Issue Brief

Adding Structure to AI Harm

An Introduction to CSET's AI Harm Framework

Authors
Mia Hoffmann
Heather Frase

CSET CENTER *for* SECURITY *and* EMERGING TECHNOLOGY

July 2023

Executive Summary

Harms from the use of artificial intelligence systems (“AI harms”) are varied and widespread. Monitoring and examining these harms (AI harm analyses) are a critical step towards mitigating risks from AI. Such analyses directly inform AI risk mitigation efforts by improving our understanding of how AI systems cause harm, enabling earlier detection of emerging types of harm, and directing resources to where prevention is needed most.

This paper introduces the CSET AI Harm Framework, a standardized conceptual framework to support and facilitate analyses of AI harm. This framework improves the comparability of harm monitoring efforts by providing a common foundation that consistently identifies AI harms, while providing modularity to adapt to different analytical needs.

The CSET AI Harm Framework lays out the key elements required for the identification of AI harm, their basic relational structure, and definitions without imposing a single interpretation of AI harm. Specifically, this framework:

- **Defines “AI harm” as when an entity experiences harm (or potential for harm) that is directly linked to the behavior of an AI system.**
- **Groups harm into either tangible or intangible harm.** Tangible harm is harm that is observable, verifiable, and definitive. Intangible harm is harm that cannot be directly observed or does not have any material or physical effect. Because of its observability, tangible harm is inherently easier to detect and identify. This means that tangible harm data is more consistent, less noisy, and easier to analyze.
- **Allows users to define additional categories of tangible and intangible harm.** The CSET AI Harm Framework provides some common categories of harm, such as harm to physical health or safety, financial loss, property damage, detrimental content, bias and differential treatment, and violation of privacy, human and civil rights, or democratic norms. This framework also allows for the inclusion of new categories since new harm types could emerge in the future or be more relevant in another incident data-source.

- **Distinguishes harm that actually occurred from harm that may occur.** Parsing and differentiating between harm that occurred and may not occur allow for the tracking of realized harms, while also enabling research and analysis on potential harms that are risks and vulnerabilities.

In addition to providing introductions to the definitions and concepts of the CSET AI Harm Framework, this report also:

- **Discusses how users can adapt the framework.** In order to apply the framework to data, users should create a customized framework. This requires specifying the framework's components to such a degree that it can be used to extract all the information needed to identify and characterize AI harms according to the user's analytic interests and the limitations of the data source.

- **Provides an example customized framework.** As an example, this report shows how the CSET AI Harm Framework was customized for use in the CSET AI Harm Taxonomy for AI Incident Database (AIID). Since modifications and definitions are centrally documented in the CSET taxonomy, database users are able to retrace the underlying framework and compare it to other taxonomies built on the CSET AI Harm Framework.

- **Details future additions to the framework.** Future versions of the CSET AI Harm Framework will incorporate content on the severity and spread of AI harm. When combined, these factors can inform our understanding of the aggregated impact of a particular harm.

Table of Contents

Glossary

AI harm: An AI harm occurs when an entity experiences a harm event or harm issue that can be directly linked to a consequence of the behavior of an AI system.

Direct link to AI: CSET's definition of AI harm requires a clear chain of events through which the AI is linked to the tangible or intangible harm. It is not sufficient that the AI is part of a system that caused harm. The AI functionality itself must be linked to the harm and the harm would not have occurred without the behavior of the AI.

Entity: An entity is a person, place, or thing. Common entities involved in AI incidents are individuals, groups of people, companies, locations, products, infrastructure, government agencies, or the natural environment.

Harm event: An entity experienced a harm event when harm definitely occurred.

Harm issue: An entity experienced a harm issue when harm did not occur, but there is a reasonable probability that it could have occurred.

Harm near-miss: An entity experienced a harm near-miss when harm did not occur, but there was an imminent potential for harm; harm near-misses are a subset of harm issues.

Intangible harm: Intangible harm is harm that is typically not observable, including psychological harm, pain and suffering, and damage to intangible property.

Tangible harm: Tangible harm is harm that is material in nature, and therefore observable, verifiable, and definitive. Common categories of tangible harm are financial loss, physical injury and damage to property or the environment.

Taxonomy: Taxonomies classify, document and organize information, in this case about harmful incidents involving AI systems. Users of the CSET AI Harm Framework can customize the framework's structure to develop their own AI harm taxonomy. In this paper, the term taxonomy is often used interchangeably with "customized framework."

Introduction

We are presenting the CSET AI Harm Framework, a customizable framework to facilitate analyses of harms from artificial intelligence. This framework is based upon extensive experience annotating AI incidents. It supports data aggregation across different databases and efforts, which can improve the AI community's ability to understand AI harm. It is database-agnostic and modular, allowing it to be customizable and support a variety of analytic goals. At the end of the document, we provide an illustrative example of customization by describing the CSET AI Harm Taxonomy for the AI Incident Database (AIID). As the number of AI harm data-sources and taxonomies increases, a structured approach that allows for combining data becomes more useful.

There are several publicly available data-sources for AI incidents, including:

- the AI Incident Database,[1]
- the AI, Algorithmic, and Automation Incidents and Controversies repository (AIAAIC[2]),
- the AI Vulnerability Database[3] (AVID), and;
- the Emerging Technology Institute's AI Litigation Database.[4]

The Organisation for Economic Cooperation and Development (OECD) has also recently announced the development of a global AI incidents monitor (AIM).[5] With the exception of the AIM, these databases are maintained by small, non-profit organizations or even groups of volunteers interested in promoting transparency around the use, risks and harms of artificial intelligence.

Taxonomies, though not particularly common, can improve the utility of incident databases such as those listed above. Taxonomies classify, document and organize things, which in this case is AI harm data. Different taxonomies can be applied to the same incident database. For instance:

- Three taxonomies have been applied to AIID: the Goals, Methods, and Failures (GMF[6]) taxonomy and two editions of the CSET AI Harm Taxonomy.[*]
- AIAAIC currently provides high-level structured information and has announced the launch of a harm taxonomy development process starting in June 2023.[7]
- AVID has developed a high-level taxonomy that categorizes cases into issues of security, ethics and performance, building on MITRE's work for adversarial attacks and cybersecurity.

Importance of Understanding AI Harm

Harms caused by the deployment and use of artificial intelligence systems ("AI harms") are varied and rapidly increasing as AI systems proliferate across different sectors of the economy and society.[8] Because AI is a general-purpose technology that can be used for a wide range of tasks in varied contexts, the types of harms that these systems create are multi-faceted. Autonomous driving accidents,[9] privacy violations,[10] wrongful incarceration,[11] biased healthcare decisions,[12] flawed student evaluations,[13] discriminatory hiring[14] and digital sexual violence[15] illustrate some of the harms in which AI has been implicated. Tracking efforts by AIID and AIAAIC suggest that the number of harms experienced in relation to AI systems has grown rapidly over the past 5-10 years.

A better understanding of the range of harms linked to the use of AI systems, and the mechanisms that contribute to their occurrence is critical to producing AI systems that are less harmful and to learning how to use them more safely.

Difficulty of Defining, Categorizing, and Combining Harm Data

Analyses of AI harm depend on reliable data on harm incidents. Combining data on AI harm is difficult because there are many possible interpretations of AI harm and data sources often gather different information and code it differently. Definitions of harm

[*] The CSET AI Incident Taxonomy, CSET's first edition taxonomy for annotating incidents, was created in 2021 and 100 annotations using this taxonomy are available on AIID. The second edition of this taxonomy is called the CSET AI Harm Taxonomy for AIID.

often depend on local laws, societal norms, and communal experiences, which means that interpretations of harm will differ across individuals, organizations and governments.

