CSET AI Harm Taxonomy for AIID and Annotation Guide

Authors

Mia Hoffmann
Mina Narayanan
Ankushi Mitra
Yu-Jie Liao
Heather Frase, PhD

Overview

CSET is updating version 1 of its AI Harm Taxonomy for characterizing harm associated with AI incidents. The impetus behind the update was to (1) decrease inter-annotator variability for AI incident harm annotation, (2) better distinguish between tangible harm events, near-misses, issues, and (3) identify three specific, special-interest intangible harm types.

A goal of the CSET harm taxonomy is to create a structure for extracting AI harm information that will allow people to draw conclusions from the dataset without having to compile and interpret the individual incident reports themselves. It can provide relevant information to AI stakeholders to help them manage risk and prevent harm from AI systems, track trends in AI incidents, and identify emerging types and vectors of AI harm.

CSET's taxonomy defines and differentiates between tangible and intangible harm. CSET considers tangible harm to be incidents definitively involving observable injury, loss, or damage. Intangible harm cannot be directly observed. What constitutes intangible harms depends upon societal norms or cultural context, while tangible harms are often easily quantifiable. CSET also distinguishes three categories of tangible harm (events, near-misses, and issues). At this time, CSET is not distinguishing or defining similar levels (event, near-miss, and issue) for intangible harms. This approach to separating tangible and intangible harms aims to reduce inter-annotator variability for tangible harms.

CSET's taxonomy focuses on harm that occurred and harm that has a reasonable probability of occurring now. Thus, if a possible harm require the creation of an AI system that does not yet exist or is only used in controled 'lab-like' environments, CSET does not view there to be a resonable probablity of harm occuring now and CSET's definition of harm is not met. CSET's taxnomy include a series of questions to help annotator assess if there is a reasonable probablity of harm occuring now. These questions are about the domain or use conditions of an AI system.

Additionally, CSET attempts to capture details about the AI system, sector, environment, entities, locations, dates, and type of harm that were involved in the AI incident. Frequently, the reports do not provide annotators with enough information to completely record all of this information. CSET's taxonomy prioritizes recording more details about the type and quantities of tangible harm. It records fewer details about the type of intangible harm.

Because intangible harms are generally more subjective, they are inherently more difficult to annotate, resulting in increased inter-annotator variation. Definitions of intangible harms can vary depending upon cultural and societal norms. Thus, there can also be variation in the types of intangible harm (e.g. psychological harm, harm to social norms, reputation harm, etc.) that different groups prioritize. This version of the CSET taxonomy focuses on identifying three types of special interest intangible harms; 1) harm to civil liberties, civil rights, human rights, or democratic norms, 2) detrimental content (misinformation, hate-speech, etc), and 3) differential

treatment based upon a protected characteristic. Future taxonomy versions may expand upon these.

Initially, this taxonomy will be applied to incidents in the Artificial Intelligence Incident Database (AIID).[1] AIID collects and indexes reports of AI Incidents. Thus, this taxonomy focuses on extracting data likely to be present in the reports. The taxonomy could be adapted for other sources of AI incident information.

Definitions of AI Harm

CSET divides AI harm into two groups, tangible and intangible harm. Further, CSET distinguishes three categories of AI tangible harm imminency: event, near-miss, and issue. A single instance can involve multiple types of tangible harm (death, financial loss, damage to private property, etc.) and/or intangible harm (harmful content, damage to social institutions, etc.).

Tangible harm is harm that is observable, verifiable, and definitive. It typically encompasses physical and economic harm through injury (including death), loss, or damage, and may be expressed in monetary terms. Examples of tangible harm include a damaged car, an injured person, or loss of income.

CSET's taxonomy places instances of tangible harm into one of three categories of imminency; event, near-miss, and issue. These three categories of AI harm differ only in the immediacy of harm. For an incident to constitute an AI tangible harm event, the harm needs to have definitely occurred. For example, situations where an AI system could have caused harm but did not due to a safety mechanism are not AI harm since the safety mechanisms worked as expected to prevent the harm. Near-misses record incidents with an imminent risk of harm. Harm issues record incidents exhibiting a non-imminent risk of harm.

CSET applies the following definitions;

> An **AI Tangible Harm Event** occurs when 1) a potentially identifiable, specific entity 2) experiences an event that causes tangible harm (injury, loss, or damage) which 3) can be directly linked to a consequence of 4) an AI's behavior.
>
> An **AI Tangible Harm Near-Miss** occurs when 1) a potentially identifiable, specific entity 2) experiences an event where tangible harm (injury, loss, or damage) does not occur but where there is an imminent risk of tangible harm which 3) can be directly linked to a consequence of 4) an AI's behavior.

[1] The Artificial Intelligence Incident Database can be accessed at https://incidentdatabase.ai/.

> An **AI Tangible Harm Issue** occurs when 1) a potentially identifiable, specific entity experiences an event where 2) tangible harm (injury, loss, or damage) does not occur, but where there is a non-imminent risk of tangible harm which 3) can be directly linked to a consequence of 4) an AI's behavior.

An incident can be moved to a different category after more information is gained. For example, based upon the information in the reports about the incident, it is often not possible to determine if an AI was involved versus some other type of automated process. This would not meet CSET's definitions for an AI harm event, near-miss, or issue.[2] However, if additional information later makes it clear that an AI was definitely involved, then the incident would move to a different AI tangible harm category.

Intangible harm is harm that cannot, even with additional information, be directly observed or does not have any material or physical effect. This can include (but is not limited to) mental/psychological harm, harm to opportunity, harm to intangible property (for example, IP theft, damage to a company's reputation), and loss of trust or belief.

CSET applies the following definition:

> A **Special Interest Intangible AI Harm** occurs when 1) a characterizable class or subgroup of entities 2) experiences or has a risk of experiencing a designated intangible harm that 3) can be directly linked to the consequences of 4) an AI's behavior. In reference to #2 above, CSET has designated three categories of intangible harm: a) harm to civil liberties, civil rights, human rights, or democratic norms, b) detrimental content (misinformation, hate-speech, etc), and c) differential treatment based upon a protected characteristic.

CSET recognizes that these three categories do not reflect all the possible types of intangible harm. What qualifies as intangible harm is often subjective and dependent upon cultural perspectives. As such, other organizations could easily adopt the part of CSET's taxonomy associated with tangible harm, but adapt what qualifies as intangible harm to their specific needs. Later versions of CSETs taxonomy may specify and record more intangible harm types.

Note, it is possible for tangible harm to result from intangible harm. For example, 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. Moreover, one incident may cause both tangible and intangible harm. For example, a facial recognition AI system could

[2] Compared to CSET, the AIID has more permissive in what it considers an AI incident of interest and inclusion in its dataset. This is beneficial to researchers, because it allows researchers to have custom and different definitions and enables focused, nuanced research.

work poorly on women, leading to false arrest and the loss of wages. This would cause both tangible and special interest intangible AI harms, because financial harm was inflicted and a subgroup of people were disproportionately affected by the harm.

Table 1: Summary of AI tangible harm categories and special interest intangible AI harm.

Harm type	AI Tangible Harm			Special Interest Intangible Harm
Harm category	**Event**	**Near-Miss**	**Issue**	**Categories currently not defined**
AI	An AI or a system with an embedded AI can be definitively identified.	An AI or a system with an embedded AI can be definitively identified.	An AI or a system with an embedded AI can be definitively identified.	An AI or a system with an embedded AI can be definitively identified.
Harm	Tangible harm ***definitively occurred***.	Tangible harm did not occur, but there was an ***imminent risk of tangible harm***.	Tangible harm did not occur, but there was a ***non-imminent risk of tangible harm***.	Intangible harm occurred, or there was a risk of imminent or non-imminent intangible harm.
Chain of harm	AI can be ***directly linked*** to harm.	AI can be ***directly linked*** to imminent risk of harm.	AI can be ***directly linked*** to non-imminent risk of harm.	AI can be ***directly linked*** to the special interest intangible harm.
Entity	A ***potentially identifiable specific entity*** exists which experienced the harm. The entity's name may not be known, but it exists.	A ***potentially identifiable specific entity*** exists which experienced the near miss. The entity's name may not be known, but it exists.	A ***potentially identifiable specific entity*** exists which experienced the non-imminent risk of harm. The entity's name may not be known, but it exists.	***A class or subgroup of people can be characterized*** who experienced or has risk of experiencing the special interest intangible harm.

AI Designed to Produce Harm

One value of collecting and annotating AI incidents is that information on potential AI risks, vulnerabilities, and unintended consequences can be gathered. To enable better analysis and

more effective forecasting of undesirable impacts, CSET's AI harm taxonomy distinguishes between harm created unintentionally versus by design.

CSET's AI harm taxonomy records if an incident was linked to an AI designed to produce the observed harmful behavior. This distinction allows separate analyses of accidental and intentional AI harm for a more nuanced picture of vulnerabilities and risks of AI systems designed for different purposes. Examples of incidents with AI expected harm include AI-enabled weapon systems, AI built by one company to hurt another, an AI designed to support criminal activity, or an AI produced by a researcher to understand harmful AI behavior. Included are situations where an individual or an organization develops or deploys an AI system to perpetrate harm or undermine the security or well-being of another entity. While incidents with this type of system can produce harm, they are tagged as incidents where the harm was expected.

Still, it is possible for an AI designed to be harmful to not behave as expected and produce unintentional harm. Thus, if an AI in a weapon system was expected to harm target A, but instead unexpectedly harmed target B, it would be a tangible AI harm event. For this reason, AI incidents with AI designed to be harmful are also tagged with whether or not the observed harm was linked to intended behavior.

Importance of domain

When assessing an AI Incident, determining if an AI system can be directly linked to the harm, risk of harm, or special interest intangible harm can be difficult. CSET developed a series of questions that help determine if a direct link exists. These questions center around the domain or conditions in which the potential AI harm occurred. Answering these questions helps the annotator focus on characteristics and details that inform the determinations of harm and/or issue. While the questions greatly improve annotation consistency[3], they are not sufficient to cover all possible domains or conditions. Annotators must consider the totality of information and not assume that answering these questions is sufficient for understanding the potential for harm in the given domain.

