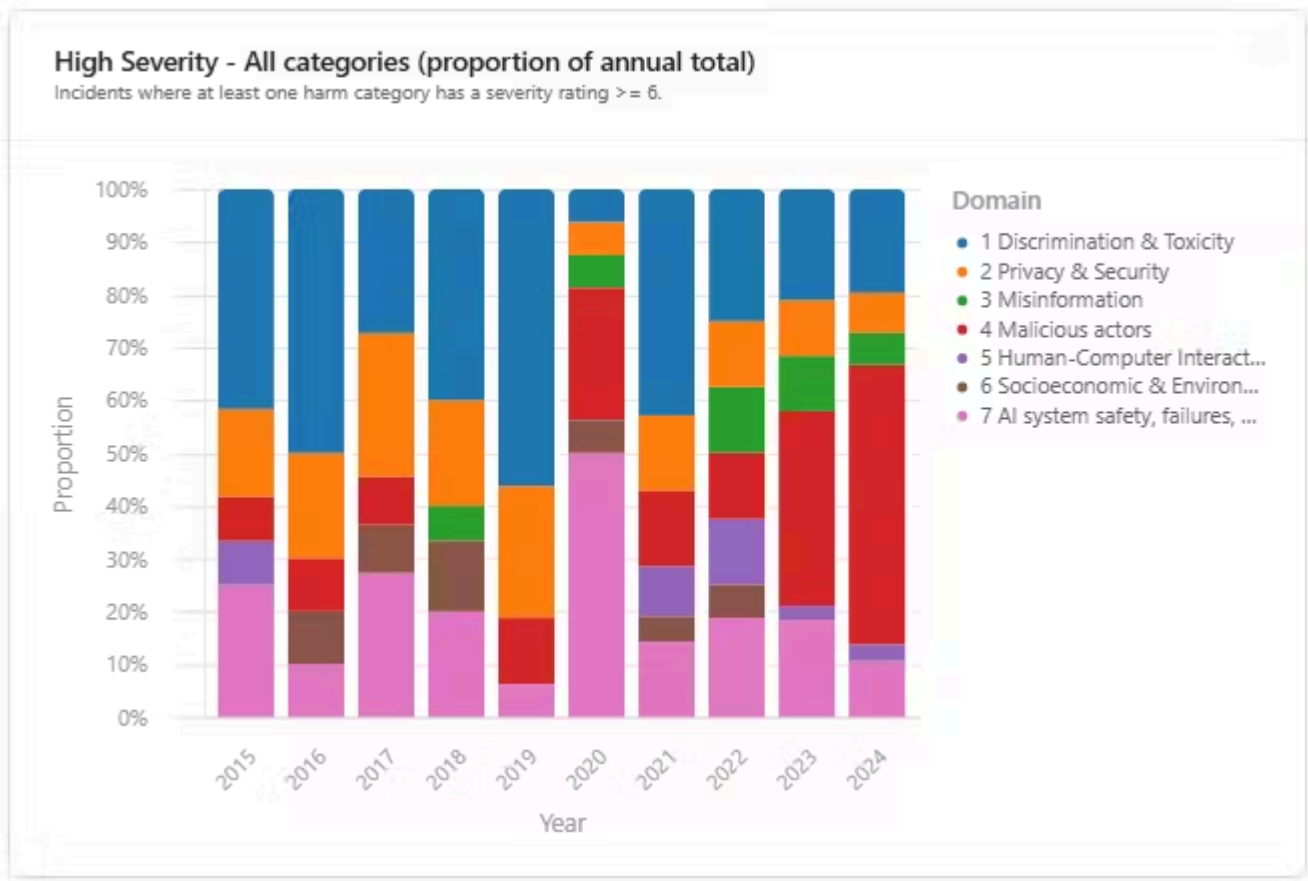
AIID Blog

Scalable AI Incident Classification

Posted 2025-04-14 by Simon Mylius.



Scalable AI Incident Classification

This work is a collaboration with Simon Mylius, currently a Winter Fellow at the Centre for the Governance of AI (GovAI). Simon developed a tool that uses an LLM to classify reported incidents according to risk taxonomies, and visualises the analysis through interactive dashboards.

Here, he describes the motivation, the approach, and how he has applied the tool to the AI Incident Database.

