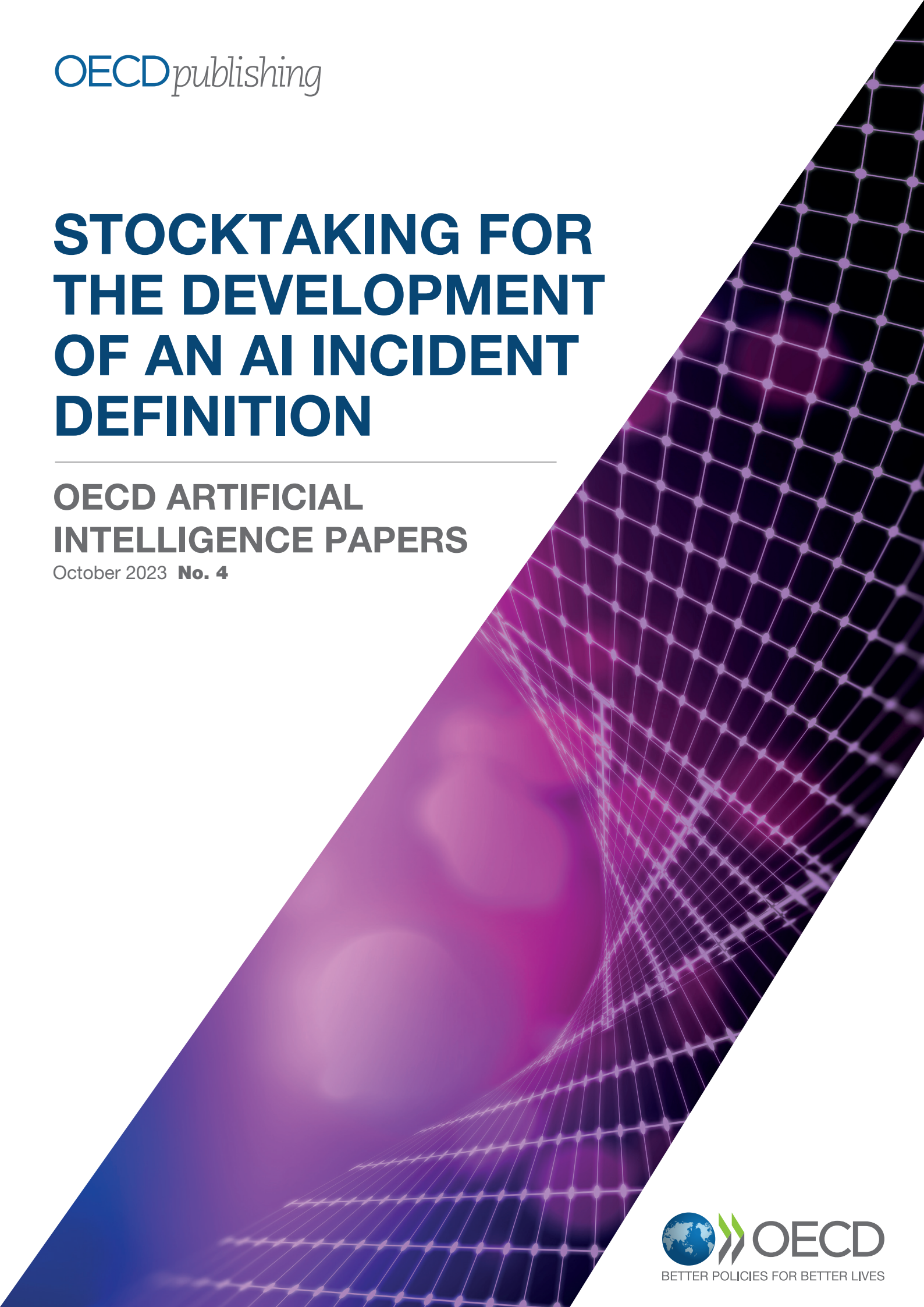
OECD publishing

STOCKTAKING FOR THE DEVELOPMENT OF AN AI INCIDENT DEFINITION

OECD ARTIFICIAL INTELLIGENCE PAPERS

October 2023 **No. 4**



OECD
BETTER POLICIES FOR BETTER LIVES

Foreword

This report takes stock of terminology and practices related to incident definitions, both specific to AI and across other fields. The report provides a foundational knowledge base to identify commonalities and specificities to inform the development of an AI incident definition and related terminology.

