Why It’s So Hard To Know How Often AI Goes Wrong – And What That Means For Accountability When Algorithms Harm People

Lucia de la Torre (Gonzalez Mantecon), 2025

Supervisor: Steve Eder

In 2020, on a gray January morning in Detroit, Robert Williams was in the driveway with his wife and two young daughters when police pulled up. Without warning or explanation, officers handcuffed him in front of his children and took him away for a theft he hadn't committed. The evidence? A blurry police mugshot generated by facial recognition software. An algorithm – improperly trained to accurately recognize darker skinned people – had decided he and the suspect looked alike. Williams spent 30 hours locked in a crowded cell. His confusion turned to anger when he asked police officers if they had used facial recognition to incriminate him, and he realized technology alone had wrongfully accused him.

Williams' story made headlines that led to public outcry, lawsuits, and new regulations on police use of facial recognition in Detroit. But for every case like his, many others may go unnoticed: a lost job or denied loan from an opaque scoring model, a welfare payment quietly slashed by an automated system, medical negligence caused by a large language model incorrectly summarizing a patient's record.

Another artificial intelligence (AI) injustice case that caught the attention of the media and technology ethics circles was that of Arkansas resident Tammy Dobbs. A few years earlier, in 2016, Dobbs found her state-funded home care suddenly cut from 56 to 32 hours per week. No one from Medicaid visited, and Dobbs – who uses a wheelchair due to cerebral palsy – was offered no explanation. The decision had been generated quietly and automatically, deep inside a Medicaid algorithmic calculation. The cuts left her to pick and choose between the basic needs she wanted to get help with: bathing, eating, dressing. When Dobbs tried to appeal the Medicaid decision, her questions disappeared into a maze of bureaucracy – no one could explain what happened, and no one could be held accountable.

Across the Atlantic, similar algorithmic failures also devastated families. A 2021 Amnesty International report titled “Xenophobic machines: Discrimination through unregulated use of algorithms in the Dutch childcare benefits scandal” documented how an AI system used by Dutch tax authorities to detect childcare benefit fraud systematically targeted parents from immigrant and low-income backgrounds. Thousands of families like Batya Brown's were wrongly accused and forced to repay benefits they had legitimately received, plunging them into debt and poverty. “Since we've been acknowledged as victims of what I call the ‘benefits crime’, even four years later, we’re still being treated as a number,” Brown told Amnesty International. Years after being officially recognized as victims, many families continue struggling with the financial and emotional aftermath of what became known as the “Dutch childcare benefits scandal”.

All of these incidents have something in common: a line of code made a decision with real-life consequences. The processes that made this happen are nearly impossible to track, challenge, or even detect. Unlike a plane crash or a medication error, algorithmic failures rarely trigger public disclosure or government investigation. The systems at fault are often hidden by corporate secrets or shrouded in technical complexity, their consequences scattered and mostly invisible.

Who, then, is responsible for noticing? For counting? For learning from these AI-generated mistakes before they’re repeated?

Why Most AI Failures Go Uncounted

Every day, people interact with automated systems that decide high-stakes outcomes, sometimes even without their knowledge. Yet unlike plane accidents or medical negligence – where industry-specific regulation requires that every crash, near miss, or harmful side effect is logged, shared, and studied – AI still operates without any universal incident reporting framework. In other domains, meticulous documentation exists so that catastrophes aren't repeated. Why, then, is the AI field any different?

One of the attempts to create a collective memory of AI gone wrong is the Artificial Intelligence Incident Database (AI Incident Database or AIID), launched by researcher Sean McGregor.

"I was inspired by safety-critical fields like aviation and medicine, where incident databases help prevent repeated failures. AI lacked a collective memory for failures, leading to repeated mistakes in design and deployment," McGregor explains.

The AI Incident Database is crowdsourced: journalists, technologists, advocates, and even affected individuals are all encouraged to submit reports. This way, the AI Incident Database strives to aggregate incidents ranging from minor errors to major failures across sectors and geographies. Each record is annotated with details on harm type, sector, year, and where available, geographic location. Yet McGregor candidly admits:

"The impact [of AI] is typically experienced by lay people, or at least by people not usually in the AI industry," McGregor observes. "So to understand what harms are occurring in the world, you actually need people to report those harms and have them entered into a database."

The crowdsourced idea sounds promising, in theory. In practice, though, you're only likely to submit a case to the AI Incident Database if you're already aware of both

AI-related harms and the database itself – meaning you're a journalist, probably someone working in the AI industry, or someone who has connected with an advocacy group that supports people affected by these systems. In short, someone in the tech or tech advocacy space.