There are currently eight questions about the domain. They focus on understanding if the system was undergoing testing/demonstration, if the AI system could interact with the physical environment, and who was using the system when the potential AI harm occurred. AI systems that interact with physical objects are more likely to create injury or damage. Incidents that are tests or demonstrations are less likely to lead to harm – but they still can. Some of the below questions are designed to help the annotator figure out if harm could have occurred even during testing or demonstration of the AI system. **Ultimately, domain questions help annotators**

[3] Catherine Aiken, "Classifying AI Systems," (Center for Security and Emerging Technology, November 2021). https://doi.org/10.51593/20200025

make determinations about the potential harm's immediacy and inform incident categorizing, but they are not absolute determinants of AI harm category.

Did the incident occur in a domain with physical objects (roads, factories, medical facilities, etc.)?
Incidents that involve physical objects are more likely to have damage or injury. However, AI systems that do not operate in a physical domain can still lead to harm. For example a large language model could be misused by criminals to get identifying information from a person, which is then used for identity theft.

Was the AI used for entertainment purposes?
AI systems used for entertainment are less likely to involve physical objects and hence unlikely to be associated with damage, injury, or loss. Additionally, there is a lower expectation for truthful information from entertainment, making detrimental content less likely (but still possible). For example, a deepfake of a pet put into a YouTube video is considered harmless entertainment, while a deepfake of a presidential candidate on social media is likely a misinformation special interest intangible harm and in some context could lead to tangible harm.

Was the incident about a report, test, or study of data instead of the AI itself?
The quality of AI training and deployment data can potentially create harm or risks in AI systems. However, an issue in the data does not necessarily mean the AI will cause harm or increase the risk for harm. It is possible that developers or users apply techniques and processes to mitigate issues with data. In fact, it is best practice to assume that there are issues with the training or input data and take appropriate mitigation measures. Unless the data issues clearly contributed to an instance of an AI harming or otherwise endangering an entity or group of entities, these reports are unlikely to be AI harm events or near-misses.

Was the reported system (even if AI involvement is unknown) deployed or sold to users?
AI systems that have not been sold or provided to the user are more likely to be in the development stage and less likely to result in AI harm. However, this is not a hard and fast rule. For example, harm could occur if a company develops a prototype AI system for hiring, uses it internally before selling it, and later discovers that the AI prototype had gender biased results. In this case, the system is in development, under testing and not sold to the targeted user – but it was still used in a manner in which harm could occur.

Was this a test or demonstration of an AI system done by developers, producers or researchers (versus users) in controlled conditions?
AI tests or demonstrations by developers, producers, or researchers in controlled environments are less likely to expose people, organizations, property, institutions, or the natural environment to harm. Controlled environments may include situations such as an isolated compute system, a regulatory sandbox, or an autonomous vehicle testing range.

Was this a test or demonstration of an AI system done by developers, producers or researchers (versus users) in operational conditions?
While almost every AI system undergoes testing or demonstration in a controlled environment, some also undergo testing or demonstration in an operational environment. Testing in operational environments still occurs before the system is deployed or sold to end-users. However, relative to controlled environments, operational environments try to closely represent real-world conditions and end-users that affect use of the AI system. Therefore, testing in an operational environment typically poses a heightened risk of harm to people, organizations, property, institutions, or the environment.

Was this a test or demonstration done by users in controlled conditions?
Sometimes, prior to deployment, the users will perform a test or demonstration of the AI system. The involvement of a user (versus a developer, producer, or researcher) increases the likelihood that harm can occur even if the AI system is being tested in controlled environments because a user may not be as familiar with the functionality or operation of the AI system.

Additionally, these are formal tests or demonstrations done by the user. They have structured test plans and demonstration goals. These are different from informal or ad hoc research done by users after a system is deployed. Acceptance testing, which is a quality assurance process where the system is tested to see if it meets the end-user's needs, is an example of a formal test done by users.

Was this a test or demonstration done by users in operational conditions?
The involvement of a user (versus a developer, producer, or researcher) increases the likelihood that harm can occur even if the AI system is being tested. Relative to controlled environments, operational environments try to closely represent real-world conditions and end-users that affect use of the AI system. Therefore, testing in an operational environment typically poses a heightened risk of harm to people, organizations, property, institutions, or the environment.

Guidance for Annotating

What to do

- Pick an incident from the database.
- Review the reports associated with the incident. If the reports are voluminous (>100 pages total), review until you're confident you have a firm understanding of the incident and will be able to annotate it thoroughly, prioritizing more recent reports and those from reputable sources.
- Based on those reports, fill out the form for that incident, referring to the field (column) descriptions in Appendix B as you go.

General pointers

- **Before annotating incidents, review the overview section, the AI harm definitions, the definition of terms in Appendix A**, which explains some important concepts relevant to this annotation task (such as "AI system" and "entity").
- **Don't speculate.** Everything you put in each row of the spreadsheet should have a clear basis in one or more incident reports. That means that the information you're adding to the sheet should either be *explicitly stated* or *directly and clearly implied* in at least one report on the relevant incident. Determining whether this is the case will sometimes require your best judgment. In some instances, you'll have the option to mark "Unclear" (or similar) in a field - use this if and only if, after careful inquiry, you think that's the best answer. In other words, please don't use it as a crutch in cases when there's decent information available to you and a judgment call is required. When reasonable, we want you to make that call.
- **It's OK to leave blanks.** We don't expect to be able to fill out every field for every incident. For some fields, especially those related to the details of AI systems, you'll *rarely* have the information you need to fill them out. This is OK. We would rather have more authoritative data with large gaps in it than more comprehensive data based on a lot of guesswork and unnecessary subjective judgments.
- **Incident variants[4] are grouped together:** AIID identifies related reports of incident variants and groups them together, creating a consolidated incident that covers all variants. An incident variant 'that shares the same causative factors, produces similar harms, and involves the same intelligent systems as a known AI incident.' This approach reduces 'the multiplicity of events that are likely to repeat, reducing the human labor

[4]Sean McGregor, Kevin Paeth, Khoa Lam, Indexing AI Risks with Incidents, Issues, and Variants, To be published in Human-Centered AI Workshop at NeurIPS 2022, https://arxiv.org/pdf/2211.10384.pdf

required for subsequent event processing.' Thus, a single incident in the AIID may have multiple reports, harmed parties, and harm occurrences at different times or locations.

- **Multiple and potentially conflicting reports.** The AI Incident Database (AIID) is currently the main source for potential AI incidents that are being annotated. In the AIID, each incident can have multiple reports and sources. Sometimes the reports can conflict or provide different information. Additionally sometimes an individual report, in the group of reports, may be discussing multiple AI incidents.
    - **Stick to the incident reports.** All of our annotations need to be based on information in the database - not on your background knowledge, stuff you Googled on the side, etc. You can, of course, find additional incident reports that fill gaps in the information currently in the database, and add them to the database as well. But the basic principle holds: all of our annotations need to be based on information in the database.
    - **Resolve conflicts carefully.** From time to time, different sources may have conflicting information. It's part of your job as an annotator to resolve these conflicts using your best judgment. As you do, consider the credibility and recency of each source. When uncertainty still remains, err on the side of classifying the account as an issue or the lowest clearly known level of harm category. We should have a high degree of confidence that harm did indeed involve harm tied to an entity and an AI system.
    - **Some reports have multiple incidents.**[5] While most incidents contain multiple reports and variants on the same incident, there are cases where a report may discuss multiple distinct incidents. This tends to happen in high-level reports that are discussing the breadth, depth, or variety of AI harm. With such reports it can make it difficult to know which potential incident to annotate. The best approach is to look at the incident title and then identify the report's content that is associated with the title.
- **Provide Notes.** Frequently it is useful to understand why an annotator entered/chose a value for an annotation field. There are multiple fields where annotators can add notes. Annotators should add any notes that they feel are helpful, however, we strongly recommend adding notes when
    - The incident does not meet any of CSET's definitions for AI tangible harm or AI special interest intangible harm.
    - The annotator finds that no AI system was involved (e.g. statistical methods were used instead of AI) or if they are unclear if an AI system was involved.
    - The annotator does not find that the AI was directly linked to the harm.

[5] As AIID develops, this situation may be ameliorated and handled before these reports get to the public facing website.

- There was actual or a risk of tangible harm, but the CSET's definition for AI tangible harm was not met because of the domain in which the incident occurred.
- The annotator's input is based upon an item in the reports that may be missed by a casual reader.

Detailed Guidance

- **Conducting an AI harm assessment**: In order to determine if the incident was an AI harm you need to answer four questions that each address a different component of the definitions described above.
    1) Did harm occur?
    2) Does the incident involve an AI system?
    3) Can an AI be directly and clearly linked to harm?
    4) Is there an entity that experienced the harm?

  Questions 1, 2 and 4 are relatively straightforward: *Did harm occur?* establishes if the adverse outcome described in the incident meets CSET's definition of a tangible or special interest intangible harm. *Does the incident involve an AI system?* checks if the technology involved is a simple automation technology or a rule-based software system, or if it is based on machine-learning. And *Is there an entity that experienced the harm?* verifies that the adverse outcome described in the incident happens to a type of entity that meets CSET's definition.

  The spirit of the third question is to establish the relationship between the technology and the harm without assessing the nature of both components. In other words, it asks if the behavior of the technology described in the incident can be directly and clearly linked to the adverse or potentially adverse outcome described in the incident. Complicated incidents may contain multiple adverse outcomes and technologies that encompass AI-driven and other functionalities. This means that answering this question requires a nuanced approach.
    - If your answer to question 1 is yes, and to question 2 is no, question 3 refers to the link between the technology's behavior and the tangible/special interest intangible harm. For example, incident #24 describes a fatal workplace accident caused by a manufacturing robot in a VW plant. Death is a tangible harm, but the robot is not powered by machine-learning technology. Since the robot's behavior can be directly linked to the fatality, the answer to question 3 is Yes.
    - If your answer to question 1 is no, and to question 2 is yes, question 3 asks about the link between the AI's behavior and the adverse outcome described in the incident. For example, incident #80 describes an AI-powered camera deployed by a Scottish soccer club that is programmed to track the ball during soccer matches. During one game, the camera mistakes a referee's bald head for the

ball and fails to broadcast the match. This does not constitute a harm per CSET definitions, but the adverse outcome is directly and clearly linked to the AI's behavior.