As a result, efforts to track AI harms tend to fragment, as the incident repositories AIID and AIAAIC illustrate. While they both show a steep upward trend in incident occurrences over the recent past, the absolute numbers in the two databases vary substantially. This is likely due to definitional and interpretative differences of what constitutes an AI harm and different methods for collecting harm data, leading to data that may not be easily compatible. Analyses of harm incidents from different data sources may therefore produce recommendations that are not comparable—potentially without the ability to understand the source of the differences—making comprehensive tracking and a shared understanding of the problem difficult.

The CSET AI Harm Framework Facilitates Information Extraction

A common approach to identifying and characterizing harms from AI would facilitate comparability between different analytical efforts. Transparency as to how AI harm is defined, and how different incidents are categorized allows third-parties to adopt similar concepts and clarify where efforts diverge. We are proposing the CSET AI Harm Framework to equip people and organizations with a common approach to categorization and comparison.

The CSET AI Harm Framework facilitates the structured characterization of harm incidents and their circumstances. Its use enables a consistent identification and accounting of AI harm, which fosters a shared understanding and raises awareness of the prevalence and pervasiveness of AI harm among policymakers, analysts, AI developers, and the general public. As a starting point for taxonomy development, it serves as a baseline for targeted analyses of AI harm and the varied socio-technical circumstances from which it emerges.

Organizations can create and document customizations of the AI harm framework that reflect their analytic needs and data sources. By starting with this AI harm framework as the common underlying conceptual structure, individual customizations may be mapped to each other or back to the core. As long as the documentation for any customization details which framework components they maintain and drop and their corresponding—and where possible, common—definitions, it will be clear where and when customizations are interoperable. This is central to how the CSET AI Harm Framework facilitates combining and sharing AI harm data.

CSET AI Harm Framework Development Process

The CSET AI Harm Framework is based on discussions with outside organizations and experience annotating about 100 AI incidents in the AIID. During the year-long annotation process, we increased our understanding of the variety and complications of incidents, identified the core elements of AI harm and refined definitions. This resulting framework is centered on creating actionable, analyzable data from real-world incident reports. Additionally, CSET participated in discussions with the Responsible AI Collaborative, MITRE, the Organisation for Economic Co-operation and Development, the National Institute of Standards and Technology (NIST), and O'Neil Risk Consulting & Algorithmic Auditing (ORCAA)[16] about characterizing, identifying and tracking AI harm and risk. When possible, the framework aligns its terminology with those used by these organizations.

Building the CSET AI Harm Framework

In this section, we step through each framework component, describing how they combine into the final CSET AI Harm Framework. We first build a structure for categorizing harm in general. Then we add framework components that specifically identify "AI harm."

The components of the framework and their structure are derived from the review and assessment of numerous AI incidents. They reflect our best effort to balance two competing interests: the wish to add detail in order to best capture the variety of factors that describe an incident; and the wish to group and aggregate similar instances of AI harm. Every element of the framework was evaluated according to its added value along both dimensions, and the following section explains our rationale for its inclusion.

Structuring Harm

The generic harm framework divides harm into two high-level categories, tangible and intangible. Within these high-level categories are more specific subcategories. We provide some common subcategories, but organizations can easily create their own. We also add levels of harm realization, capturing the difference between harm that definitively occurred ("harm events") and potential for harm ("harm issues").

High-level Harm Types: Tangible and Intangible

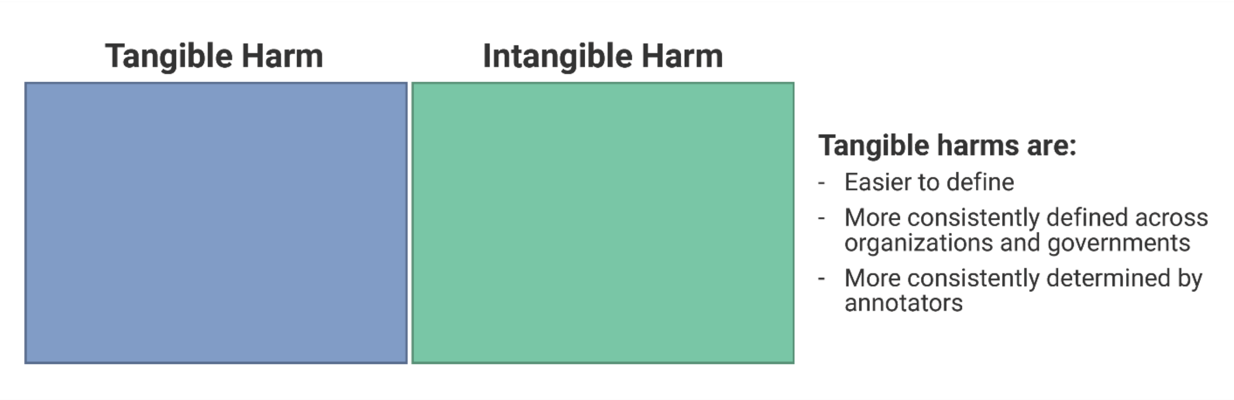
At a high level, harm can be divided into tangible and intangible harm (Figure 1). *Tangible harm* is harm that is material in nature, and therefore observable, verifiable, and definitive. A third party can observe tangible harm as it is happening, verify that harm happened even after the moment of its immediate occurrence, and judge with certainty whether harm did or did not occur. Common tangible harms include physical injury (including death), financial loss, and damage to or destruction of private or public property. Tangible harm is usually quantifiable, and may often be expressed in monetary terms. Examples of tangible harm include damage to a car, a person's broken arm, or a loss of income.

Tangible harm is harm that is material in nature, and therefore observable, verifiable, and definitive.

Intangible harm generally cannot be directly observed, but may have observable consequences.

In contrast, *intangible harm* generally cannot be directly observed. That is, while the event causing the harm may be observable, and its effects and consequences may be expressed in observable ways, the harm in itself usually is not. Intangible harm can include, but is not limited to, mental/psychological harm, pain and suffering, harm to intangible property (for example, IP theft, damage to a company's reputation), and loss of trust or belief.

Figure 1. Harm Is Divided into Two Main Groups: Tangible and Intangible



Tangible harms are:

- Easier to define
- More consistently defined across organizations and governments
- More consistently determined by annotators

Source: CSET AI Harm Framework.

Our framework separates tangible and intangible harms because tangible harms are more consistently defined and identifiable by different groups of people. Because tangible harms are usually of material nature their occurrence is more difficult to dispute. For the same reasons, it is easier to distinguish between realized harm and potential harm when considering tangible harm incidents than intangible harm incidents.[*] In contrast, whether or not intangible harm occurred or could occur is often subjective, depending on personal perspectives.[17] Reasonable people often disagree about what constitutes intangible harm.

Box 1. Example of Harm Issue

Consider the example of a child seeing an age-inappropriate, violent video on YouTube. While a third party could see the problematic content and observe the child’s tearful response, the harm itself is not observable. It could occur even if the child does not start to cry or nobody is around to see the video. This means that especially after the fact, intangible harm is hard to verify, in marked contrast to, for example, a broken arm.

By distinguishing tangible and intangible harms, the framework sections out harm that more people and organizations can agree upon, which allows for more consistent

[*] See further discussion in tangible harm types and imminency levels.

accounting of those harms. Consistency improves the quality of collected information and facilitates data sharing across organizations.