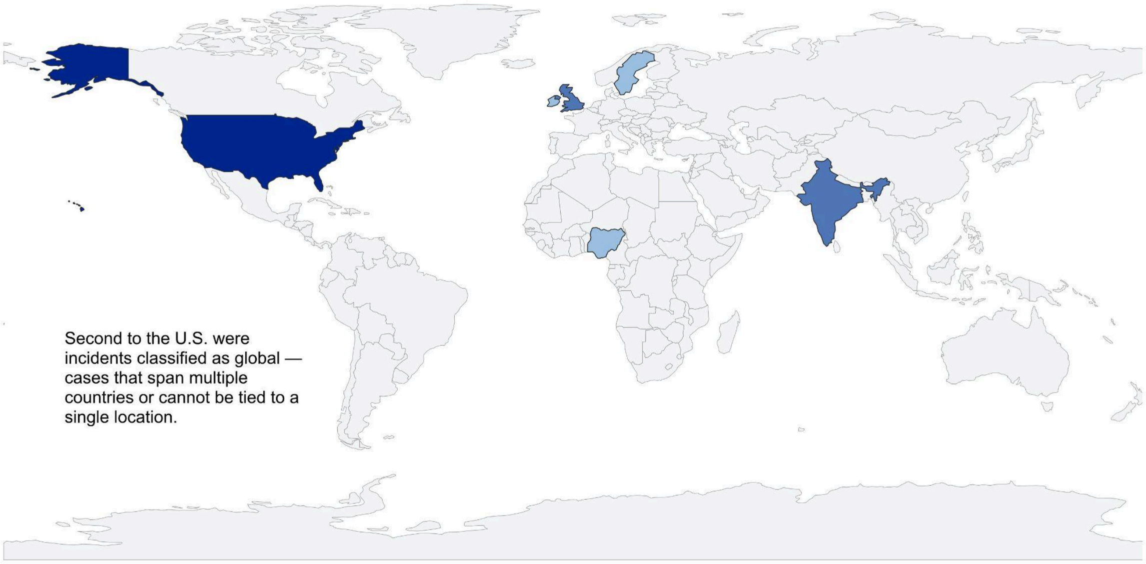
McGregor is aware of this: "We desperately need more data from outside North America and Europe. Most incidents in the Global South go unreported," he adds. "It's not that harms aren't happening – they are – but there's limited local media coverage and little systematic archiving."

I downloaded the AI Incident Database to see what patterns might emerge if I sorted incidents by country and by sector. The dataset contains thousands of entries covering a wide span of years, each describing some real or alleged failure of an AI system.

But, in practice, only a fraction of these records could be cleanly slotted into a geographic category. Many list the United States as the location, and a large share are marked "Global" – meaning the incident couldn't be tied to a single place, or the location wasn't properly entered. Entries for other countries – such as the UK, India, Nigeria, Ireland, and Sweden – exist, but in much smaller numbers. And for the majority of cases, there is no clear location tag at all, which makes precise mapping difficult.

AI Safety Incidents Cluster in Countries With Better Reporting

The AI Incident Database shows most incidents reported between 2010 and 2025 happen in America, but data gaps likely obscure the real worldwide distribution.



Second to the U.S. were incidents classified as global — cases that span multiple countries or cannot be tied to a single location.

A similar pattern came up when I looked at industry sectors. Most of the incidents logged come from areas like information technology and government – fields where there are tech-savvy workers and active watchdog groups. Meanwhile, stories about problems with AI in retail, education, banking, or hospitality are much less common, even though mistakes in those areas can have really serious consequences.

Tech and Transportation Lead in Reported AI Incidents

According to the AI Incident Database, between 2010 and 2025, the Information sector leads with 26 reported cases, though lack of data and varying disclosure practices may skew representation across industries.