- Finally, if you answer yes to both, question 3 refers to the link between the AI's behavior and the tangible/special interest intangible harm. For example, incident #31 describes a driverless metro train in Delhi, India, crashing into a wall. While the train was AI-powered, the crash happened because staff failed to deploy the train's brakes after a maintenance check. Therefore, the AI functionality cannot be linked to the tangible harm outcome and the answer to question 3 is No.

- **Annotating entities**: Annotators should record all entities involved in an incident. At the very least, ever incident includes the following two:
  1. The AI system or technology involved
  2. The allegedly harmed entity or entities.

  But most incidents will involve a much larger number of entities. Some are more common, such as the developer and/or deployer of the AI system; others are more rare, like researchers or government watchdogs discovering a harm from an AI.
  - **When there are many entities:** Occasionally, there is a large number of named entities that have the same role and relationship to the AI system. In order to limit the workload of annotation, annotators can group them as one entity. For example, incident #13 lists five online media outlets that have used or adopted a particular content moderation tool (Wired, New York Times, Vox Media, OpenWeb, Disqus). Instead of adding them individually as users, annotators may group them and annotate in the following way: Describe the entities as a group, such as 'Various online media outlets'. Select 'Named Entity' as True, and list the names in the entity notes.
    As a rule of thumb, annotators may group named entities when there are <u>four or more</u> entities of the same type and relationship to the AI system.

- **Incidents with multiple harms:** In a single incident, multiple entities can experience tangible harm, or intangible harm. Additionally the same entity could experience multiple harms. During annotation, you will provide both 1) an overarching harm category, 2) a harm category for each unique combination of entity and harm.

  - **Notional incident with multiple harms:** Consider an incident where an autonomous vehicle crashed, narrowly avoiding a pedestrian (Jane Doe) and injuring the driver (John Smith), who then had to miss a month of work, lost his job, and went into medical debt. In this example, Jane Doe experienced an AI tangible harm near miss for physical injury. John Smith experienced two AI tangible harms events, physical injury and financial loss.

  - **Overarching intangible harm category:** When a single incident has harms with different assessed harm levels, record the most severe or highest level of harm for the overarching harm. For the notional examples above, the overarching harm category would be 'AI tangible harm event' although both 'AI tangible harm event'

and ‘AI tangible near-miss’ occurred.

    - **Combinations of entities and harms:** You will be asked to annotate the harm category for each unique combination of entity and harm. Thus, for the above notional example there would be 3 combinations: 1) John Smith experienced AI tangible harm that was a physical injury, 2) John Smith experienced AI tangible harm that was financial loss, and 3) Jane Doe experienced an AI tangible harm near-miss that was a physical injury. If there are unnamed entities, all of which experienced the same harm, lump them all together with a descriptive term (e.g. passengers, pedestrians, job applicants, etc.) and provide a single overarching AI tangible harm category for the group.

- **Recording harmed parties and financial loss when there is damage to property.** It is often possible to estimate a financial loss to the owner when a property is damaged. In these cases it is important to not double count a harm as both damage property and financial loss. For annotation purposes the harmed entity was the damaged property (product, privately owned space, etc) and the tangible harm is the property damage cost. This does not mean that any damage to property can not result in a second occurrence of harm. For example, if the AI in an autonomous vehicle caused damage to a mailbox and then many customers canceled contracts for vehicle purchases, then there would be 1) property damage to the mailbox and 2) financial loss to the vehicle producer. In this case, the property cost of the mailbox would be recorded separately from the financial cost to the vehicle producer.

- **Harm when financial loss is refunded or compensated.** Even if the harmed entity is refunded or compensated for the harm, CSET still considers the harm to have occurred.

- **Tangible harm quantities:** Where possible, we are interested in quantifying the harms that AI has created. Due to the difficulty of measuring intangible harms, we limit this exercise to tangible harms for now. In Section 8, you are asked to annotate the number of injuries and fatalities associated with each incident. While we primarily aim to capture quantities of harm directly linked to an AI’s behavior, you should provide information on all injuries and deaths that can be linked to the technology, even if AI involvement is unknown.

- **Level of autonomy:** Autonomy is an AI’s capability to operate independently. Levels of autonomy differ based on whether or not the AI makes independent decisions and the degree of human oversight. The level of autonomy does not depend on the type of input the AI receives, whether it is human- or machine-generated.
  Currently, CSET is annotating three levels of autonomy.
    - Level 1: the system operates independently with no simultaneous human oversight.
    - Level 2: the system operates independently but with human oversight, where the system makes a decision or takes an action, but a human actively observes the behavior and can override the system in real time.
    - Level 3: the system provides inputs and suggested decisions or actions to a

human that actively chooses to proceed with the AI's direction.

The degree of independence of an AI or level of human interaction can vary. It is possible for a product with an AI system to have multiple operational modes, being able to switch between different levels of autonomy. When this occurs, try to annotate for the level of autonomy that the AI system was operating in at the time of the incident.

- **Uncertainty if harm occurred:** It may be that for an incident, the annotator does not have enough information to determine if harm has occurred. In these cases, the incident should not be classified as tangible or intangible harm. We should have a high degree of confidence that incidents annotated as 'AI harm' did indeed involve tangible or intangible harm tied to an entity and an AI system.

  However, if a legal or regulatory decision (e.g. a decision on a suit or a finding of guilt, but not an arrest or prosecution) has determined that the harm has occurred or financial damages have been awarded to a harmed party, the annotator should assume that harm has occurred.

- **Harm designation could change over time:** It is possible that at the time of annotation, there is insufficient information to determine if a tangible AI harm event, near-miss, issue, or special interest intangible AI harm occurred. However, after the initial annotation more information may become available which adds clarity. Thus, the designation and characteristics of harm can change.

  For example, if a recruiting algorithm is reported to give preferential treatment to men, it would be a special interest intangible AI harm for differential treatment based upon a protected characteristic. Some testing of the system was done internally with the company's human resources department, but it is unclear if the software was used in actual hiring decisions. With this limited information, it may not be possible to tell if any identifiable, specific candidates were harmed, meaning that the incident could not be an AI tangible harm event, near-miss, or issue. However, if there was a class action lawsuit and the court awards female candidates damages for discrimination, there would then be an additional tangible AI harm beyond the special interest intangible AI harm: financial loss to the sued company.

- **Annotating harmed entities:** Identifying and describing harmed entities can be difficult, because it requires drawing a line between harmed and not harmed that can feel fuzzy, especially when it comes to intangible algorithmic harms. The below points provide guidance for general and edge cases:
  - **In general**, any entity that you add as a harmed entity in section 7 must be described in a way to meet the applicable thresholds for characterizable or potentially identifiable. Named harmed entities should always be listed individually, even if they are part of a characterizable subgroup.
  - **Cases of differential treatment:** for incidents involving differential treatment, annotators should include all groups affected by differential treatment as

individual harmed entities. In addition, any individual or named entity that experiences the differential treatment should be added.

- **Cases of detrimental content:** For incidents involving detrimental content, annotators need to distinguish between content that is discriminatory against protected groups, such as hate speech, and content that is harmful in other ways, like disinformation. In the latter case, only people directly (at risk of being) exposed to the content should be added as harmed entities. For example, suppose malicious actors use a generative AI to produce fake news articles and distribute them via Facebook. Facebook users are (at risk of being) exposed to this disinformation and are therefore a harmed entity. Harm is not considered to extend to people beyond the platform. However, if the content that is distributed, for example, were to depict members of a religious group in derogatory and stereotypical ways, it is harmful to all members of this group, not just those on Facebook.
- **Grouping harmed entities:** Some AI systems exhibit bias towards many different groups at once. In order to limit the workload of annotation, annotators may group them under a single entity, where they list the details of the harm distribution in the notes. For example, suppose an AI system is found to be biased against Muslims, women, people with disability and Hispanics. Annotators might create an entity called 'Affected groups with protected characgteristics', indicate harm by differential treatment and list the specific groups in the entity notes. As a rule of thumb, harmed entities should be grouped when there are more than 3 experiencing the same harm.
- **Individual, collective and societal level harms from AI**: We recognize that the harms of algorithmic systems often extend beyond those that directly experience them and may have collective and societal level impacts[6]. The purpose of our taxonomy is to create a dataset of harms from AI by drawing on individual incidents, in order to enable analysis at an aggregate level that will hopefully inform the study of collective harms. This means that for each individual incident we limit the annotation of harmed entities to those described as directly affected (or at direct risk of being affected) in the incident reports and do not extrapolate to collective entities. For example, although the impacts of misinformation on social media platforms can be felt offline, unless such repercussions are explicitly described in an incident report, our taxonomy considers only users of the disseminating platform as harmed entities.

- **Incident locations in disputed territories:** The naming of disputed territories is a complex and politically sensitive issue that often involves conflicting claims and differing perspectives from various parties involved. Annotators should follow guidelines by the United Nations. The UN provides guidelines on naming conventions related to (i) member states as well as (ii) non-member observer states, which are recognized as sovereign states by the UN General Assembly. A list of UN member states is found here,

[6] Smuha (2021) Beyond the individual: governing AI's societal harm. Internet Policy Review, 10(3). https://doi.org/10.14763/2021.3.1574

and observer states is found here. The appropriate naming convention for each sovereign state is listed in the UN Multilingual Terminology Database here. With regard to disputes territories, UN guidelines recommend referencing maps produced by UN geoservices, found here, which delineate jurisdictions based on current international standards. If desired, annotators can provide additional disputed territory names in the appropriate notes field.

- **Types of harm:** Below is a list of the types of harms most likely to occur during annotation. Because this is not an exhaustive list, CSET does have an 'other' category.