Harm is either tangible or intangible—not both. It is also possible for tangible harm to result from intangible harm. For instance, psychological harm can result in medical treatment that results in financial loss. Misinformation or IP theft can lead to legal or civil actions that result in fines or monetary damages.

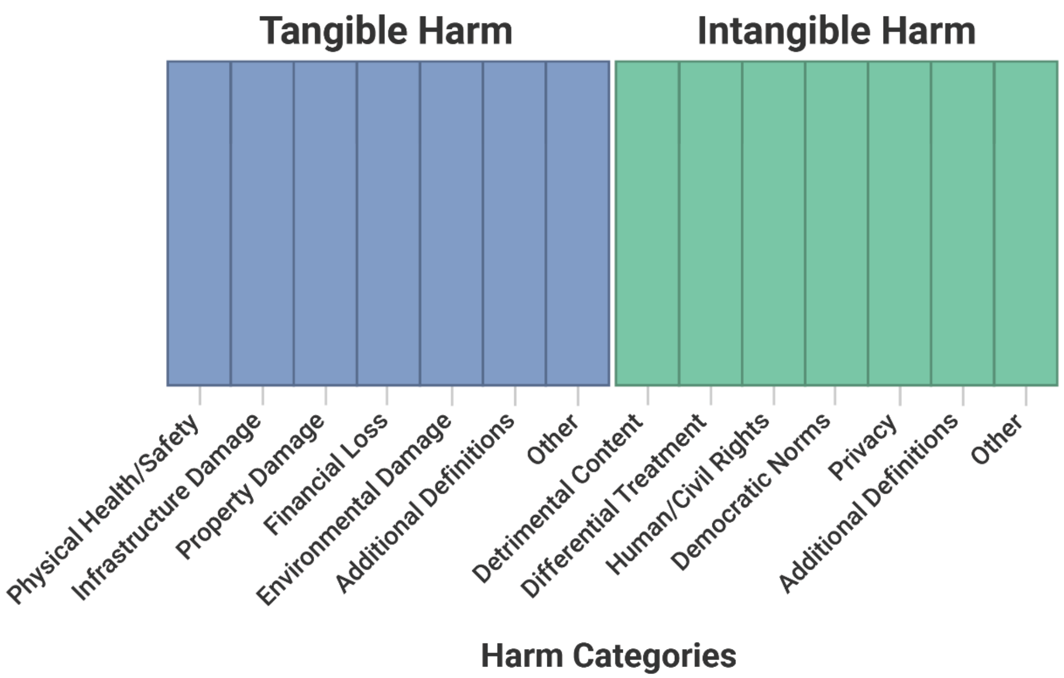
Box 2. Example of Multiple Harms

Multiple harms can occur simultaneously and be a mix of tangible and intangible harm. For example, in 2018, investigative journalists discovered that the Dutch Tax Authority had been using a risk prediction system to detect welfare fraud that wrongfully accused thousands of families of defrauding child care benefits.[18] In addition to terminating benefits payments and plunging thousands of households into debt (a tangible harm in the form of financial loss), the model disproportionately accused immigrant families because it considered a second nationality a high-risk factor (an intangible harm in the form of disparate treatment).

Categories of Harm

The CSET AI Harm Framework divides tangible and intangible harms into categories. It provides some suggested harm categories and allows for adding custom categories (Figure 2). When customizing the framework, organizations should clearly define all harm categories, whether or not they are custom or suggested. Whenever possible, the definitions should point to common references or standard definitions. Because new types of harms continue to emerge, the framework includes an “Other’’ category to capture harms which do not fall neatly into one of the defined harm categories.

Figure 2. Tangible and Intangible Harms are Further Divided into Categories



Source: CSET AI Harm Framework.

Note: The tangible and intangible harm type "Other" reflects that no customization of the framework will likely be able to describe every type of harm present in a data-source.

Tangible Harm Categories

The suggested tangible harm categories are common and fairly easy to identify: harm to physical health or safety, financial loss and damage to property, infrastructure or the environment. There are many examples of these tangible harms involving algorithmic systems. Algorithmic trading systems have been involved in a number of stock and foreign exchange market crashes, and in 2019, a single deep fake-enabled scam defrauded several hundreds of thousand dollars from a UK firm.[19] One of the most public incidents in which an algorithmic system has been implicated to date are the crashes of the Boeing 737 Max airplanes in 2018 and 2019 that destroyed two aircrafts (damage to property) and killed 346 people (harm to physical health).[20]

Organizations interested in tracking other types of tangible harm can create additional categories. For example, a vehicle maintenance depot that uses AI to identify and prioritize repair needs could view time delays or time loss as a tangible harm.*

Intangible Harm Categories

The framework's suggested categories of intangible harm are: the creation and spread of detrimental content, bias and differential treatment, the violation of human or civil rights, harm to democratic norms and infringement of privacy.[21] These intangible harms are explicitly shown in Figure 2 because CSET annotators encountered them while reviewing incidents. They do not encompass a complete list of possible intangible harms. For example, reputational harm and psychological harm are two other types of intangible harm that could be used.

Our experience annotating and discussing AI incidents revealed that different communities, representing different backgrounds and values, may understand the meaning of harm categories differently. These differences tend to be smaller for tangible harms, with death—included under "harm to physical health or safety"—being the most consistently identified harm. Conversely, annotators more frequently disagreed in their assessments of whether intangible harm occurred. For example, one reviewed incident involved a Scottish soccer club that installed an AI-powered camera to broadcast its matches by automatically detecting and following the ball. However, during one game in October 2020, the camera continuously misidentified a referee's bald head for the ball and failed to broadcast the game's actual action. While some might consider this to be an inconvenience or disappointment, but not actual harm, others might argue that the club's fans or the referee experienced emotional distress and, therefore, harm.[22]

The framework is intentionally flexible to account for varying perspectives. Organizations that develop customized frameworks may choose to define categories specific to their needs. The above examples demonstrate how important clear definitions of harm categories are to the application of the framework. Organizations

* Because time delays often imply financial loss, it can be appropriate to designate delay as a financial harm in some contexts. In other contexts, however, there are harmful non-monetary consequences to time loss, for example when delays affect mission availability and readiness.

creating a customization must clarify, for example, whether their category for harm to physical health shall include minor injuries such as a scraped knee alongside more severe harms and death. Those interested in tracking disinformation harm must document what does and does not qualify as disinformation. Whenever possible, we encourage referencing commonly accepted and established definitions. When definitions of harm categories in different customizations align it enhances the interoperability of their data.