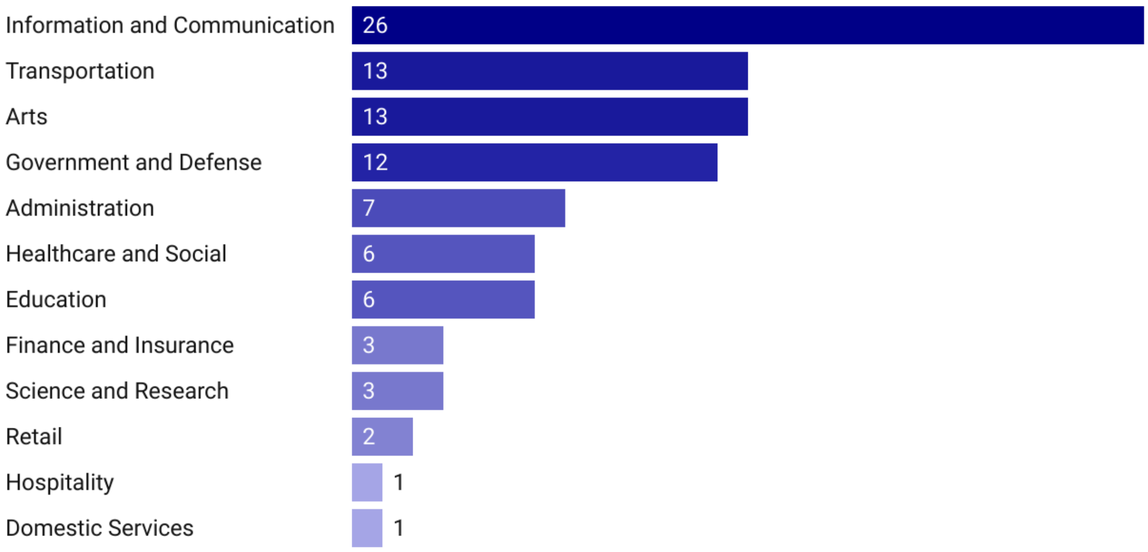


Chart: By Lucia de la Torre • Source: AI Incident Database • Created with Datawrapper

These gaps shaped the scope of the analysis: they don't diminish the value of the AI Incident Database as a growing archive of AI incidents, but they do mean that statistical trends by geography or sector reflect the subset of entries where those details are provided. As McGregor puts it, even this partial dataset is instructive:

"The [AIID] is a mirror, but a warped one. It reflects incidents logged, not incidents occurred, amplifying wherever there's a keen observer and a policy environment that encourages disclosure."

In this sense, the AI Incident Database is as much about who counts as it is about what's counted.

Others are trying to make sense of AI failures with more organized research. For example, AI researcher Jamie Bernardi helped MIT create The MIT AI Risk Repository, a risk database that builds on the cases collected by the AI Incident Database. What MIT is doing is similar to what transportation agencies do when they analyze plane or car crashes – they look for patterns in how and why things go wrong, and how serious the impact is. The MIT database sorts problems by what part of society they affect, what caused them, and how much harm they did. The goal is to help researchers and policymakers spot trends and think about solutions.

But as Bernardi and his team have found, even the best-organized research runs into some of the same challenges: the information is still patchy, and it's difficult to compare incidents when details are missing or reported differently each time.

“We wanted to move beyond anecdotes and create a tool that could classify AI incidents by risk domain, causal factors, and harm,” Bernardi explains. “But the hard part is standardization: every domain, from health to elections, has unique risks.”

The Challenge of Definitions and Legal Frameworks

Suresh Venkatasubramanian, a Brown University professor and former White House AI advisor, often points out that AI regulation stalls on disagreements over what counts as harm. Is it when an ad algorithm excludes someone from a job based on race? When a credit scoring model wrongly denies a mortgage? When an autonomous car narrowly avoids hitting a pedestrian?

These cases also raise the question of who should be held accountable: the company that built the AI model, or the organisation that chose to deploy it? In healthcare, for example, should errors fall on the developer of a diagnostic tool or the hospital using it? With AI embedded across finance, policing, medicine, and more, there's no single regulatory approach that fits all. The diversity and ubiquity of AI technologies demand flexible frameworks that share responsibility between developers and users, tailored to context and sector.

As challenging as finding a unified framework is, global organizations have begun groping for consensus. A recent publication by the Organisation for Economic Co-operation and Development (OECD) lays out 29 points on how to define, capture, and share AI failures. Its hope: that harmonized rules will let researchers and policymakers compare incident patterns across borders and industries. In the EU, the new AI Act and the Digital Services Act introduce formal requirements to disclose, and in some cases, prevent, certain types of AI failures in high-risk areas.

But laws on paper don't always change what happens in practice. In the U.S., there's still no nationwide system for tracking problems or mistakes involving AI, which means oversight is often inconsistent and depends on what state you are in. Congress has discussed bills like the AI Incident Reporting and Security Enhancement Act (H.R.9720) – meant to create voluntary standards for reporting AI mishaps – but nothing like this has become law yet.

Some states have stepped up on their own. California has passed the AI Transparency Act (SB 942), which makes companies disclose when content is AI-generated and follow certain rules. New York approved the Responsible AI Safety and Education Act (RAISE

Act, S6953B), which requires big AI companies to report and plan for major AI incidents – though some protections, like those for whistleblowers, didn't make it into the final version.