  CSET's list of harms:
  - **Physical health/safety:** This includes death, injury, or a reduction in lifespan.
  - **Financial loss:** This includes any loss which is economic in nature and not consequent upon injury or damage. It is the inability to keep, have, or get something that is monetary in nature. For example, this would include loss of wages from wrongful detainment.
  - **Physical property:** The damaging of a physical object. Note this does not include damage to the environment, which has a separate designation.
  - **Intangible property:** It is possible to experience tangible harm (financial loss) from the loss of intangible property. Examples of intangible property include patents, reputation, copyrights, trade secrets. Within the AIID reports, this should be uncommon and would only be an AI tangible harm even if there were legal actions or damages indicating harm to intangible property had occurred. This high bar for harm to intangible property is necessary, otherwise it could be claimed that almost any negative report about a company's AI resulted in reputation damage.
  - **Infrastructure:** Infrastructure includes the network of roads, railways, utilities, and buildings necessary to maintain commerce, transportation, political structures, and the normalcy in daily life. Infrastructure can be harmed through destruction, diminished capability, or reduced effectiveness.
  - **Natural environment:** Pollution is the most likely type of environmental harm that an annotator will encounter.
  - **Violation of human rights, civil liberties, civil rights, or democratic norms:** The definition of these are covered under the terms section. It can often be difficult for an annotator to differentiate between violations of civil liberties, civil rights, human rights, and democratic norms. For this reason we grouped them together.

    Common examples (non-exhaustive list) are
    - The restriction of access to government services or benefits
    - Election interference

- Invasion of privacy, which can include theft of likeness or identity, public disclosure of private facts, or portraying someone in a false light.
- Censorship
- Bias in employment, imprisonment, or medical treatment based upon race or gender.

Civil rights, civil liberties, and human rights are distinct concepts that relate to the protection and promotion of fundamental rights and freedoms. While bias or discrimination more generally may involve unjust or unequal treatment, civil rights, civil liberties, and human rights are grounded in legal, constitutional, and international frameworks.[7] They are specifically designed to protect individuals and communities from discrimination and uphold the principles of equality, fairness, and dignity. Some common areas where violations of civil rights, civil liberties, and human rights can occur include:

- Government Actions: Violations can occur when governments engage in practices such as censorship, surveillance, arbitrary arrests or detentions, denial of due process, or restrictions on freedom of expression, assembly, or association.
- Discrimination and Inequality: Violations of rights and liberties in the context of discrimination and inequality can manifest in unequal access to education, employment, housing, healthcare, and public services. It can also include instances of hate crimes, harassment, or systematic marginalization.
- Criminal Justice System: The criminal justice system can be a domain where violations of rights and liberties occur. This can involve excessive use of force by law enforcement, racial profiling, denial of fair trial, torture, cruel or inhumane treatment, or use of forms of punishment that violate international human rights standards.
- Privacy and Surveillance: In the digital age, privacy and surveillance have become significant concerns. Violations of rights and liberties can occur through unwarranted surveillance, data breaches, or the collection and use of personal information without consent.
- Labor Rights: Violations of rights and liberties can also occur in the context of labor rights. This may involve exploitative working conditions, forced labor, child labor, restrictions on the right to organize and bargain collectively, or denial of fair wages and benefits.

- **Detrimental content:** The definition of this is covered under the terms section. What constitutes detrimental content is very subjective. Instances of detrimental content may lead to tangible harm, but infrequently do so. However, because it is a high-interest harm type, CSET is tracking and annotating it.

[7] US Department of Justice. 2022. "Civil Rights Division at 65: A Report on Recent Cases and Highlights." *Department of Justice.* https://www.justice.gov/crt/page/file/1555981/download.
UN. n.d. "International Human Rights Law." *United Nations.* https://www.ohchr.org/en/instruments-and-mechanisms/international-human-rights-law.

    - **Differential treatment based upon a protected characteristic:** This special interest intangible harm covers bias and fairness issues concerning AI. However, the bias must be associated with a group having a protected characteristic. A bias in the treatment between cats and dogs would not qualify because animal species are not a protected characteristic. Differential treatment based upon protected characteristics often overlaps or occurs with a civil rights violation.
    - **Other tangible harms:** Tangible harms that do not fall under any of the aforementioned categories.
    - **Other intangible harms:** Intangible harms that do not fall under any of the aforementioned categories.

- **AI task or core application area:** The application of an AI is the high level task that the AI is intended to perform. It does not describe the technical methods by which the AI performs the task. Considering what the AI's technical methods enable it to do is another way of arriving at what an AI's application is.

  It is possible for multiple application areas to be involved. When possible pick the principle or domain area, but it is ok to select multiple areas.

  According to the OECD[8], the core application areas are defined as follows:
    - **Human language technologies:** Analyze, modify, produce or respond to human text and speech. Human language technologies may combine tasks like recognition, personalisation and interaction support.
    - **Computer vision:** Feeding digital images and videos into models, machines can identify and classify objects and react to what they "see." Computer vision may include tasks like object recognition and event detection.
    - **Robotics**: A system that contributes to the movement of robots. This involves e.g., recognition and goal-driven optimisation – to perform an action in the real world. Autonomous vehicles are supported by robotic systems.
    - **Automation and/or optimization:** Process automation and/or simulation using structured data such as data mining, pattern recognition, a recommendation system or forecasting/prediction. Numerical optimization may be used to optimize a business process such as scheduling, process controlling or operational research.

[8] Adapted from OECD (2022), "OECD Framework for the Classification of AI systems", *OECD Digital Economy Papers*, No. 323, OECD Publishing, Paris, https://doi.org/10.1787/cb6d9eca-en.

Appendix A: Terms

For the purpose of this document and CSET's annotations, the following terms and definitions apply.

AI: "AI" means the capability of machines to perform functions typically thought of as requiring human intelligence, such as reasoning, prediction, detecting patterns or recognizing and generating coherent natural language. AI includes, but is not limited to, machine learning - a set of techniques by which a computer system learns how to perform a task through recognizing patterns in data and inferring decision rules, rather than through explicit instructions.

Tangible harm: Tangible harm is harm that is observable, verifiable, and definitive. It typically encompasses physical and economic harm through injury (including death), loss, or damage, and may be expressed in monetary terms.

AI System: "AI systems" are technologies and processes in which AI plays a meaningful role. These systems may also include components that do not involve artificial intelligence, such as mechanical components (e.g. self-driving cars and Alexa).

Caused: When we say that an AI system "caused" harm, we mean that it played an important role in the chain of events that led to harm. The AI system doesn't need to be the only factor, or even the major factor, in causing the harm. But it should at least be a "but-for" cause - that is, if the AI system hadn't acted in the way it did, the specific harm would not have occurred.

Imminent risk of harm: Harm would have occurred had it not been for randomness , luck, or atypical intervention that prevented the harm. A risk of imminent harm often, but not always, occurs in an operational environment.

Non-imminent risk of harm: Harm could not have nearly occurred but it could plausibly occur in the future. A risk of non-imminent harm often, but not always, occurs in a controlled or test environment.

Potentially identifiable specific entity: An entity that can be described in detail such that the name (Mr. Joe Smith, Acme Inc, etc.) or a unique identifier (e.g. 100 Main Street, Oakland, California, USA) of the entity could be found. We may not know the name or identifier of the entity from the reports, but it does exist and could be found. For example, the general public is not an identifiable specific entity. However, incarcerated people in the Springfield penitentiary would be specific entities because it would be possible to get a list of all the incarcerated people in the facility.

Civil liberties: CSET defines civil liberties as the personal freedoms that are referenced in the US Bill of Rights. At a high level these are speech, religion, press, assembly, and the right to petition the government. For annotation purposes, this is how CSET is defining civil liberties.

Other organizations could define it differently or use a different reference. Civil liberties, civil rights, human rights, and democratic norms often overlap.

Civil rights: According to the Cornell Legal Information Institute, civil rights are legal provisions that originate from notions of equality and can be enforced by law. They are not in the Bill of Rights but may be established through court decisions and include the right to vote, the right to a fair trial, the right to government services, the right to a public education, and the right to use public facilities. For annotation purposes, this is how CSET defines civil rights. Other organizations could define it differently or use a different reference. Civil liberties, civil rights, human rights, and democratic norms often overlap.

Human rights: According to the United Nations, human rights are rights inherent to all human beings, regardless of race, sex, nationality, ethnicity, language, religion, or any other status. They include the right to life and liberty, freedom from slavery and torture, freedom of opinion and expression, and the right to work and education. CSET uses the UN's Universal Declaration of Human Rights to define human rights. Other organizations could define it differently or use a different reference. Civil liberties, civil rights, human rights, and democratic norms often overlap.

Democratic norms: Democratic norms are traditions, customs, and best practices that support democracy. An example of a democratic norm is accepting election results and facilitating a peaceful transfer of political power. Civil liberties, civil rights, human rights, and democratic norms often overlap. In 2002, the UN Commission for Human Rights declared the items below as essential elements of democracy. While CSET is using this list to define democratic norms, other organizations could define it differently.

- Respect for human rights and fundamental freedoms
- Freedom of association
- Freedom of expression and opinion
- Access to power and its exercise in accordance with the rule of law
- The holding of periodic free and fair elections by universal suffrage and by secret ballot as the expression of the will of the people
- A pluralistic system of political parties and organizations
- The separation of powers
- The independence of the judiciary
- Transparency and accountability in public administration
- Free, independent and pluralistic media

Protected characteristics: Protected characteristics include religion, geography, age, sex, sexual orientation or gender identity, familial status (e.g., having or not having children) or pregnancy, disability, veteran status, genetic information, financial means, race, ideology, nation of origin, citizenship, and immigrant status. In the US, age is a protected characteristic for people over the age of 40. At the federal level, minors are not considered a protected class. For this reason the CSET annotation taxonomy has a separate field to note if a minor was involved.

Characterizable class or subgroup: These are descriptions of different populations of people. Often they are characteristics by which people qualify for special protection by a law, policy, or similar authority. For example, the following characteristics can define subgroups: religion, geography, age, sex, sexual orientation or gender identity, familial status (e.g., having or not having children) or pregnancy, disability, veteran status, genetic information, financial means, race or creed, ideology, nation of origin, citizenship, or immigrant status.