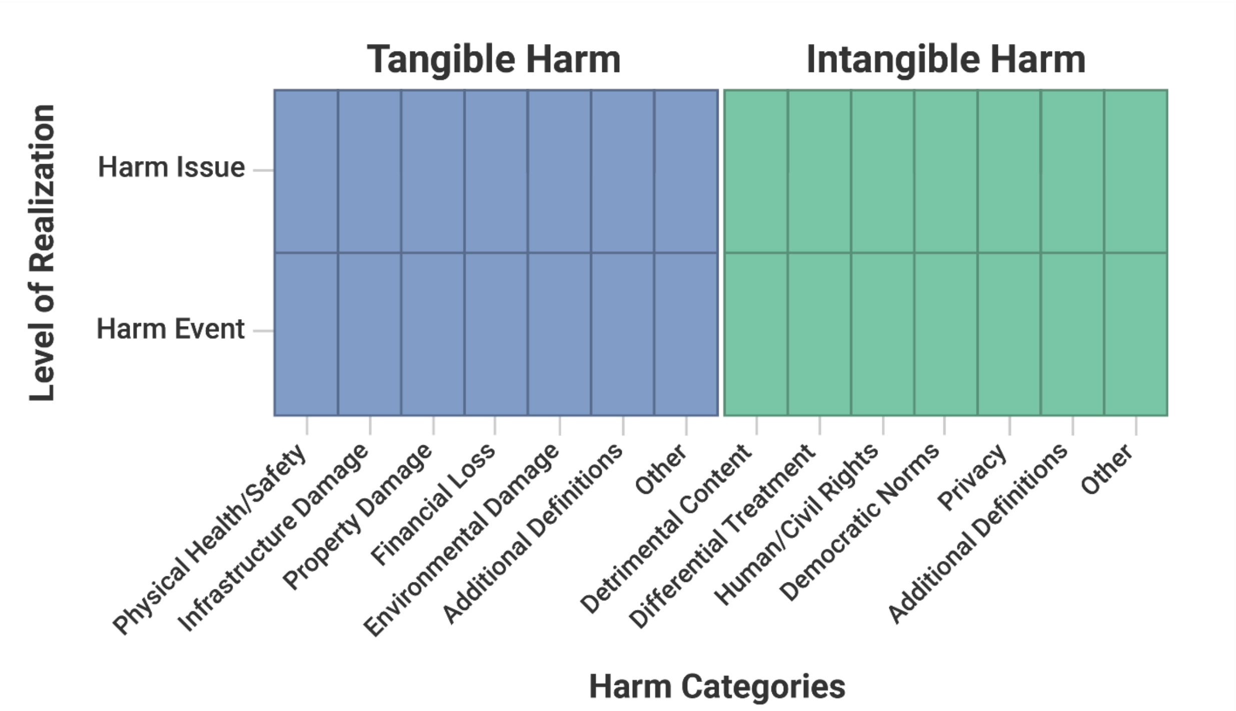
Imminency: Events v. Issues

Harm issues are instances where harm nearly happened or could reasonably happen.

Often reports about AI behavior are not about a harm that happened, i.e., a "harm event," but about harm that nearly happened or could reasonably happen, "harm issues." These potential harms can often be predicted before real harm occurs. In order to reduce AI harm events in the future, differentiating harm issues from harm events is essential because it allows AI system developers, deployers, and users to assess vulnerabilities and develop mitigations to prevent occurrence. In contrast, tracking and analyzing realized harm events provides information about the effectiveness of existing harm mitigation and may reveal harms that were not previously anticipated.

Thus, the framework differentiates between a harm event that has occurred and harm issues where harm has not yet occurred (Figure 3).

Figure 3. Differentiating Between Harm Events and Harm Issues



Source: CSET AI Harm Framework.

Box 3. Example of Harm Issue

Harm issues can be problems and failures that do not rise to the level of a harm event because they were identified in development, training, or testing before the risk materializes. For example, developers of autonomous vehicles are working on the challenge of "snow-blindness," the phenomenon that the quality of both the data collected by sensors on the car and the AI model outputs tend to deteriorate during bad weather, which makes the detection of road lanes and obstacles difficult.[23] While fully self-driving cars are still largely confined to testing or controlled environments, semi-autonomous vehicles are already widely used in a variety of weather conditions. Because this vulnerability could lead to harm in the future, it is a harm issue.

Organizations may want to distinguish levels of harm potential. For this reason, the framework can be customized with additional categories associated with the imminency of harm. For example, our customization of the framework (discussed later in the document) further defines "near-misses," which are cases where harm almost occurred or there was an imminent potential for harm.

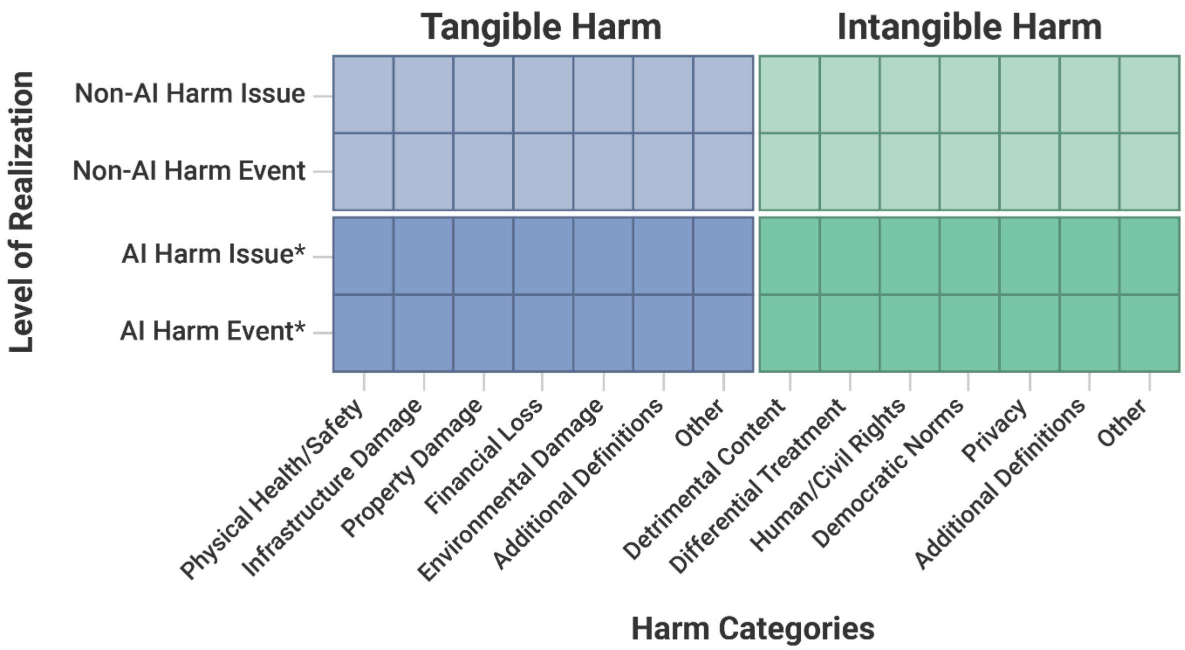
Incorporating Specific Elements of AI Harm

The framework incorporates four elements which, once appropriately defined, enable the precise identification of an AI harm. These key components serve to distinguish harm from non-harm and AI harm from non-AI harm. To be an AI harm, there must be:

1) an **entity** that experienced
2) a **harm event** or **harm issue** that
3) can be **directly linked** to a consequence of the behavior of
4) an **AI** system.

Under the CSET AI Harm Framework, all four elements need to be present in order for there to be AI harm. The following sections explain the rationale of each component and how they serve to detect AI harms (Figure 4).

Figure 4. Adding Elements of "AI Harm"



*An AI harm occurred when an **entity** experienced a **harm event** or **harm issue** that can be **directly linked** to a consequence of the behavior of an **AI system**.

Source: CSET AI Harm Framework.

Entities

A harmed entity can be a person, place, or thing.

Common entities are a person, a group of people, companies, locations, products, infrastructure, a government agency, or the natural environment. Harm is unlikely to occur without an entity that experiences it, which is why we consider an entity experiencing the harm event or issue to be a key component of AI harm.

Organization may develop differing definitions of "entity" when creating customizations of the framework. For example, one organization's customization of the framework could require all entities be “named entities,”[*] while another organization could adopt a looser definition.[†]

AI systems

The presence of an AI system is clearly paramount to an AI harm, and is therefore core to the CSET AI Harm Framework. However, defining what constitutes AI is not as straightforward. There is currently no universally adopted or agreed-upon definition of AI. Existing definitions by governments and governance bodies are often intentionally vague to avoid inadvertent exclusions of some technologies from proposed regulation.[24] Academic definitions tend to focus on technical system functionality, information that is often unavailable from public incident sources.[25] This is precisely why the framework itself does not provide a definition of AI. Instead, it lets organizations define AI as part of framework customization.