So while California and New York now have laws on AI transparency and safety, most states do not. Without a national law, the rules on how AI problems are reported and addressed are still uneven across the country.

Accountability's Dead Ends

The promise of reporting is that mistakes won't be repeated. Yet the legal system, as the ultimate arbiter of responsibility, adds another layer of complexity to the problem. Individuals harmed by AI often face significant legal challenges that make seeking justice difficult or nearly impossible.

Take the United States. Under employment law (e.g., Title VII), proving discrimination means a plaintiff must first show a "disparate impact" (that a group was harmed more than another) – a tall order in a black-box regime. AI systems are frequently proprietary; their decision processes are shielded by trade secrets. Legal discovery moves at snail's pace, if at all, and companies routinely argue the code, data, or entire system isn't theirs but leased or licensed from a third party.

Sarah Cen, a postdoctoral researcher at Stanford University who specializes in machine learning, law, and AI accountability, outlined what happens next:

“One of the biggest problems is the burden of proof. Typically, a plaintiff must produce evidence of a less discriminatory alternative algorithm. This is a huge, often impossible demand, because rebuilding and testing these systems requires expertise, access to data, and computational resources that most claimants simply don’t have. As a result, companies rarely have to provide transparency or allow outside scrutiny.”

Yet even when transparency is mandated, Cen observes that progress is often incremental and met with skepticism. “There is potentially some so-called ‘transparency fatigue’ happening – every time there’s a new transparency push, you hear the counter that it doesn’t really lead to meaningful accountability. People start to wonder if all this reporting and disclosure is changing outcomes, or just piling up unread documentation.”

The consequence, she reflects, is an ecosystem where reports of harm are sparse, successful legal cases even rarer, and meaningful reform slow to materialize.

Journalism As The Last Line of Defense

When regulatory bodies and courts struggle to keep pace with rapidly evolving AI technologies, much of the responsibility for documenting and publicizing harm falls to journalists and civil society organizations. Through detailed investigations, freedom of information (FOIA) requests, and collaboration with affected communities, these actors work to connect isolated incidents into larger patterns that might otherwise remain invisible.

William Owen, Communications Director at the Surveillance Technology Oversight Project (S.T.O.P.) in New York, an organization devoted to providing legal help to those affected by rights violations by facial recognition technologies, describes the painstaking effort involved:

“We maintain our own records of facial recognition and surveillance abuses because public agencies often don’t. We rely on community tips, whistleblower disclosures, FOIA requests, and lawsuits to gather pieces of what is often a larger, hidden story.”

Owen notes that tracking instances of AI-related harm requires a network of committed individuals and organizations. Yet, outside well-resourced cities, these networks are often fragile or absent:

“In places with limited press freedom or fewer watchdog groups, reports are scarce or non-existent. This creates significant gaps in understanding how AI technologies are deployed and the harms they cause globally.”

Organizations such as Amnesty International, S.T.O.P., and legal clinics compile incident files that highlight systemic risks like discriminatory policing or biased welfare algorithms. Still, their efforts are constrained by resource limits, time, and often that the most vulnerable voices rarely reach the public eye.

Lessons from Medicine: Can AI Learn from VAERS?

Healthcare offers an instructive example of how incident reporting can surface emerging hazards. The U.S. Vaccine Adverse Event Reporting System (VAERS) allows health professionals and the public to report suspected vaccine side effects, enabling timely detection of rare but significant risks – such as myocarditis among young men during the COVID-19 vaccination campaign.

Deborah Raji, a PhD student at UC Berkeley, computer scientist and activist known for her work on AI fairness and ethics, has drawn parallels to this model:

“Incident databases provide a unique means to capture harms that only become apparent after wide deployment. They’re essential for ongoing learning – many AI harms cannot be fully understood until after real-world use.”

Raji points to current limitations in corporate approaches to incident data:

“While companies sometimes solicit feedback through red teaming or user reports, many ‘incidents’ are treated as complaints or misuse of terms of service, not as evidence of systemic harm. This leads to underreporting and missed opportunities to detect risks to vulnerable groups.”

Facing the Gaps: The Stakes of What Goes Unseen

As AI systems become embedded in decisions about employment, healthcare, criminal justice, and social services, a critical gap in oversight is becoming apparent. Unlike aviation, medicine, and other high-stakes industries that maintain comprehensive incident tracking, AI operates without systematic safety mechanisms to detect and prevent repeated failures.

The consequences are already visible. Robert Williams spent 30 hours in jail because of faulty facial recognition. Tammy Dobbs lost essential care due to an opaque Medicaid algorithm. Thousands of Dutch families were wrongly accused of fraud by an AI system that targeted immigrants and low-income parents. These cases made headlines, but researchers acknowledge they represent only a fraction of actual incidents.