Entity: At a general level, an entity is a person, place, or thing. For AI incidents, common entities are a person, group of people, companies, locations, product, infrastructure, government agency, or natural environment.

AI designed to be harmful: An AI is malicious or designed to be harmful if it is specifically developed or deployed to compromise security; cause or lead to injury, damage or loss, or create conditions where the likelihood of such harm is increased.[9]

Embedded AI: An AI is embedded if it is incorporated into a physical device.

Definitively identified AI: All the tangible harm categories (event, near-miss, and issue) and the special interest intangible harm types require that an AI was clearly involved in the incident versus automation or some other process. Sometimes the information provided is not sufficient to determine if an AI is present. In these cases, no AI harm occurred by CSET's definition. However, if additional information later becomes available that makes an AI's involvement clear, the category could change from no tangible AI harm to a tangible AI harm event, near-miss, or issue. Clear involvement of an AI system does not necessarily require an AI action. Instead, lack of activity that was expected from the AI system may also qualify as clear involvement. For example, if an AI system was designed for safety or to intervene and prevent harm, but failed to do so, then it was clearly linked to the harm.

Directly linked through chain of harm: CSET's definition requires a clear chain of events or mechanism 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. Note that this requirement precludes questions of liability. An AI's behavior may be directly linked to the harm even if it is not considered to be liable for it.

Detrimental content: Detrimental content can include deepfakes, disinformation, cyberbullying, identity misrepresentation, insults, threats of violence, eating disorder or self harm promotion, extremist content, misinformation, sexual abuse material, and scam emails.[10]

[9] Brundage et al. (2018). The malicious use of artificial intelligence: Forecasting, prevention, and mitigation. *arXiv preprint* arXiv:1802.07228.

[10] Based on Banko et al. (2020). A Unified Typology of Harmful Content. Proceedings of the Fourth Workshop on Online Abuse and Harms: 125-137. https://aclanthology.org/2020.alw-1.16.pdf. Watanabe et al. (2018). Hate speech on Twitter: a pragmatic approach to collect hateful and offensive expressions and perform hate speech detection. IEEE Access: 6 (13825–13835).

Intangible harm: Harm that cannot be directly observed. This 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. Note, it is possible for tangible harm to result from intangible harm. For example, 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.

Whether or not intangible harm occurs is often subjective and dependent upon cultural perspectives. Reasonable people might disagree about whether harm has actually occurred (or almost occurred). For this reason, CSET has placed intangible harms into special interest categories. CSET currently only defines and tags three types of special interest intangible harm but recognizes that they do not make up a comprehensive list of intangible harms.

Kawintiranon, K. & Singh, L. (2022). DeMis: Data-efficient misinformation detection using reinforcement learning. In Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases. Grenoble, France.
Singh et al. (2020). Understanding high-and low-quality URL Sharing on COVID-19 Twitter streams. Journal of Computational Social Science, 1-24. https://doi.org/10.1007/s42001-020-00093-6.

Appendix B: Schema and Field Descriptions

Field number	Short field name	Long field name	Description for annotators	Public field description (on incidentdatabase.ai)	Example	Notes
1. Metadata						
1.1	Incident Number	Incident number	The number of the incident in the AI Incident Database.	id	1	
1.2	Annotator	Who is responsible for this annotation?	The ID for the person annotating.	An ID number for the individual who classified this incident according to the CSET taxonomy	001	
1.3	Annotation Status	What is the status of this annotation?	Select from: 1. Annotation in progress 2. Initial annotation complete 3. In peer review 4. Peer review complete 5. In quality control 6. Complete and final	status	6. Complete and final	Once an annotation is marked “Initial annotation complete,” we will assume that any remaining blanks were left deliberately - that is, you looked, but couldn’t find enough information to fill out the blank fields. For this reason, please don’t mark this field “Initial annotation complete” until you’ve truly finished filling it out. When peer review begins, the assigned reviewer should switch the status to “In peer review.” When the review is complete and all comments have been resolved, either the peer reviewer or the original annotator should switch the status to “Peer review complete.” Options 5 and 6 should only ever be selected by the project lead.

1.4	Peer reviewer	Peer reviewer	The ID number for the peer reviewer	An ID designating the individual who classified this incident according to the CSET taxonomy	002	The project lead will assign a peer reviewer to each incident.
1.5	Quality Control	Selected for quality control?	Leave this blank. The project lead uses it to flag a random sample of incidents for additional quality control.	Designates incidents randomly selected to receive additional quality control.	n/a	
2. Incident Domain						
2.1	Physical Objects	Did the incident occur in a domain with physical objects ?	"Yes" if the AI system(s) is embedded in hardware that can interact, affect, and change with the physical objects (roads, factories, medical facilities, etc.). Mark "No" if the system cannot. This includes systems that inform, detect, predict, or recommend. Select ● yes ● no ● maybe	"Yes" if the AI system(s) is embedded in hardware that can interact with, affect, and change the physical objects (roads, factories, medical facilities, etc.). Mark "No" if the system cannot. This includes systems that inform, detect, predict, or recommend.	yes	Context matters. AI systems embedded in hardware that can physically interact are more likely to cause death, injury, or damage.
2.2	Entertainment Industry	Did the AI incident occur in the entertainment industry?	"Yes" if the sector in which the AI was used is associated with entertainment. "No" if it was used in a different, clearly identifiable sector. "Maybe" if the sector of use could not be determined. Select ● yes ● no ● maybe	"Yes" if the sector in which the AI was used is associated with entertainment. "No" if it was used in a different, clearly identifiable sector. "Maybe" if the sector of use could not be determined.	yes	Context matters. AI systems used for entertainment are less likely to result in harm. For example a deepfake used in a movie is less likely to cause harm than a deepfake used for political misinformation.
2.3	Report, Test,	Was the incident	"Yes" if the incident is about	"Yes" if the incident is about a	no	Sometimes there are reports about issues

	or Study of data	about a report, test, or study of data instead of the AI itself?	a report, test, or study of the data and does not discuss an instance of injury, damage, or loss. “Maybe” if it is unclear. Otherwise mark “No.” Select ● yes ● no ● maybe	report, test, or study of the data and does not discuss an instance of injury, damage, or loss. “Maybe” if it is unclear. Otherwise mark “No.”		with the data that could be used to develop AI systems. Since there are mitigation approaches, data issues do not automatically mean that the associated AI will have issues that lead to harm. A projection or hypothesis of the harm resulting from data issues is not sufficient. There must be harm that can be clearly linked to an AI.
2.4	Deployed	Was the reported system (even if AI involvement is unknown) deployed or sold to users?	“Yes” if the involved system was deployed or sold to users. “No” if it was not. “Maybe” if there is not enough information or if the use is unclear. Select ● yes ● no ● maybe	“Yes” if the involved system was deployed or sold to users. “No” if it was not. “Maybe” if there is not enough information or if the use is unclear.	maybe	Systems that are not deployed or sold to users tend to still be in the development stage and hence are less likely to cause harm. However, harm can still be possible.
2.5	Producer Test in Controlled Conditions	Was this a test or demonstration of an AI system done by developers, producers or researchers (versus users) in controlled conditions?	“Yes” if it was a test/demonstration performed by developers, producers or researchers in controlled conditions. “No” if it was not a test/demonstration. “No” if the test/demonstration was done by a user. “No” if the test/demonstration was in operational or uncontrolled conditions. “Maybe” otherwise. Select ● yes ● no ● maybe	“Yes” if it was a test/demonstration performed by developers, producers or journalists in controlled conditions. “No” if it was not a test/demonstration. “No” if the test/demonstration was done by a user. “No” if the test/demonstration was in operational or uncontrolled conditions. “Maybe” otherwise.	no	AI system tests or demonstrations by developers, producers, or researchers in controlled environments are less likely to expose people, organizations, property, institutions, or the natural environment to harm. Controlled environments may include situations such as an isolated compute system, a regulatory sandbox, or an autonomous vehicle testing range.

2.6	Producer Test in Operational Conditions	Was this a test or demonstration of an AI system done by developers, producers or researchers (versus users) in operational conditions?	"Yes" if it was a test/demonstration performed by developers, producers or researchers in controlled conditions. "No" if it was not a test/demonstration. "No" if the test/demonstration was done by a user. "No" if the test/demonstration was in controlled or non-operational conditions. "Maybe" otherwise. Select ● yes ● no ● maybe	"Yes" if it was a test/demonstration performed by developers, producers or journalists in controlled conditions. "No" if it was not a test/demonstration. "No" if the test/demonstration was done by a user. "No" if the test/demonstration was in controlled or non-operational conditions. "Maybe" otherwise.	no	While almost every AI system undergoes testing or demonstration in a controlled environment, some also undergo testing or demonstration in an operational environment. Testing in operational environments still occurs before the system is deployed or sold to end-users. However, relative to controlled environments, operational environments try to closely represent real-world conditions and end-users that affect use of the AI system. Therefore, testing in an operational environment typically poses a heightened risk of harm to people, organizations, property, institutions, or the environment.
2.7	User Test in Controlled Conditions	Was this a test or demonstration done by users in controlled conditions?	"Yes" if it was a test/demonstration performed by users in controlled conditions. "No" if it was not a test/demonstration. "No" if the test/demonstration was done by developers, producers or researchers. "No" if the test/demonstration was in operational or uncontrolled conditions."Maybe" otherwise. Select ● yes ● no ● maybe	"Yes" if it was a test/demonstration performed by users in controlled conditions. "No" if it was not a test/demonstration. "No" if the test/demonstration was done by developers, producers or researchers. "No" if the test/demonstration was in operational or uncontrolled conditions."Maybe" otherwise.	yes	Sometimes, prior to deployment, the user will perform a test or demonstration of the AI system. The involvement of a user (versus a developer, producer, or researcher) increases the likelihood that harm can occur even if the AI system is being tested in controlled environments.
2.8	User Test in Operational	Was this a test or demonstration	"Yes" if it was a test/demonstration	"Yes" if it was a test/demonstration performed	no	The involvement of a user (versus a developer, producer, or researcher)