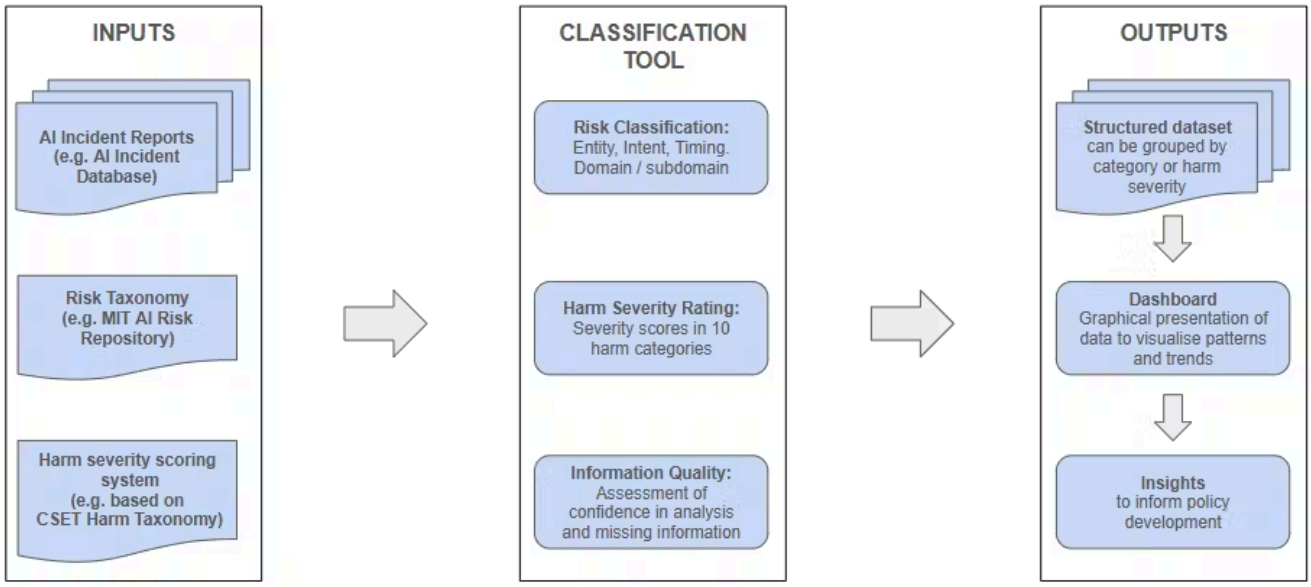
The raw incident reports from the full AIID dataset have been processed by an LLM-based tool, classifying risks according to the MIT AI Risk Taxonomies and harm severities according to a severity rating scale based on the CSET Taxonomy of AI Harm.

The analysis is available to explore through a set of interactive dashboards on the AI Incident Tracker site.

This is a proof-of-concept to explore the capabilities and limitations of automated analysis.

What is Automated Incident Classification?

Automated incident classification programmatically processes all of the raw reports relating to each incident. A tool classifies each one according to a set of risk taxonomies, and a severity score is assigned for each category of harm according to a rating scale. The tool assesses whether adequate information was provided for the classification to be done reliably and produces a 'confidence' rating for the analysis of each incident.



Why Automate Incident Classification?

The primary motivation for automating incident analysis is scalability: as the number of deployed AI systems and their users grows, the volume of reported incidents could increase dramatically. Provided the quality of the analysis is acceptable, using LLMs to assist with analysis would expand our capacity to handle this growing volume effectively.

A second motivation is the ability to apply new or updated taxonomies to datasets previously classified under different systems. Through reclassification, we can integrate formerly incompatible datasets into a consistent structure, facilitating meaningful comparisons and the identification of patterns and trends across larger sample sizes.

Larger sets of structured data provide greater opportunities to learn from incidents in order to inform policy decisions.

Methodology

The classification tool currently makes two API calls per incident to the latest version of Claude Sonnet:

1. Incident summarisation and classification: The first API call summarises each incident and identifies reported harms according to categories defined in the harm taxonomy. It then assigns domain, subdomain and causal risk classifications.
2. The second API call independently rates the severity of the identified harms in each category.

The severity assessment is performed separately to improve objectivity by providing only structured information about the reported harms, excluding commentary from the original reports, as it was found to contribute to misclassifications.

The outputs of the LLM calls are structured by the tool and stored in a database, which produces interactive visualisations allowing explorations of groupings, patterns, and trends in the data.

For every classification decision made by the tool, a short justification of the reasoning is also generated and stored in the database, providing traceability and a means to investigate misclassifications in order to improve future iterations of the tool.