Current tracking efforts remain fragmented and incomplete. The AI Incident Database, while valuable, captures incidents that get reported rather than incidents that occur. Geographic and sectoral gaps mean vast areas of AI deployment go unmonitored. Legal remedies prove difficult when companies shield their systems behind trade secrets and victims lack the technical expertise to challenge algorithmic decisions.

Without systematic oversight, the field risks repeating the same mistakes that have already cost people their freedom, care, and livelihoods. Other high-stakes industries offer proven models: aviation’s mandatory incident reporting has prevented countless crashes, while healthcare’s adverse event systems have identified emerging drug risks before they became widespread. As AI deployment accelerates across sectors from healthcare to criminal justice, the absence of comprehensive tracking mechanisms becomes increasingly consequential.

The technology industry has often moved fast and fixed problems later. With AI systems now making decisions about who gets hired, who receives medical care, and who faces police scrutiny, this approach carries risks that extend far beyond individual companies. Building the institutional capacity to track, analyze, and prevent AI-related harm may prove as important as the technological advances themselves.

The question facing policymakers, technologists, and society is not whether AI will cause unintended consequences, but whether institutions will develop the tools necessary to detect them before they become widespread.

Source List

Interviews

- Sean McGregor, Founder, AI Incident Database (smcgregor@seanbmcgregor.com)
- Jamie Bernardi, Contributor, MIT AI Risk Repository (jamie@jamiebernardi.com)
- Serena Booth, Computer Science Professor, Brown University (serena_booth@brown.edu)
- Sarah Cen, Postdoctoral Researcher, Stanford University (shcen@stanford.edu/cen.sarah@gmail.com)
- Deborah Raji, Computer Scientist and Activist, Mozilla Fellow, PhD Student at UC Berkeley (deborahraji1@gmail.com)
- William Owen, Executive Director, Surveillance Technology Oversight Project (S.T.O.P) (+12404767866)

Data Sources

- AI Incident Database

Methodology

Detailed information on methodology is available in the GitHub Repo's ReadMe (link below). This analysis employed a systematic examination of the AI Incident Database (AIID), utilizing the full database snapshot in CSV format containing detailed records of AI failures and incidents reported globally across various sectors and years. The data processing methodology involved cleaning and standardizing inconsistent date, location, and sector formats, followed by temporal analysis to identify reporting patterns over time, geographical analysis to map incident distribution by country and US state, and sectoral breakdown to examine industry-specific patterns. The analysis was conducted using Python in Jupyter notebooks, with data manipulation performed using pandas.

GitHub Repo

- lg3394/aiincidenttracking

Postscript: How The Story Came Together

When I first set out to write my master's thesis, I was drawn to the topic of AI data centers and water scarcity. However, in early May, Bloomberg published a comprehensive interactive data piece titled "How AI Is Draining Water From Areas That Need It Most." I realized I would have very little unique to add to the topic, and for me, a priority with the master's thesis was finding an underreported area in the AI field where I could make an original contribution. I decided to shift focus toward AI accidents and incident reporting.

Looking for a data story, I was inspired by the AI Incident Database. I initially thought I could analyze vast amounts of AI incident data and write a story based on observed trends. However, a quick analysis of the dataset revealed that the data was too insufficient for that approach – there were simply too many gaps. After discussing this with my advisor, Steve Eder, I realized the lack of data itself could be a story: what is needed for better AI incident reporting?

I then set out to speak with experts in the field. This summer, while working on the thesis, I was fortunate to attend Brown University's AI policy summer school and the Aspen Institute's Ideas Festival, which put me in touch with a lot of experts in the field and broadened my understanding of the legal, social, and political complexities that shape how and whether these incidents are tracked. I was able to interview people in the academic, advocacy, and legal fields.

However, one big challenge was making this piece journalistic rather than report-style, and engaging, given the technical nature of the topic. Another of the most challenging aspects was reaching out to people directly affected by AI incidents. These conversations would have humanized the story, put a face to the issue, and added nuance and urgency that data alone cannot provide. I wasn't able to speak to anyone directly, given time constraints, but I believe the work would have really benefited from them. Gathering those voices remains an important goal for me beyond this thesis.

I am proud that this thesis became not just a data analysis of AI failure trends but a broader examination of the complex world of preventing AI accidents. In the future, I would prioritize reaching out to affected people from the start. I now plan to expand on this story through reporting, listening, and building toward a more complete, inclusive record of AI's real-world harms with the voices of those affected at the center.