due to failed moderation and false generation. This highlights the need for developers to consider public perceptions of their AI systems to ensure social acceptance, and underscores the importance of integrating judgments from the general population alongside institutional guidelines from the very outset of AI deployment.

LIMITATIONS AND FUTURE WORK

This work has two main limitations. First, most of the AI incidents come from the US because the AIID is a US-based project. This might make our research more focused on the US, even though some incidents come from places like Europe and Asia. Second, the people

who add incidents to the database might have their own ideas about what counts as an AI incident and the news sources they use, which can cause bias. One key editor added 22% of the incidents, and others added different amounts, which can also lead to bias. In the future, researchers could use other incident databases to refine our Ethical AI Framework.

Secondly, the choices made in shaping harm taxonomies might favor some technology areas over others [1], [3], [19]. This aligns with our concern about current reporting strategies, where some incidents get highlighted while others do not. This limitation hampers our ability to define what counts as an incident and could marginalize certain social groups or regions further [3], [20].

CONCLUSION

Despite the growing attention towards real-world AI incidents across academia, industry, and the media, a systematic and comprehensive analysis has been lacking. Our examination of AIID incidents reveals that most harms are unintended, underscoring the challenge AI developers face in anticipating risks. We also discovered that while there is often a focus on technical issues, most harms actually arise during human interaction with AI systems. This highlights the need for interdisciplinary research on these downstream effects. Additionally, we found that AI subjects are typically the ones harmed, but there is a notable lack of reported harms to companies and institutions, likely due to confidentiality concerns. We suggest exploring new, potentially anonymous reporting methods for AI incidents, which also allow for confidential reporting. Lastly, discrepancies between lenient official risk assessments and unlikely social acceptance are exemplified by challenges in moderating viral content (a complex task) and AI-generated content (an emerging field with many uncertainties). As we navigate this complex landscape, as Albert Einstein put it: “We cannot solve our problems with the same thinking we used when we created them.” We should learn from past incidents, prioritize interdisciplinary research, and explore innovative reporting methods to ensure the responsible development and deployment of AI technologies.

REFERENCES

1. L. Weidinger *et al*, “Sociotechnical safety evaluation of generative AI systems,” 2023. [Online]. Available: arXiv preprint arXiv:2310.11986. (URL)
2. R. Shelby *et al*, “Sociotechnical harms of algorithmic systems: scoping a taxonomy for harm reduction,” in *Proc. of the 2023 AAAI/ACM Conference on AIES*, 2023, pp. 723-741. (Conference proceedings)
3. M. Boyarskaya, A. Olteanu and K. Crawford. "Overcoming failures of imagination in AI infused system development and deployment.", 2020, arXiv preprint arXiv:2011.13416. (Workshop)
4. M. Feffer, N. Martelaro and H. Heidari, “The AI Incident Database as an Educational Tool to Raise Awareness of AI Harms: A Classroom Exploration of Efficacy, Limitations, & Future Improvements,” in *Proc. of the 3rd ACM Conference on EAAMO*, 2023, pp. 1-11. (Conference proceedings)
5. V. Turri and R. Dzombak, “Why we need to know more: exploring the state of AI incident documentation practices,” In *Proc. of the 2023 AAAI/ACM Conference on AIES*, 2023, pp. 576-583. (Conference proceedings)
6. H. P. Lee, Y. J Yang, T. S. Von Davier, J. Forlizzi, and S. Das, “Deepfakes, Phrenology, Surveillance, and More! A Taxonomy of AI Privacy Risks,” In *Proc. of the ACM Conference on CHI*, 2024. (Conference proceedings)
7. S. McGregor, “Preventing repeated real world AI failures by cataloguing incidents: the AI Incident Database,” in *Proc. of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 17, 2021, pp. 15458-15463. (Conference proceedings)
8. M. Wei and Z. Zhou, “AI ethics issues in real world: evidence from AI incident database,” 2022. [Online]. Available: arXiv preprint arXiv:2206.07635. (Unpublished manuscript)
9. A. Domínguez Hernández *et al*, “Mapping the individual, social, and biospheric impacts of Foundation Models,” in *Proc. of the 2022 ACM Conference on FAccT*, 2024. (Conference proceedings)
10. C. Thomas, H. Roberts, J. Mökander, A. Tsamados, M. Taddeo and L. Floridi, “The case for a broader approach to AI assurance: addressing “hidden” harms in the development of artificial intelligence,” *AI & SOCIETY*, 2024. (Journal)
11. D. Golpayegani, H. J. Pandit and D, Lewis, “To be high-risk, or not to be—semantic specifications and implications of the AI Act’s high-risk AI applications and harmonised standards,” in *Proc. of the 2023 ACM Conference on FAccT*, 2023, pp. 905-915. (Conference proceedings)
12. T. Shaffer Shane, “AI incidents and ‘networked trouble’: The case for a research agenda,” *Big Data & Society*, vol. 10, no. 2, 2023. (Journal)
13. M. Constantinides and D. Quercia, “Good intentions, bad inventions: how employees judge pervasive technologies in the workplace,” *IEEE Pervasive Computing*, vol. 22, no. 1, 2022, pp. 69-76. (Journal)

14. Birhane, A., Steed, R., Ojewale, V., Vecchione, B., and Raji, I. D., "AI auditing: The broken bus on the road to AI accountability," in *IEEE Conference on SaTML*, 2024. (Conference proceedings)
15. J. Mökander, J. Schuett, H. R. Kirk, and L Floridi, "Auditing large language models: a three-layered approach," *AI and Ethics*, 2023, pp. 1-31. (Journal)
16. J. Haidt, "The new synthesis in moral psychology," *Science*, vol. 316, no. 5827, 2007, pp. 998-1002. (Journal)
17. C. Ziems, W. Held, O. Shaikh, J. Chen, Z. Zhang and D. Yang, "Can large language models transform computational social science?," *Computational Linguistics*, 2023, pp. 1-55. (Journal)
18. A. Jobin, M. Ienca, and E. Vayena, "The global landscape of AI ethics guidelines," *Nature machine intelligence*, vol. 1, no. 9, 2019, pp. 389-399. (Journal)
19. M. Jakesch, Z. Buçinca, S. Amershi and A. Olteanu, "How different groups prioritize ethical values for responsible AI," in *Proc. of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 2022, pp. 310-323. (Conference proceedings)
20. J. Burrel, "Automated decision-making as domination," *First Monday*, vol. 29, no. 4, 2024. (Journal)

Julia De Miguel Velázquez is a PhD student at the Department of Informatics at King's College London. She works in areas of computational social science, responsible AI and gender studies. Contact her at julia.de_miguel_velazquez@kcl.ac.uk.

Sanja Šćepanović is a Senior Research Scientist at Nokia Bell Labs Cambridge (UK). She works in the areas of social computing, earth observation, and responsible AI. Contact her at sanja.scepanovic@nokia-bell-labs.com.

Andrés Gvirtz is an Assistant Professor of Marketing Technology & Innovation at King's Business School, King's College London and a Research Affiliate at King's Institute for Artificial Intelligence, King's College London. He works at the intersection of computational social science, marketing and psychology. Contact him at andres.gvirtz@kcl.ac.uk.

Daniele Quercia is the Department Head at Nokia Bell Labs in Cambridge (UK) and Professor of Urban Informatics at King's College London. He works in the areas of computational social science, urban informatics, and responsible AI. Contact him at quercia@cantab.net.