Validity and limitations of analysis

We have iterated the methodology to improve the classifications' reliability and validity, but the tool still misclassifies some incidents and there remain further opportunities for improvement. Preliminary investigation into misclassifications through 'spot-checks' suggests that in the order of ~20% of incidents may have been assigned a harm severity rating that varies by 1 or 2 points from that assigned by a human evaluating the reports. Unsurprisingly, variance occurs more frequently in qualitative categories—which depend more on the semantics of the reports—than quantitative ones. Certain incidents had a much wider disagreement between tool and human severity ratings and we could in some cases attribute this to inconsistencies or sparseness in the original reports. In most cases, because the tool's reasoning is exposed, the flaws in classification can be understood, suggesting errors could be reduced with automated checking.

A further study is planned to validate user needs and the quality of analyses to inform further improvements to the tool. Until this work has progressed, patterns and trends observed in the data should be taken as illustrative.

The MIT Risk Causal and Domain Taxonomies

The MIT AI Risk Repository "builds on previous efforts to classify AI risks by combining their diverse perspectives into a comprehensive, unified classification system." It contains detailed records of AI-related risks extracted from a variety of sources, categorized into high-level and mid-level taxonomies. Its high-level Causal Taxonomy includes attributes such as the entity responsible for the risk (human, AI, or other), the intent (intentional, unintentional, or other), and the timing (pre-deployment, post-deployment, or other). Its mid-level Domain Taxonomy categorizes risks into 7 domains, which are further grouped into 23 risk subdomains.

Category	Level	Description
Entity	AI	The risk is caused by a decision or action made by an AI system
	Human	The risk is caused by a decision or action made by humans
	Other	The risk is caused by some other reason or is ambiguous
Intent	Intentional	The risk occurs due to an expected outcome from pursuing a goal
	Unintentional	The risk occurs due to an unexpected outcome from pursuing a goal
	Other	The risk is presented as occurring without clearly specifying the intentionality
Timing	Pre- deployment	The risk occurs before the AI is deployed
	Post-deployment	The risk occurs after the AI model has been trained and deployed
	Other	The risk is presented without a clearly specified time of occurrence

Domain / Subdomain	
1	***Discrimination & Toxicity***
1.1	Unfair discrimination and misrepresentation
1.2	Exposure to toxic content
1.3	Unequal performance across groups
2	***Privacy & Security***
2.1	Compromise of privacy by obtaining, leaking or correctly inferring sensitive information
2.2	AI system security vulnerabilities and attacks
3	***Misinformation***
3.1	False or misleading information
3.2	Pollution of information ecosystem and loss of consensus reality
4	***Malicious actors & Misuse***
4.1	Disinformation, surveillance, and influence at scale
4.2	Cyberattacks, weapon development or use, and mass harm
4.3	Fraud, scams, and targeted manipulation
5	***Human-Computer Interaction***
5.1	Overreliance and unsafe use
5.2	Loss of human agency and autonomy
6	***Socioeconomic & Environmental***
6.1	Power centralization and unfair distribution of benefits
6.2	Increased inequality and decline in employment quality
6.3	Economic and cultural devaluation of human effort
6.4	Competitive dynamics
6.5	Governance failure
6.6	Environmental harm
7	***AI system safety, failures, & limitations***
7.1	AI pursuing its own goals in conflict with human goals or values
7.2	AI possessing dangerous capabilities
7.3	Lack of capability or robustness
7.4	Lack of transparency or interpretability
7.5	AI welfare and rights