The CSET AI Harm Framework does not define AI, but organizations need to define AI when creating a customized framework.

Directly Linking Harm to AI

While the presence of an AI system is required, it is not sufficient. The harm event or issue has to be directly linked to a consequence of an AI’s behavior. This does not mean that the AI must be the *sole* cause for the harm, but it needs to be instrumental enough that the harm would not have occurred had the system not been involved or behaved in a different way.

The exact manner of AI involvement is intentionally left undefined in the framework. While much focus today lies on preventing harm incidents from internal AI failures like poor performance, bias or misspecification, there are several links from AI systems

[*] Data scientists are often interested in “named entities,” which are phrases or terms that clearly identify a specific entity. For example, a “university” is an entity and “Georgetown University” is a named entity. Most proper nouns are named entities, but not all named entities are proper nouns. For example, the latitude/longitude location for Washington DC is (38.9072° N, 77.0369° W), which is a clearly identified specific location (thus a named entity) but is not a proper noun.

to harm that do not require system failures. Users can intentionally employ AI systems to cause harm, or make mistakes when operating them. And even when an AI system functions as expected and is operated as intended, its use can have harmful consequences for those affected by its behavior.

Box 4. Examples of AI-Harm and Non-AI Harm

Consider the crash of a driverless metro train into a wall in Delhi in 2017. While the train was AI-powered, the accident happened because staff failed to deploy the train's brakes after a maintenance check. Therefore, the train's AI functionality cannot be linked to the crash, and the incident does not constitute an AI harm (*non-AI harm*).[26] In contrast, the fatal collision between a pedestrian and a self-driving Uber in Arizona in 2018 was an AI harm, because the car's AI model struggled to classify jaywalkers as pedestrians, which led to the delayed recognition of the pedestrian by the AI (*AI harm*).[27]

Finally, deepfake pornography generated using AI-tools allows for the non-consensual depiction of individuals in compromising contexts.[28] Deepfake pornography is an example of an AI harm that is intangible in its outcome (reputation and psychological harm), and represents a misuse of AI technology by users that results in harm, rather than a failure of the AI itself.

Future Framework Features: Severity and Spread

In future work we plan to extend the CSET AI Harm Framework to capture the severity and spread of harm. While some incidents occur locally and impact only those in their immediate vicinity, others affect the whole user base of an online service such as a social network or a cloud data storage. Likewise, some incidents have consequential impacts, while others only have minor effects. Adding both of these dimensions will enable a better understanding of the scale and severity of AI harms. Research into the design of metrics and indicators for severity and spread of harm, in particular for intangible harm, is ongoing.

Creating CSET AI Harm Framework Customizations

Researchers and organizations interested in tracking and analyzing AI harm can use the structure offered by the CSET AI Harm Framework as a starting point. After identifying a data source, users should create a framework customization to ensure consistent extraction of all the information needed to identify and classify AI harms according to the user's analytic interests.

Developing a customized framework involves adapting the basic framework structure shown in Figure 4, developing suitable definitions for all components and describing additional incident information to be recorded. All choices and adaptations should be clearly documented along with data sources and any associated data limitations.

The main steps for customization are:

1) Review available dataset(s)[*] and develop analytic goals
    - Understand dataset's content, strengths, and limitations
    - Make analytic goals that incorporate your organizational needs and understanding of the dataset
2) Define key elements and categories
    - AI harm elements: entity, harm event, harm issue, directly linked, and AI
    - Subcategories for tangible and intangible harm; e.g., financial loss, injury, privacy, etc.
    - Additional incident information that an organization wants extracted
3) Modify basic framework structure (Figure 4)
    - Alter the base, modular structure of the CSET AI Harm Framework
    - Add, combine, or remove harm categories or levels to reflect definitions and analytic goals
4) Document
    - Analytic goals, definitions, and modified framework structure
    - Provide clarifying use-case and edge-case examples

[*] Creating your own dataset is an option. However, it will probably result in more time developing analytic goals, determining your modified framework structure, and defining additional incident information that you want to extract. You may have to discover the strengths and limitations gradually as data is collected. Thus, it may be advisable to do many quick iterations through the steps.

5) Apply and assess customization
    - Apply the customization to representative incident data
    - Assess results
    - Refine the definitions, analytics goals, and customization
    - Identify additional guidance to enable consistent application
6) Iterate

Key Considerations for Effective Framework Customizations

Good Definitions

When possible, use existing definitions that are broadly acceptable to a wide range of legal systems, organizations, or people. When defining harm categories or levels of harm realization, specify the boundaries between each. Ideally, the definitions will be clear and provide enough detail for consistent application.

Defining "AI"

Among the most important definitions is what is considered to be an AI system. Analyzing incidents requires a decision-rule that distinguishes harm situations caused by AI from those caused by traditional digital technology; and data comparability requires this decision-rule be made explicit. How narrowly or broadly AI should be defined depends on the questions framework users intend to answer. Organizations that want to study the harms of algorithmic systems may want to include rule-based, statistical and machine-learning systems. Others may be interested in harms caused by a specific subset of AI, such as generative models. Defining AI based on the user's analytic goals is one of the most important elements of any customization of this framework.

Adding Additional Details

A customization can describe the additional incident data that an organization wants extracted. These details often are characteristics of dimensions like:

- Harm type: the specific civil right impinged, the group or nature of the differential treatment

- Temporal/location information: when/where the harm occurred, environmental conditions (e.g., rainy night, high temperatures, etc.)
- Tangible harm quantities: the number of people injured, the size of the financial loss
- Deployment: how many people or organizations use or are affected by the system, the sector in which the system is used
- System: the level of system autonomy, the methods or tasks of the AI

Example: CSET AI Harm Framework Customization for AIID

We customized the CSET AI Harm Framework to support our research on AI harms. The resulting customization incorporates our analytic goals[*] and is specific to the AI Incident Database (AIID). Figure 5 summarizes how we customized the presented framework. The annotation guidance for this taxonomy provides specific definitions, illustrative examples, guidance on complicated situations, and details on supplementary characteristics of interest[†] that the CSET AI Harm Taxonomy for AIID records.[29]

AIID, The Incident Data Source

Our data-source is the AI Incident Database (AIID) maintained by the Responsible AI Collaborative (RAIC). Incidents in the AIID are based on publicly available reports (news, academic papers, etc.) of adverse AI behavior and concerns. Anyone can nominate a report for inclusion in the AIID. Additionally, RAIC actively searches for new AI incidents to incorporate into the database. With 2,500+ incident reports describing more than 540 distinct incidents, the AIID is a valuable public database of AI harms.[‡]

Despite its exceptionally broad coverage, the data source is limited based on the collection mode employed by RAIC. The AIID primarily logs incidents recorded in English-language news; thus, it likely undercounts total incidents worldwide and likely privileges incidents occurring in English speaking countries.[§] It also likely over-

[*] Our primary analytic goal was to extract details of AI harm events and issues in order to identify, track and mitigate risks for AI systems.