This report and previous versions of it were discussed and reviewed by members of the former OECD.AI Expert Group on Classification & Risk during a series of informal workshops between July and October 2022. It was also discussed at the OECD.AI Expert Group on AI Incidents at its March, April and June 2023 meetings and the OECD Working Party on Artificial Intelligence (AIGO) at its April and July 2023 meetings.

The report was written by Karine Perset, Luis Aranda, Orsolya Dobe and Valéria Silva under the supervision of Audrey Plonk, Head of the OECD Digital Economy Policy Division. The report also benefitted from the inputs of delegates for the OECD Working Party on Artificial Intelligence (AIGO), including the Civil Society Information Society Advisory Council (CSISAC) and Business at the OECD (BIAC). Shellie Phillips, Andreia Furtado and Angela Gosmann provided editorial support.

This paper was approved and declassified by the OECD Committee on Digital Economy Policy on 12 October 2023 and prepared for publication by the OECD Secretariat.

Note to Delegations:

This document is also available on O.N.E under the reference code:

DSTI/CDEP/AIGO(2022)11/FINAL

Acknowledgements

This report is based on the work of the OECD.AI Expert Group on AI Incidents and the former OECD.AI Expert Group on Classifying AI systems. It was prepared under the aegis of the OECD Working Party on AI Governance (AIGO). The OECD.AI Expert Group on AI incidents was co-chaired by Irina Orssich (European Commission); Elham Tabassi (National Institute of Standards and Technology), Marko Grobelnik (Jozef Stefan Institute) and Mark Latonero (White House Office of Science and Technology Policy). Luis Aranda and Orsolya Dobe (OECD Digital Economy Policy Division) led the report development and drafting with contributions from Valéria Silva and Kevin Paeth.

Many experts provided invaluable feedback and suggestions, including Sean McGregor (Responsible AI Collaborative); Heather Frase (CSET); Irina Orrssich, Tatjana Evas and Yordanka Ivanova (European Comission). as well as of national delegations to the OECD Working Party on AI Governance, including Roxane Sabourin (Canada); Ferenc Kasa (Hungary); Luis Ricardo Sánchez Hernández (Mexico); and Rose Woolhouse (United Kingdom).

The paper benefited significantly from the oral contributions of those associated with the AI Incidents Expert Group, including Ahmet Yildiz (Ministry of Industry and Technology, Republic of Türkiye; Andrejs Vasiljevs (Tilde); Annalore Verhagen (OECD); Aurelie Jacquet (Standards Australia); Barry O'Brien (IBM); Carlos Ignacio Gutierrez (Future of Life Institute); Carlos Muñoz Ferrandis (BigScience); Catelijne Muller (ALLAI); Coran Darling (DLA Piper); Craig Shank (independent expert); Daniel Schwabe (Catholic University in Rio de Janeiro); Elham Tabassi (NIST); Eva Thelisson (AI Transparency Institute); Florian Ostmann (The Alan Turing Institute); Ilya Meyzin (Dun & Bradstreet); Irina Orssich (European Commission); Jana Novohradska (Slovak Republic); Jessica Cussins (Center for Lon-term Cybersecurity (UC Berkeley)); John McCarthy (Arup); Judith Peterka (Germany); Leonidas Aristodemou (OECD); Mohammed Motiwala (United States); Nicolas Miailhe (The Future Society); Nicolas Moës (The Future Society); Nozha Boujemaa (IKEA); Pam Dixon (World Privacy Forum); Peter Addo (Agence Française de Développement (AFD)); Philip Dawson (independent consultant); Prateek Sibal (UNESCO); Raja Chatila (IEEE); Rayid Ghani (Carnegie Mellon University); Rebecca Anselmetti (United Kingdom); Sean McGregor (AIID and Responsible AI Collaborative); Sebastian Hallensleben (CEN-CENELEC); Yolanda Lannquist (The Future Society); Yordanka Ivanova (European Commission).

The Secretariat would also like to thank stakeholder groups at the OECD for their input, including Pam Dixon (Civil Society Information Society Advisory - CSISAC) and Nicole Primmer and Maylis Berviller (Business at OECD – BIAC).

Finally, the authors thank Orsolya Dobe, Andreia Furtado and Angela Gosmann for editorial support. The overall quality of this report benefited significantly from their engagement.