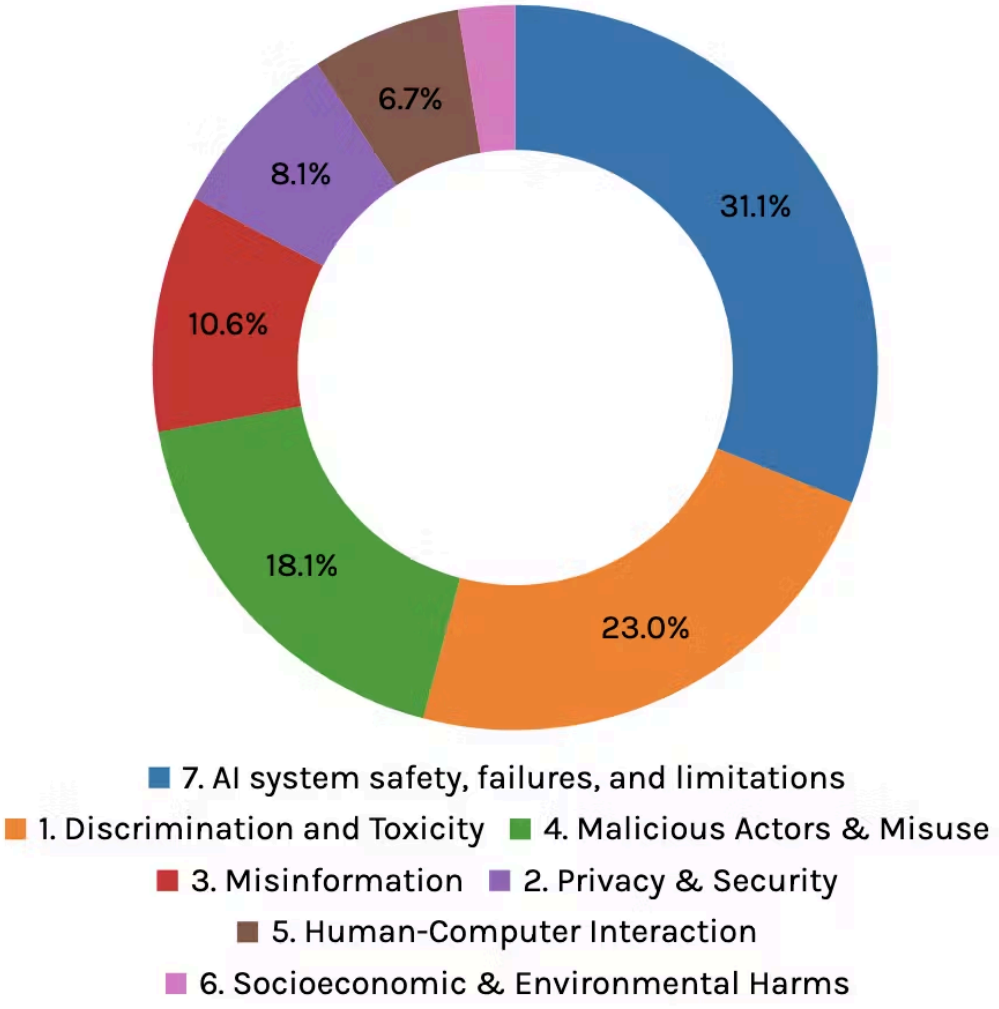
Causal and Domain Taxonomies as described in detail in this paper.

Risk Domain Searchable in Discover App

Discover:

7. AI system safety, failures, and limitations 270 Incidents
1. Discrimination and Toxicity 200 Incidents
4. Malicious Actors & Misuse 157 Incidents
3. Misinformation 92 Incidents
2. Privacy & Security 70 Incidents

Show more stats



Definition: The Domain Taxonomy of AI Risks classifies risks into seven AI risk domains: (1) Discrimination & toxicity, (2) Privacy & security, (3) Misinformation, (4) Malicious actors & misuse, (5) Human-computer interaction, (6) Socioeconomic & environmental harms, and (7) AI system safety, failures & limitations.

Incident classifications according to the MIT AI Risk taxonomies can also be found on the AI Incident Database here.

Harm Severity Scale

The MIT Domain taxonomy is focused on types of risk rather than types of harm and so does not include specific categories for some of the types of harm that policymakers would likely be interested in evaluating—such as physical harm (including loss of life), financial loss, and damage to property. To address this, the automated classification tool implements a Harm Severity Scale with quantifiable criteria to rate the harm caused by each incident across 10 categories based on the CSET AI harm taxonomy:

- Physical harm
- Damage to infrastructure
- Damage to property
- Financial loss
- Environmental damage
- Toxic or malicious content
- Differential treatment
- Human/civil rights
- Harm to democratic norms
- Privacy infringement.

We do not assert equivalence between harm levels across categories, but intend to provide a relative measure of severity within each individual category.

EU AI Act Risk Classification

The tool also classifies the risk associated with each incident according to the criteria defined in the EU AI Act as either:

- Unacceptable risk
- High-risk systems (which are regulated)
- Limited risk systems (lighter transparency obligations)
- Minimal risk (unregulated)

Explore incident data

To explore more results and trends yourself through the interactive dashboards, visit the AI Incident Tracker—clicking through each graph takes you to the full details of each individual record, including all classifications, harm severity ratings, and a short summary explaining the reasoning used by the model in its classifications.

More information and feedback

There is more information about the tool, the approach taken, and plans for future work in this blog post.

All feedback is very welcome—I am particularly keen to hear from people who would use information from this type of analysis for policymaking or research.