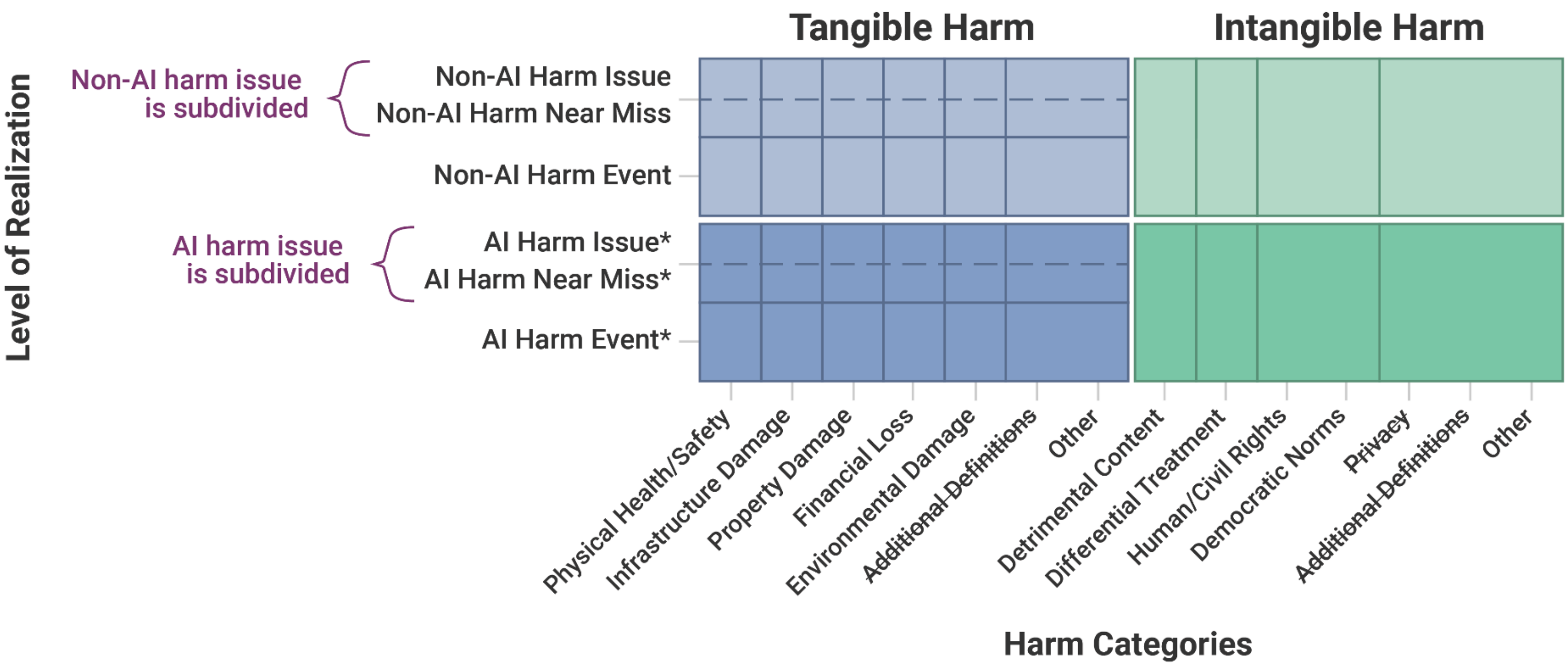
[†] For example, the CSET AI Harm Taxonomy for AIID records if the incident involved a minor child, if multiple AI systems were interacting, and if the AI system was designed to be harmful. These characteristics do not impact the assessment of harm as shown in Figure 5. They are simply items that we are interested in tracking and analyzing.

[‡] As of June 2023.

[§] English-language articles make up >99 percent of all reports in the database. Other represented languages are German, Spanish, Thai, Italian, Korean, French, Chinese, Hindi, Vietnamese and Portuguese (as of 03/20/2023).

indexes on harms that are newsworthy. Since technology developers rarely publish information about the AI development process for their systems, reports often lack technical details about the AI system involved. As a result, the AIID likely underrepresents incidents occurring before deployment.

Figure 5. An Example Customization of the CSET AI Harm Framework: The CSET AI Harm Taxonomy for AIID



*An AI harm occurred when an **entity** experienced a **harm event** or **harm issue** that can be **directly linked** to a consequence of the behavior of an **AI system**. All bold and underlined items as defined in the CSET AI Harm Taxonomy for AIID.

Source: CSET AI Harm Framework.

Tangible harm types and imminency levels

We did not alter the set of tangible harm categories provided in the CSET AI Harm Framework (Figure 4) but defined the boundaries of each type. The CSET AI Harm Taxonomy for AIID considers tangible harm to be incidents involving physical injury (including death), financial loss, or physical damage. Any injury, as minor as scraped skin, is a tangible harm, as is a reduction in lifespan. Financial loss is defined as the inability to keep, have, or get something that is monetary in nature. Physical damage covers damage to objects, infrastructure and the natural environment. Infrastructure can be harmed through destruction, diminished capability, or reduced effectiveness, while pollution is the most likely type of environmental harm.

For tangible harms, we modified the levels associated with harm events and issues, adding an additional harm immediacy level to distinguish between imminent and non-imminent potential for tangible harm. AI is considered to present an imminent potential for harm in incidents describing “near miss” situations, where harm would have occurred had it not been for randomness, luck, or atypical intervention.

Box 5. Example of an AI-harm Near Miss

Consider the scenario of a self-driving vehicle whose AI fails to detect a red-light when it faces a certain angle of the sun. If the vehicle runs through the red-light, hitting and injuring a pedestrian, it is a harm event. If the vehicle runs through the red-light and narrowly avoids the pedestrian, because the pedestrian jumps out of the way at the last second, there is an imminent potential for harm (and therefore an “AI harm near-miss”).[30] If the vehicle runs through the red-light and there are no other cars or pedestrians on or near the intersection, there is neither a harm event nor an imminent potential for harm. However, the failure to correctly recognize the traffic signal implies that harm could plausibly occur in the future, making the incident a non-imminent potential for harm (and therefore an “AI harm issue”).

Intangible harm types and imminency Levels

The CSET AI Harm Taxonomy for AIID identifies three types of intangible harm of special interest for our analytic goals:

a) detrimental content (misinformation, hate-speech, etc.),

b) differential treatment based upon a protected characteristic, and;

c) harm to civil liberties, civil rights, human rights, or democratic norms.[*]

While incidents involving discrimination imply a violation of civil rights in certain contexts, the reverse is not true. Because algorithmic bias and discrimination is of special interest to us, the distinct category allows us to study this type of harm separate from other rights violations.

We recognize that there are many other intangible harms that are relevant in the context of AI, such as privacy violations or psychological harm. Future editions of the CSET AI Harm Taxonomy for AIID may specify and record additional categories of intangible harm. We prioritize the selected categories for several reasons:

1) Our analysis of incidents in the AIID shows that they occur relatively frequently.
2) They are currently of significant interest to the larger AI and policy community.
3) Despite the intangible nature of their harms, the occurrence or potential occurrence of bias, misinformation and rights violations is relatively easy to assess with sufficient certainty.

After reviewing many AI incidents in the AIID, we decided to not differentiate between intangible harm events and issues. This is because the line between intangible harm events and issues is inherently less clear and harder to differentiate and our data-source (AIID reports) often does not contain sufficient detail to make such distinctions for intangible harms.

[*] The definition of and additional information on these intangible harms can be found in the CSET AI Harm Taxonomy Annotation Guidance.

Taxonomy definitions of entity

Our definition of entity differs based on the type of harm they experience. For the CSET AI Harm Taxonomy for AIID, tangible AI harm must affect a specific and potentially identifiable entity. Such an entity may be an individual falling victim to an AI deepfake scam, welfare recipients that are subject to erroneous algorithmic decision-making or patients at a hospital who were misdiagnosed for cancer as a result of a faulty AI-enabled diagnostic tool.