	Conditions	done by users in operational conditions?	performed by users in operational conditions. “No” if it was not a test/demonstration. “No” if the test/demonstration was done by developers, producers or researchers. “No” if the test/demonstration was in controlled or non-operational conditions.“Maybe” otherwise. Select ● yes ● no ● maybe	by users in operational conditions. “No” if it was not a test/demonstration. “No” if the test/demonstration was done by developers, producers or researchers. “No” if the test/demonstration was in controlled or non-operational conditions.“Maybe” otherwise.		increases the likelihood that harm can occur even if the AI system is being tested. Relative to controlled environments, operational environments try to closely represent real-world conditions and end-users that affect use of the AI system. Therefore, testing in an operational environment typically poses a heightened risk of harm to people, organizations, property, institutions, or the environment.
2.9	Harm Domain	Incident occurred in a domain where we could likely expect harm to occur?	Using the answers to the 8 domain questions, assess if the incident occurred in a domain where harm could be expected to occur. If you are unclear, input “maybe.” Select ● yes ● no ● maybe	Using the answers to the 8 domain questions, assess if the incident occurred in a domain where harm could be expected to occur. If you are unclear, input “maybe.”	maybe	Reflecting upon the previously answered questions, decide if the reported incident or instance occurred in a domain in which harm could possibly occur. This is not a decision on whether or not harm did occur. Just a reflection on the operating conditions or context of the system.
3. AI Tangible Harm Assessment						
3.1	Tangible Harm	Did tangible harm (loss, damage or injury) occur?	An assessment of whether tangible harm, imminent tangible harm, or non-imminent tangible harm occurred. This assessment does not consider the context of the tangible harm,	An assessment of whether tangible harm, imminent tangible harm, or non-imminent tangible harm occurred. This assessment does not consider the context of the tangible harm, if an AI was involved, or	tangible harm definitively occurred	

			if an AI was involved, or if there is an identifiable, specific, and harmed entity. It is also not assessing if an intangible harm occurred. It is only asking if tangible harm occurred and what its imminency was. Select ● tangible harm definitively occurred ● imminent risk of tangible harm (near-miss) did occur ● non-imminent risk of tangible harm (an issue) occurred ● no tangible harm, near-miss, or issue ● unclear	if there is an identifiable, specific, and harmed entity. It is also not assessing if an intangible harm occurred. It is only asking if tangible harm occurred and what its imminency was.		
3.2	AI System	Does the incident involve an AI system?	An assessment of whether or not an AI system was involved. It is sometimes difficult to judge between an AI and an automated system or expert rules system. In these cases select “maybe” Select ● yes ● no ● maybe	An assessment of whether or not an AI system was involved. It is sometimes difficult to judge between an AI and an automated system or expert rules system. In these cases select “maybe”	maybe	Note, over time more information about the incident may become available, allowing a ‘maybe’ to be changed to a ‘yes’ or ‘no.’

3.3	Clear link to technology	Can the technology be directly and clearly linked to the adverse outcome of the incident?	An assessment of the technology's involvement in the chain of harm. It is not an assessment of whether harm occurred, nor if the technology is an AI system. "Yes" if the technology was involved in harm, its behavior can be directly linked to the harm, and the harm may not have occurred if the technology acted differently. "Maybe" if the link is unclear. Otherwise, select "no." Select ● yes ● no ● maybe	"Yes" if the technology was involved in harm, its behavior can be directly linked to the harm, and the harm may not have occurred if the technology acted differently. "Maybe" if the link is unclear. Otherwise, select "no."	yes	For the technology to be directly linked to harm it must have played an important role in the chain of events that led to harm. The technology doesn't need to be the only factor, or even the major factor, in the chain of harm. However, if the technology hadn't acted in the way it did, the specific harm would not have occurred. An occurrence of harm that involves a system which contains an AI is not sufficient for calling an incident and AI harm event, near-miss, or issue. The involved AI must also be directly linked to the harm.
3.4	There is a potentially identifiable specific entity that experienced the harm	There is a potentially identifiable specific entity that experienced the adverse outcome described in incident.	"Yes" if it is theoretically possible to both specify and identify the entity. Having that information is not required. The information just needs to exist and be potentially discoverable. "No" if there are not any potentially identifiable specific entities or if the harmed entities are a class or subgroup that can only be characterized. Select ● yes ● no	"Yes" if it is theoretically possible to both specify and identify the entity. Having that information is not required. The information just needs to exist and be potentially discoverable. "No" if there are not any potentially identifiable specific entities or if the harmed entities are a class or subgroup that can only be characterized.	yes	A potentially identifiable specific entity is an entity that can be described in detail such that the name (Mr. Joe Smith, Acme Inc, etc.) or a unique identifier (e.g. 100 Main Street, Anywhere USA) of the entity could be found. We may not know the name or identifier of the entity from the reports, but it does exist and could be found. For example, the general public is not a potentially identifiable specific entity. However, incarcerated people in the Springfield penitentiary would be specific entities because it would be possible to get a list of all the prisoners in the facility.

3.5	AI Harm Level	Annotator's AI tangible harm level assessment	An assessment of the AI tangible harm level, which takes into account the CSET definitions of AI tangible harm levels, along with the inputs for annotation fields about the AI, harm, chain of harm, and entity. Select ● AI tangible harm event ● AI tangible harm near-miss ● AI tangible harm issue ● none ● unclear	An assessment of the AI tangible harm level, which takes into account the CSET definitions of AI tangible harm levels, along with the inputs for annotation fields about the AI, harm, chain of harm, and entity.	AI tangible harm near-miss	Special interest intangible harm is determined in a different field. The determination of a special interest intangible harm is not dependent upon the AI tangible harm level
3.6	AI Tangible Harm Level Notes	AI tangible harm level notes	If for 3.5 you select unclear or leave it blank, please provide a brief description of why. You can also add notes if you want to provide justification for a level	Notes about the AI tangible harm level assessment	While there was an AI in the system involved in harm, it is unclear if the AI can be directly linked to harm. Although someone was injured by the factory robot with an AI, the injury occurred because of operator error where the person ignored safety procedures.	
4. Special Interest Intangible Harm						

4.1	Impact on Critical Services	Did this impact people's access to critical or public services (health care, social services, voting, transportation, etc)?	Indicate if people's access to public services was impacted. Select ● yes ● no ● maybe	Indicates if people's access to public services was impacted.	yes	Public services include healthcare, social services, voting, public transportation, education, and consumer protection. Note, if 'yes' is selected then there was likely a violation of civil liberties and there was a special interest intangible harm.
4.2	Rights Violation	Was this a violation of human rights, civil liberties, civil rights, or democratic norms?	Indicate if a violation of human rights, civil rights, civil liberties, or democratic norms occurred. Select ● yes ● no ● maybe	Indicates if a violation of human rights, civil rights, civil liberties, or democratic norms occurred.	no	It can often be difficult for the typical annotator to differentiate between violations of civil liberties, civil rights, human rights, and democratic norms. For this reason CSET grouped them together. Human rights are rights inherent to all human beings, regardless of race, sex, nationality, ethnicity, language, religion, or any other status. They include the right to life and liberty, freedom from slavery and torture, freedom of opinion and expression, and the right to work and education. Civil rights are legal provisions that originate from notions of equality and can be enforced by law. Civil liberties are personal freedoms that are referenced in the Bill of Rights. Democratic norms are traditions, customs, and best practices that support democracy. An example of a democratic norm is accepting election results and facilitating a peaceful transfer of political power.
4.3	Involving Minor	Was a minor involved in the incident (disproportionally treated or specifically targeted/affected)	Select ● yes ● no ● maybe	Indicate if a minor was disproportionately targeted or affected	unclear	Generally, governments have an interest in establishing heightened protections for minors. These protections are often associated with media content or privacy. For example, if an AI system illegally tracked a minor's activity online, then answer "yes" to this question. There are instances where an AI system causes indiscriminate harm to a group of people,

						and it is plausible that some of those people are minors. However, in this case the entire group of people, adults and children alike, shared the distribution of harm equally and therefore the answer to this question would be “no.”
4.4	Detrimental Content	Was detrimental content (misinformation, hate speech) involved?	Select ● yes ● no ● maybe		no	Detrimental content can include deepfakes, identity misrepresentation, insults, threats of violence, eating disorder or self harm promotion, extremist content, misinformation, sexual abuse material, and scam emails.
4.5	Protected Characteristic	Was a group of people or an individual treated differently based upon a protected characteristic (e.g. race, ethnicity, creed, immigrant status, color, religion, sex, national origin, age, disability, genetic information)?	Select ● yes ● no ● maybe		no	Protected characteristics include religion, commercial facilities, geography, age, sex, sexual orientation or gender identity, familial status (e.g., having or not having children) or pregnancy, disability, veteran status, genetic information, financial means, race or creed, Ideology, nation of origin, citizenship, and immigrant status. At the federal level in the US, age is a protected characteristic for people over the age of 40. Minors are not considered a protected class. For this reason the CSET annotation taxonomy has a separate field to note if a minor was involved. Only mark yes if there is clear evidence discrimination occurred. If there are conflicting accounts, mark unsure. Do not mark that discrimination occurred based on expectation alone.