In contrast, for our taxonomy, entities experiencing intangible AI harm need to be characterizable. Characterizable is a lower threshold than potentially identifiable, which better reflects the instances of intangible harms in the incident database. The types of special interest intangible harms that our taxonomy prioritizes, in particular harmful content and discrimination, frequently affect groups of people rather than specific entities. Group affiliation is often defined by a shared characteristic, which may be related to identity (e.g., age, race, gender, religion) or shared experience (physical or virtual presence at the time of the incident, medical diagnoses, criminal history, etc.).

Box 6. Examples of Characterizable Entities

Google's ad-placement algorithm AdSense was found to deliver ads for services checking individuals' criminal history at significantly higher rates on searches for Black-identifying names than white-identifying names.[31] In this case, group affiliation is determined by race as assessed via names suggestive of African American ethnicity.

Another example of a characterizable subgroup are people whose pictures have been used to train image diffusion models. Scientists discovered they were able to extract training images from those models, which presents a potential violation of data protection and privacy rights.[32]

Taxonomy definition of AI systems

For the purposes of the CSET AI Harm Taxonomy for AIID, we define AI as the capability of machines to learn and perform functions that typically require human intelligence, such as reasoning or generating coherent language. Our definition of AI encompasses technologies based on machine learning and other contemporary AI techniques and excludes statistical, rule-based, or theoretically derived algorithms or traditional automation technology.

Box 7. Example of Harm without AI System

A 2019 study conducted in the Mass General Brigham health system demonstrated that a popular algorithm for estimating kidney function included a race multiplier, which underestimated the risk to African-American patients.[33] The equation had been used for decades and likely affected the health care for tens of thousands of Black Americans. While the formula is directly linked to harm it is not an AI system, and therefore this was not an AI harm event.

Additional incident information

Beyond the information captured within the CSET AI Harm Framework grid structure, as depicted in Figure 5, the CSET AI Harm Taxonomy for AIID extracts details from the AI incidents that provide valuable context for our analyses. We collect information on the entities responsible for the AI's development and deployment, the sector of use and details about the harm, including the basis for differential treatment or how many people were injured. We collect information about the AI's functionality (data inputs, task, etc.) and the level of autonomy an AI system operates in at the time of the incident. It distinguishes three levels:

- A "human-in-the-loop" level where the AI provides a cue but requires human approval or action to continue with its course of action.
- An intermediate "human-on-the-loop" level, where the AI executes actions based on its assessment but humans provide oversight and are able to intervene in real-time.
- A "human-out-of-the-loop" level where no human is involved in the AI's behavior.

Differentiating autonomy levels could help improve our understanding of the potential and pitfalls of human-AI-interaction and human-in-the-loop requirements for risky AI systems.

Conclusion

Keeping track of AI incidents is a key part of AI monitoring and harm mitigation, both in aggregate and at the level of individual systems. An important obstacle to developing responsible AI is the difficulty of understanding how current systems fail in the real world. In addition, even when functioning as intended, an AI's behavior or use can cause unintended harm. Therefore, data on AI incidents represents a vital source of information about failure modes, mechanisms of harm and particularly risky applications and techniques. This is especially true when such data can be shared and integrated across researchers and organizations.

Studying and analyzing AI harms requires comparing and combining data on individual incidents. This paper introduces the CSET AI Harm Framework, a conceptual structure for the definition, identification and characterization of AI harm that may be used as a foundation for taxonomy development. The framework lays out the key elements of any AI harm framework customization and provides their basic relational structure and definitions.

The CSET AI Harm Taxonomy for AIID (a customization of the AI harm framework) illustrates the framework's adaptability. Through specific definitions, the targeted inclusion and exclusion of the framework's building blocks, and the extraction of additional information relevant to our analytic goals, the resulting taxonomy is tailored to capture the most information from the AI Incident Database (AIID).

The resulting analyses of AI harms can support AI governance and risk mitigation efforts on many levels. Findings may inform the research and development of assessment and testing tools for uncovered vulnerabilities of AI systems. They may also contribute to monitoring and audit strategies while systems are in operation. Organizations planning to deploy a specific system can look up records of AI harms that have occurred in connection with similar technology in the past and take targeted steps to mitigate this risk. At a higher level, the AI harm framework facilitates identifying patterns, such as particularly problematic use cases. This, in turn, can help policymakers set priorities for regulation and enforcement.

The AI harm framework is a shared resource for researchers and organizations interested in monitoring and studying harms from artificial intelligence. By providing the analytical groundwork ,we hope the framework contributes to growing a community of AI harm researchers and a body of interoperable research. We invite users to reach out with questions, ideas and share lessons learned during their customization process, so that we may further strengthen the framework and its value as a shared resource.

Authors

Heather Frase is a senior fellow at CSET and leads the AI Assessment line of research, where Mia Hoffmann is a research fellow.

Acknowledgements

We would especially like to thank incident annotators Ankushi Mitra, Yu-Jie (Grace) Liao, Manda Xie, Reid Benjamin and Reut Medalion for their engagement and critical thought throughout the development and evolution of the CSET AI Harm Framework and CSET AI Harm Taxonomy for AIID. We would further like to express our gratitude to Helen Toner and Zachary Arnold whose earlier work on the prior edition of the CSET taxonomy was fundamental to this framework, and whose thoughtful comments substantially improved our work. Mina Narayanan's efforts were instrumental in the initial efforts of development and annotation. This framework and taxonomy would not exist without the continuous support of the Responsible AI Collective, and Sean McGregor and Luna McNulty in particular, for which we are very grateful. For their comprehensive and valuable reviews, we would like to thank Steph Batalis, Violet Turri, Jill Crisman and Nikiforos Pittaras. Finally, we would like to thank Jason Ly for his assistance in designing the framework figures, and Margarita Konaev and Igor Mikolic-Torreira for their feedback and support.

Document Identifier: doi: 10.51593/###

Endnotes

[1] "Welcome to the Artificial Intelligence Incident Database," Artificial Intelligence Incident Database, https://incidentdatabase.ai/

[2] "AIAAIC Repository," AIAAIC, https://www.aiaaic.org/aiaaic-repository

[3] "Home," AI Vulnerability Database, https://avidml.org/

[4] "AI Litigation Database," Ethical Tech Initiative of DC at George Washington Law School, https://blogs.gwu.edu/law-eti/ai-litigation-database/ and "Home," Emerging Technologies Institute, https://www.emergingtechnologiesinstitute.org/

[5] “Expert Group on AI Incidents,” OECD Working Party and Network of Experts on AI, https://oecd.ai/en/network-of-experts/working-group/10836

[6] “Taxonomy: GMF,” Artificial Intelligence Incident Database, https://incidentdatabase.ai/taxonomy/gmf/

[7] "AI, algorithmic, and automation risks and harms taxonomy," AIAAIC, https://www.aiaaic.org/projects/ai-algorithmic-risks-harms-taxonomy

[8] Artificial Intelligence Index Report 2023, (Stanford Institute for Human-Centered Artificial Intelligence, 2023), https://aiindex.stanford.edu/report/

[9] See David Shepardson, “U.S. safety agencies to investigate fatal Tesla crash in Florida,” *Reuters*, March 1, 2019, https://www.reuters.com/article/us-tesla-crash/us-safety-agencies-to-investigate-fatal-tesla-crash-in-florida-idUSKCN1QJ031; and “Autonomous Vehicle Collision Reports,” State of California Department of Motor Vehicles, https://www.dmv.ca.gov/portal/vehicle-industry-services/autonomous-vehicles/autonomous-vehicle-collision-reports/