4.6	Harm Distribution Basis	If harms were potentially unevenly distributed among people, on what basis?	Indicate how the harms were potentially distributed. Select ● none ● Age ● Disability ● Familial status (e.g., having or not having children) or pregnancy ● Financial means ● Genetic information ● Geography ● Ideology ● Nation of origin, citizenship, immigrant status ● Race ● Religion ● Sex ● Sexual orientation or gender identity ● veteran status ● unclear ● other	Indicates how the harms were potentially distributed.	religion age	Multiple can occur Genetic information refers to information about a person's genetic tests or the genetic tests of their relatives. Genetic information can predict the manifestation of a disease or disorder.
4.7	Notes (special interest intangible harm)	Notes (special interest intangible harm)	Input any notes that may help explain your answers.			
5. AI Special Interest Intangible Harm Assessment						

5.1	Special Interest Intangible Harm	Was there a special interest intangible harm or risk of harm?	An assessment of whether a special interest intangible harm occurred. This assessment does not consider the context of the intangible harm, if an AI was involved, or if there is a characterizable class or subgroup of harmed entities. It is also not assessing if any other intangible harm occurred. It is only asking if a special interest intangible harm occurred . Select ● yes ● no ● maybe	An assessment of whether a special interest intangible harm occurred. This assessment does not consider the context of the intangible harm, if an AI was involved, or if there is a characterizable class or subgroup of harmed entities. It is also not assessing if any other intangible harm occurred. It is only asking if a special interest intangible harm occurred	yes	
5.2	AI System	Does the incident involve an AI system?	An assessment of whether or not an AI system was involved. It is sometimes difficult to judge between an AI and an automated system or expert rules system. In these cases select “maybe” Select ● yes ● no ● maybe	An assessment of whether or not an AI system was involved. It is sometimes difficult to judge between an AI and an automated system or expert rules system. In these cases select “maybe”	maybe	Note, over time more information about the incident may become available, allowing a ‘maybe’ to be changed to a ‘yes’ or ‘no.’
5.3	Clear link to Technology	Can the technology be directly and clearly linked to the adverse outcome of the incident?	An assessment of the technology’s involvement in the chain of harm. It is not an assessment of whether harm occurred, or if the technology is an AI system. “Yes” if the technology was	“Yes” if the technology was involved in harm, its behavior can be directly linked to the harm, and the harm may not have occurred if the technology had acted differently. “Maybe” if the link is unclear. Otherwise,	yes	For the technology to be directly linked to harm it must have played an important role in the chain of events that led to harm. The technology doesn’t need to be the only factor, or even the major factor, in the chain of harm. However, if the technology hadn’t acted in the way it did, the specific harm

			involved in harm, its behavior can be directly linked to the harm, and the harm may not have occurred if the technology acted differently. “Maybe” if the link is unclear. Otherwise, select “no.” Select ● yes ● no ● maybe	select “no.”		would not have occurred. An occurrence of harm that involves a system which contains an AI is not sufficient for calling an incident and AI harm event, near-miss, or issue. The involved AI must also be directly linked to the harm.
5.4	Harmed Class of Entities	There is a characterizable class or subgroup of entities that experienced the harm.	“Yes” if the harmed entity or entities can be characterized. “No” if there are not any characterizable entities. Select ● yes ● no	“Yes” if the harmed entity or entities can be characterized. “No” if there are not any characterizable entities.	yes	A characterizable class or subgroup are descriptions of different populations of people. Often they are characteristics by which people qualify for special protection by a law, policy, or similar authority. Sometimes, groups may be characterized by their exposure to the incident via geographical proximity (e.g., ‘visitors to the park’) or participation in an activity (e.g.,‘Twitter users’). A ‘potentially identifiable specific entity’ is required for AI tangible harm. A ‘characterizable class or subgroup’ is required for AI special interest intangible harm. A ‘characterizable class or subgroup’ is an easier threshold to meet.
5.5	Annotator’s AI special interest intangible harm assessment	Annotator's AI special interest intangible harm assessment	The annotator’s assessment of if an AI special interest intangible harm occurred. Select ● yes ● no ● unclear	The annotator’s assessment of if an AI special interest intangible harm occurred.	no	AI tangible harm is determined in a different field. The determination of a special interest intangible harm is not dependant upon the AI tangible harm level

5.6	Notes (AI special interest intangible harm)	AI special interest intangible harm notes	If for 5.5 you select unclear or leave it blank, please provide a brief description of why You can also add notes if you want to provide justification for a level	Notes about the AI special interest intangible harm level assessment		
6. Environmental and Temporal Characteristics						
6.1	Date of Incident Year	Date of incident Year	The year in which the incident occurred. If there are multiple harms or occurrences of the incident, list the earliest. If a precise date is unavailable, but the available sources provide a basis for estimating the year, estimate. Otherwise, leave blank. Enter in the format of YYYY	The year in which the incident first occurred.	2021	
6.2	Date of Incident Month	Date of incident Month	The month in which the incident occurred. If a precise date is unavailable, but the available sources provide a basis for estimating month, estimate. Otherwise, leave blank. Enter in the format of MM	The month in which the incident first occurred.	12	
6.3	Date of Incident Day	Date of incident Day	The day on which the incident occurred. If a precise date is unavailable, leave blank. Enter in the format of DD	The day on which the first incident occurred.	25	
6.4	Estimated Date	Is the date estimated?	“Yes” if the date was estimated. “No” otherwise.	“Yes” if the date was estimated. “No” otherwise.	yes	If you know the incident year, but not the month or day, select…

			Select ● yes ● no			… No if you are leaving month and day blank. … Yes if you input an estimate for the month. If you use the publication date of the first article, select Yes and add a note.
6.5	Multiple AI Interaction	Was the AI interacting with another AI?	"Yes" if two or more independently operating AI systems were involved. "No" otherwise. Select ● yes ● no ● maybe	"Yes" if two independently operating AI systems were involved. "No" otherwise.	no	This happens very rarely but is possible. Examples include two chatbots having a conversation with each other, or two autonomous vehicles in a crash.
6.6	Embedded	Is the AI embedded in a physical system or have a physical presence?	"Yes" if the AI is embedded in a physical system. "No" if it is not. "Maybe" if it is unclear. Select ● yes ● no ● maybe	"Yes" if the AI is embedded in a physical system. "No" if it is not. "Maybe" if it is unclear.	yes	This question is slightly different from the one in field 2.1.1. That question asks about there being interaction with physical objects–an ability to manipulate or change. A system can be embedded in a physical object and able to interact with the physical environment, e.g. a vacuum robot. A system can be embedded in a physical object and not interact with a physical environment, e.g. a camera system that only records images when the AI detects that dogs are present. AI systems that are accessed through API, web-browser, etc by using a mobile device or computer are not considered to be embedded in hardware systems. They are accessed through hardware.
6.7	Location City	Location City	If the incident occurred at a specific known location, note the city.			If there are multiple relevant locations, enter multiple city/state/country values.
6.8	Location State/Province (two letters)	Location State/Province (two letters)	If the incident occurred at a specific known location, note the state/province.			If there are multiple relevant locations, enter multiple city/state/country values.

6.9	Location Country (two letters)	Location Country (two letters)	If the incident occurred at a specific known location, note the country.			Follow ISO 3166 for the 2-letter country codes. If there are multiple relevant locations, enter multiple city/state/country values. Please follow the above guidelines when annotating an incident that takes place in disputed territory.
6.10	Location Region	Location Region	Select the region of the world where the incident occurred. If it occurred in multiple, choose 'Global'. Select ● Global ● Africa ● Asia ● Caribbean ● Central America ● Europe ● North America ● Oceania ● South America ● unclear	Select the region of the world where the incident occurred.	North America	Use this reference to map countries to regions https://www.dhs.gov/geographic-regions
6.11	Infrastructure Sectors	Which critical infrastructure sectors were affected, if any?	Select ● chemical ● commercial facilities ● communications ● critical manufacturing ● dams ● defense-industrial base ● emergency services ● energy ● financial services ● food and agriculture ● government facilities ● healthcare and public health ● information technology ● nuclear		chemical critical manufacturing	A critical infrastructure sector is affected when the incident impacts its infrastructure physically or disturbs the normal operations in that critical sector. This is different from the sector of deployment, because deployment itself does not imply an impact on the infrastructure. For example, an AI-powered train derailing because its computer vision system does not recognize a signal has an impact on the transport sector. However, a performance-evaluation AI for employees in the retail sector unfairly penalizing staff does not affect critical commercial facilities infrastructure.

			● transportation ● water and wastewater ● other ● unclear			
6.12	Operating Conditions	Operating conditions	A record of any abnormal or atypical operational conditions that occurred. If there are no known abnormal conditions leave blank. *If there are multiple conditions, list them separated with semicolons.*	A record of any abnormal or atypical operational conditions that occurred.	raining; night; low visibility	This field is most often blank
6.13	Notes (Environmental and Temporal Characteristics)	Notes (Environmental and Temporal Characteristics)	Input any notes that may help explain your answers.			
7. Characterizing Entities and the Harm In a single incident multiple entities can experience harm. Additionally the same entity could experience multiple types of harm, which could be at different harm categories (intangible harm, tangible harm near-miss, etc.). For each unique combination of entity and harm you will input: 1) Entity 2) Named Entity indicator 3) Entity type 4) Entity relationship to AI 5) The harm level 6) The type of harm Additionally, all entities mentioned in the reports and involved in the harm will have #1-4 inputs. For #5 and #6 'not applicable' should be selected when the entity is not the harmed party.						
7.1	Entity	Entity	A short 1 to 2 word description of the entity. When possible use a proper name for the entity, making it a Named Entity.	A short 1 to 2 word description of the entity that experienced the harm. When possible use a proper name for the entity, making it a Named Entity.	**Non-named entity examples:** passenger owner students welfare applicants **Named entity**	Annotate information for each entity involved in the report. Try to capture every entity directly linked to the harm. Think about the entity that experienced the harm, all of the entities between them and the AI, and then all of the entities involved in producing and deploying the AI.

					examples: John Smith Jane Doe Xerox Amazon	Employees representing a company in a media or public relations capacity should not be included as an entity.
7.2	Named Entity	Named Entity Indicator	"Yes" if the entity is a named entity. "No" otherwise. Select ● yes ● no	Indicates if the entity is a Named Entity.	no	
7.3	Entity type	Entity type	Indicate the type of entity Select ● individual ● group of individuals ● for-profit organization ● non-profit organization ● government entity ● privately owned space ● public space ● infrastructure ● social or political system ● product ● other ● unclear	Indicates the type that best describes the entity	individual	If multiple selections could characterize the entity, select the primary function of the entity.
7.4	Entity Relationship to the AI	Entity Relationship to the AI	Indicate the entity's relationship to the AI. Select ● developer ● deployer ● government oversight ● user ● Affected non-user ● AI ● geographic area of use ● researcher ● product containing AI ● Watchdog	Indicates the entity's relationship to the AI.	user	You should only provide your own description of the relationship if none of the predefined categories is adequate. Note, the smallest possible chain of harm has just two elements; an AI and an entity experiencing harm, near-miss, or issue.