[10] Chloe Xiang, “AI Spits Out Exact Copies of Training Images, Real People, Logos, Researchers Find,” *Vice*, February 1, 2023, https://www.vice.com/en/article/m7gznn/ai-spits-out-exact-copies-of-training-images-real-people-logos-researchers-find

[11] Kashmir Hill, “Wrongfully accused by an algorithm,” *New York Times*, June 24, 2022, https://www.nytimes.com/2020/06/24/technology/facial-recognition-arrest.html

[12] Ziad Obermeyer, Brian Powers, Christine Vogel and Sendhil Mullainathan, "Dissecting racial bias in an algorithm used to manage the health of populations," *Science* 366, no. 6464 (2019):447-453, DOI:10.1126/science.aax2342

[13] Karlin Lillington, "Leaving Cert: Why the Government deserves an F for algorithms," *The Irish Times*, October 8, 2020, https://www.irishtimes.com/business/technology/leaving-cert-why-the-government-deserves-an-f-for-algorithms-1.4374801

[14] Jeffrey Dastin, "Amazon scraps secret AI recruiting tool that showed bias against women," *Reuters*, October 10, 2018, https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G

[15] Suzie Dunn, "Women, not politicians, are targeted most often by deepfake videos," *Center for International Governance Innovation*, March 3, 2021, https://www.cigionline.org/articles/women-not-politicians-are-targeted-most-often-deepfake-videos/

[16] "Home," O'Neil Risk Consulting and Algorithmic Auditing, https://orcaarisk.com/

[17] For example, interpretations for what qualifies as harm from libel vary so much that people often "shop" for the best country in which to bring a libel lawsuit. Trevor C. Hartley, "'Libel tourism' and conflict of laws," *International and Comparative Law Quarterly* 59, no. 1 (January 2010): 25-38, https://www.proquest.com/docview/236612390

[18] Anonymous. (2018-09-01) Incident Number 101. in McGregor, S. (ed.) *Artificial Intelligence Incident Database.* Responsible AI Collaborative. Retrieved on June 26, 2023 from www.incidentdatabase.ai/cite/101

[19] See Anonymous. (2010-05-08) Incident Number 28. in McGregor, S. (ed.) *Artificial Intelligence Incident Database*. Responsible AI Collaborative. Retrieved on May 10, 2023 from https://incidentdatabase.ai/cite/28; and Rob Davies, "What caused the pound's flash crash?," *The Guardian*, October 7, 2016, https://www.theguardian.com/business/2016/oct/07/what-caused-pound-flash-crash-brexit-fallen-sterling . See also Khoa Lam, (2019-03-01) Incident Number 200. in McGregor, S. (ed.) *Artificial Intelligence Incident Database.* Responsible AI Collaborative. Retrieved on May 11, 2023 from https://incidentdatabase.ai/cite/200/

[20] See "Maneuvering Characteristics Augmentation System," *Wikipedia*, accessed June 2023, https://en.wikipedia.org/wiki/Maneuvering_Characteristics_Augmentation_System; and Catherine Olsson, (2018-10-27) Incident Number 3. in McGregor, S. (ed.) *Artificial Intelligence Incident Database.* Responsible AI Collaborative. Retrieved on May 11, 2023 from https://incidentdatabase.ai/cite/3/

[21] Sarah-Jane Dobson, Paula Margolis, Katherine Ciclitira, and Yaseen Altaf, "Intangible Risks of Modern Products," T*homson Reuters Practical Law*, January 11, 2023, https://uk.practicallaw.thomsonreuters.com/Document/I603b6eea54a411ed8636e1a02dc72ff6/View/FullText.html?transitionType=Default&contextData=(sc.Default)&firstPage=true

[22] Ingrid Dickinson, (2020-10-24) Incident Number 80. in McGregor, S. (ed.) *Artificial Intelligence Incident Database*. Responsible AI Collaborative. Retrieved on May 11, 2023 from https://incidentdatabase.ai/cite/80/

[23] See Allison Mills, "Driving in the Snow is a Team Effort for AI Sensors," *Michigan Tech News*, May 27, 2021, https://www.mtu.edu/news/2021/05/driving-in-the-snow-is-a-team-effort-for-ai-sensors.html; and Anonymous, (2016-02-10) Incident Number 70. in McGregor, S. (ed.) *Artificial Intelligence Incident Database*. Responsible AI Collaborative. Retrieved on May 9, 2023 from www.incidentdatabase.ai/cite/70

[24] Matt O'Shaughnessy, "One of the biggest problems in regulating AI is agreeing on a definition," *Carnegie Endowment for International Peace*, October 6, 2022, https://carnegieendowment.org/2022/10/06/one-of-biggest-problems-in-regulating-ai-is-agreeing-on-definition-pub-88100

[25] Peter M. Krafft, Meg Young, Michael A. Katell, Karen Huang and Ghislain Bugingo, "Defining AI in policy versus practice," In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* (February 2020): 72-78, https://doi.org/10.1145/3375627.3375835

[26] Anonymous, (2017-12-03) Incident Number 31. in McGregor, S. (ed.) *Artificial Intelligence Incident Database*. Responsible AI Collaborative. Retrieved on May 9, 2023 from www.incidentdatabase.ai/cite/31

[27] Catherine Olsson, (2018-03-18) Incident Number 4. in McGregor, S. (ed.) *Artificial Intelligence Incident Database*. Responsible AI Collaborative. Retrieved on May 9, 2023 from www.incidentdatabase.ai/cite/4

[28] See e.g., Daniel Atherton, (2023-01-30) Incident Number 480. in Lam, K. (ed.) *Artificial Intelligence Incident Database*. Responsible AI Collaborative. Retrieved on June 9, 2023 from www.incidentdatabase.ai/cite/480

[29] See Mia Hoffmann, Mina Narayanan, Ankushi Mitra, Yu-Jie Liao, and Heather Frase, "CSET AI Harm Taxonomy for AIID and Annotation Guide," https://github.com/georgetown-cset/CSET-AIID-harm-taxnomy

[30] Such a near-miss occurred in 2016, when an Uber in self-driving mode ran a red-light and nearly collided with another car on the intersection. See Anonymous, (2014-08-15) Incident Number 8. in McGregor, S. (ed.) *Artificial Intelligence Incident Database*. Responsible AI Collaborative. Retrieved on May 10, 2023 from www.incidentdatabase.ai/cite/8

[31] See Latanya Sweeney, "Discrimination in online ad delivery: Google Ads, Black Names and White Names, Racial Discrimination, and Click Advertising," *Queue* 11, no. 3 (2013): 10-29, https://doi.org/10.1145/2460276.2460278; and Roman Yampolskiy, (2013-01-23) Incident Number 19. in McGregor, S. (ed.) *Artificial Intelligence Incident Database*. Responsible AI Collaborative. Retrieved on May 9, 2023 from www.incidentdatabase.ai/cite/19

[32] Nicholas Carlini, Jamie Hayes, Milad Nasr, Matthew Jagielski, Vikash Sehwag, Florian Tramèr, Borja Balle, Daphne Ippolito and Eric Wallace, "Extracting Training Data from Diffusion Models," *arXiv* preprint (2023), arXiv:2301.13188v1

[33] Ingrid Dickinson, (2020-07-17) Incident Number 87. in McGregor, S. (ed.) *Artificial Intelligence Incident Database*. Responsible AI Collaborative. Retrieved on July 20, 2023 from https://incidentdatabase.ai/cite/79/#r1736