			● collaborator Or enter a short (1-2 words) description of the relationship.			
7.5	Harm Category Experienced	Was an AI special interest intangible harm, tangible harm event, near-miss, or issue experienced by this entity	For each recorded entity, indicate the harm category that they experienced. If the recorded entity does not experience the intangible harm or the tangible harm event, near-miss, or issue, 'not applicable' should be selected. Select ● AI special interest intangible harm ● AI tangible harm event ● AI tangible harm near-miss ● AI tangible harm issue ● Other harm not meeting CSET definitions ● not applicable ● unclear	For each recorded entity, indicate the harm category that they experienced. Because recorded entities have a variety of roles in the AI incident, not every recorded entity will experience harm.	AI tangible harm near-miss	If there is only 1 instance of harm, this will be the same as the entry in 2.8. However, it is possible for there to be multiple harms or special interest intangible harm for the same incident. When this happens each unique entity-harm combination should have their own entity section filled out.
7.6	Harm Type Experienced	Type of harm experienced by entity	Only entities experiencing harm should have an assigned type. If the entity did not experience the harm, 'not applicable' should be selected. Choose from: ● physical health/safety ● financial loss ● physical property ● intangible property ● infrastructure ● natural environment ● social or political systems	Indicates the type of harm experienced by the harmed entity	financial loss	See the Guidance section for Annotating section for explanations of the types of harm. It is possible for the same entity to experience multiple types of harm. Each reported combination of harm level and harm type needs to be recorded.

			● violation of human rights, civil liberties, civil rights, or democratic norms ● detrimental content ● disproportionate treatment based upon a protected characteristic ● other tangible harm ● other intangible harm ● not applicable			
7.7	Notes (Characterizing Entities and the Harm)	Notes (Characterizing Entities and the Harm)	Input any notes that may help explain your answers.			
8. Tangible Harm Quantities This section captures the number of injuries and deaths occurring in the incident, including those that do not strictly meet CSET's definition of a tangible AI harm event.						
8.1	Lives Lost	How many human lives were lost?	Indicate the number of deaths reported Input: ● a number if lives were lost ● '0' if no lives were lost or if the question is not relevant to that type of harm. ● Leave empty if the number is unknown	Indicates the number of deaths reported	2	This field cannot be greater than zero if the harm is anything besides 'Physical health/safety.'
8.2	Injuries	How many humans were injured?	Indicate the number of injuries reported. Input: ● a number if it is known how many were injured ● '0' if there were no injuries or if the question is not relevant to that type of harm. ● Leave empty if the	Indicates the number of injuries reported	3	This field cannot be greater than zero if the harm is anything besides ● Physical health/safety All reported injuries should count, regardless of their severity level. If a person lost their limb and another person scraped their elbow, both cases would be considered injuries. Do not include the number of deaths in this count.

			number is unknown			
8.3	Estimated Harm Quantities	Are any quantities estimated?	Select ● yes ● no	Indicates if the amount was estimated.	no	
8.4	Notes (Tangible Harm Quantities Information)	Notes (Tangible Harm Quantities Information)	Input any notes that may help explain your answers.			
9. Information about involved technology						
9.1	AI System Description	Description of the technology involved	Describe the technology involved in the incident in as much detail as the reports will allow.	A description of the technology (when possible)	Recommendation system for images to go along with a poem	A high level description of the technology is sufficient, but if more technical details are available, include them in the description as well.
9.2	Data Inputs	Description of data inputs to the technology	This is a freeform field that can have any value. There could be multiple entries for this field. Common ones include ● still images ● video ● text ● speech ● Personally Identifiable Information ● structured data ● other ● unclear	A list of the types of data inputs for the technology.	still images text	Still images are static images. Video images consist of moving images. Text and speech data are considered an important category of unstructured data. They consist of written and spoken words that are not in a tabular format. Personally identifiable information is data that can uniquely identify an individual and may contain sensitive information. Structured data is often in a tabular, machine readable format and can typically be used by an AI system without much preprocessing. Avoid using ‘unstructured data’ data in this field. Instead specify the type of unstructured data; text, images, audio files, etc. It is ok to use ‘structured data’ in this field. Record what the media report explicitly states. If the report does not explicitly state an input modality but it is likely that a particular kind of input contributed to the

						harm or near-miss, record that input. If you are still unsure, do not record anything.
9.3	Sector of Deployment	Sector of deployment	Indicate the sector in which the technology is deployed There could be multiple entries for this field. ● agriculture, forestry and fishing ● mining and quarrying ● manufacturing ● electricity, gas, steam and air conditioning supply ● water supply ● construction ● wholesale and retail trade ● transportation and storage ● accommodation and food service activities ● information and communication ● financial and insurance activities ● real estate activities ● professional, scientific and technical activities ● administrative and support service activities ● public administration ● defense ● law enforcement ● Education ● human health and social work activities ● Arts, entertainment and recreation ● other service activities ● activities of households as employers ● activities of extraterritorial organizations and bodies ● other	Indicates the sector in which the technology is deployed	water supply construction	

			● unclear			
9.4	Public Sector Deployment	Public Sector Deployment	Indicate whether the technology is deployed in the public sector Select ● yes ● no ● maybe	Indicates whether the technology is deployed in the public sector	yes	The public sector is the part of the economy that is controlled and operated by the government.
9.5	Autonomy Level	Autonomy1: Does the system operate independently, without simultaneous human oversight, interaction or intervention? Autonomy2: Does the system operate independently but with human oversight, where the system makes decisions or takes actions but a human actively observes the behavior and can override the system in real time? Autonomy3: Does the system provide inputs and suggested decisions to a human that	Indicate the level of autonomy the system operated in at the time of the incident Select ● Autonomy1 ● Autonomy2 ● Autonomy3 ● unclear	Indicates if the system operates independently with no human oversight, interaction, or intervention	Autonomy2	Considering the chain of harm and autonomy level 1, there would be no person between the AI and the entity that experienced the harm. In other words, no human can intervene in the time between an AI system generating outputs and taking action. An example of Autonomy level 1 would be an AI-enabled robotic vacuum system that is designed to work independently and without oversight, cleaning the office floors at night when everyone is gone. In autonomy level 2, there would be a person observing the AI's behavior between the AI and the entity that experienced the harm. There is a possibility that the overseeing person could stop an action that would otherwise result in harm. In other words, a human could intervene in the time between an AI system generating outputs and taking action. An example of an AI system that exhibits the second level of autonomy is a vehicle with autopilot capabilities. A vehicle on autopilot can drive independently for periods of time, but a human can take control of the vehicle and reorient if needed. To be considered level two autonomy, evidence must show that a human was overseeing the system at the time of the event and the human could

		actively chooses to proceed with the AI's direction?				have or did override the system in real time. In autonomy level 3, there would be a person between the AI and the entity that experienced the harm. While the AI can still be directly linked, the harm would not have occurred if a person had not acted. In other words, a human must have acted in the time between an AI system's computation and taking action. An example of an AI system that exhibits the third level of autonomy is a large language model that is prompted by a user to write a short story. The large language model produces the story, which can then be used as is, edited, or discarded by the user. Mark the autonomy level based on information available in the reports and not based on the design specification of the system. It is possible for a system to have modes that allow it to operate at multiple levels of autonomy. Annotate for the operating autonomy level at the time of the incident.
9.8	Notes (Information about AI System)	Notes (Information about AI System)	Input any notes that may help explain your answers.			
10. AI Functionality and Techniques Leave this section blank if the technology involved in the incident is not an AI system.						
10.1	Intentional Harm	Was the AI intentionally developed or deployed to perform harm? If yes, did the AI's behavior result in	Indicate if the system was designed to do harm and whether or not the system created unexpected harm–i.e. was the reported harm the harm that the AI was expected to perform or a different unexpected	Indicates if the system was designed to do harm. If it was designed to perform harm, the field will indicate if the AI system did or did not create unintended harm–i.e. was the reported harm the harm that AI was expected to perform or a	no	Tracking and analyzing harm from AI systems designed to do harm is valuable and worthwhile. However, analysts may want to separately analyze harm from AI systems that were or were not designed to produce the observed harm.

		unintended or intended harm?	harm? Select ● Yes. Intentionally designed to perform harm and did create intended harm ● Yes. Intentionally designed to perform harm but created an unintended harm (a different harm may have occurred) ● No. Not intentionally designed to perform harm ● unclear	different unexpected harm?		
10.2	Physical System Type	Into what type of physical system was the AI integrated, if any?	Describe the type of physical system that the AI was integrated into.	Describe the type of physical system that the AI was integrated into.	trash sorting robot	
10.3	AI Task	AI task or core application area	It is likely that the annotator will not have enough information to complete this field. If this occurs, enter unclear. This is a freeform field. Some possible entries are ● unclear ● human language technologies ● computer vision ● robotics ● automation and/or optimization ● other	Describe the AI's application.	feature selection	The application area of an AI is the high level task that the AI is intended to perform. It does not describe the technical methods by which the AI performs the task. Considering what an AI's technical methods enable it to do is another way of arriving at what an AI's application is. It is possible for multiple application areas to be involved. When possible pick the principle or domain area, but it is ok to select multiple areas.
10.4	AI tools and methods	AI tools and methods	It is likely that the annotator will not have enough information to complete this	Describe the tools and methods that enable the AI's application.	neural networks	AI tools and methods are the technical building blocks that enable the AI's application.

			field. If this occurs, enter unclear This is a freeform field. Some possible entries are ● unclear ● reinforcement learning ● neural networks ● decision trees ● bias mitigation ● optimization ● classifier ● NLP/text analytics ● continuous learning ● unsupervised learning ● supervised learning ● clustering ● prediction ● rules ● random forest			
10.5	Notes (AI Functionality and Techniques)	Notes (AI Functionality and Techniques)	Input any notes that may help explain your answers.			