Table of contents

Foreword 2

Acknowledgements 3

Table of contents 4

Abstract 6

Résumé 7

Executive summary 8

1 Introduction 10
- Background 10
- Timeliness of the project 11
- Goals of a successful AI incident reporting framework and AI incident definition 11
- Complementary project to develop a global AI Incidents Monitor 12

2 Key features of incident definitions 13
- Potential harm 13
- Actual harm 15

3 Dimensions of harm 16
- Types of harm 17
- Severity of harm 19
  - Hazard 20
  - Serious hazard 21
  - Incident 21
  - Accident 22
  - Serious Incident 23

4 Assessment of findings and next steps 24

Annex A. Preliminary findings and notes from past workshops 25

Annex B. Preliminary working definition of “AI incident” 28

Annex C. Participants at OECD informal workshops and consultations on defining AI incidents 30

References 32

Tables

Table 1. Select definitions of risk 13
Table 2. Select definitions of actual harm 15
Table 3. Illustrative dimensions of harm 16
Table 4. Illustrative standards and regulations for different types of harm 18
Table 5. Select definitions of “hazard” 20
Table 6. Select definitions of “incident” 21
Table 7. Select definitions of “accident” 22
Table 8. Select definitions of “serious incident” 23
Table 9: Participation in Singapore’s AI Industry Dialogue on the OECD Global AI Incident Tracker on 29 March 2022 30
Table 10: Participation in the OECD.AI informal workshops on the AI incident strawman definition on 19 July and 11 August 2022 30

Figures

Figure 1. Common AI incident reporting framework 25
Figure 2. Illustration of AI incident concepts as tiers 28
Figure 3. Illustrating key differences between AI incidents, hazards and “near misses” 29

Boxes

Box 1. Preliminary discussions on the definition of AI incidents at the OECD.AI Network of Experts 11
Box 2. An example of a risk-based approach in relevant regulations: The proposed EU AI Act 14
Box 3. Illustrative examples of psychological harms included in legal instruments 18
Box 4. Illustrative example of ongoing efforts to classify AI harms: The CSET Taxonomy 19

Abstract

While AI offers tremendous benefits, it also poses risks. Some of these risks are already materialising into harms, broadly referred to as “AI incidents”. With the ongoing deployment of AI across various sectors, an uptick in AI incidents is expected. To effectively monitor and prevent these incidents, stakeholders require an accurate but flexible definition of AI. This report presents research and findings on terminology and practices related to the definition of an incident, encompassing both AI-specific and cross-disciplinary contexts. It establishes a knowledge foundation to identify commonalities and inspire AI-specific adaptations in terminology going forward.

Résumé

Si l'IA offre des avantages offre des avantages considérables, elle présente également des risques. Certains de ces risques se matérialisent déjà par des préjudices, que l'on appelle généralement "incidents" liés à l'IA. Avec le déploiement de l'IA dans différents secteurs, on s'attend à une augmentation des d'incidents liés à l'IA. Pour surveiller et prévenir efficacement ces incidents, les parties prenantes ont besoin d'une définition précise mais flexible de l'IA. Le présent rapport fait le point sur les recherches et les résultats concernant la terminologie et les pratiques liées à la définition des incidents, tant dans le domaine de l'IA que dans d'autres domaines. Il établit une base de connaissances pour identifier les points communs et inspirer des adaptations de la terminologie spécifiques à l'IA.

Executive summary

AI provides many benefits, but some risks are materialising and causing harms.

While AI provides tremendous benefits, it also poses risks. Some of these risks are already materialising into harms to people, organisations and the environment, like bias and discrimination, the polarisation of opinions, privacy infringements, environmental impacts and security and safety issues. These harms have been broadly referred to under the emerging term of "AI incident". As AI continues to be deployed throughout economies and societies, an increase in AI incidents is inevitable.

A common framework to enable global consistency and interoperability in AI incident reporting.

A common, consistent, framework to report AI incidents would provide the necessary information for policy makers and organisations to learn from AI harms identified elsewhere around the world and help prevent them. It would seek to align AI incident reporting terminology between jurisdictions ahead of the implementation of either mandatory or voluntary AI incident reporting schemes.

What is an "AI incident"?

Common terminology is needed to describe problems or failures of AI systems so that they may be observed, documented, reported, and learned from. These events are broadly referred to under the emerging term "AI incidents". This report takes stock of AI incident definitions and related terminology.

Harm is the starting point to define an incident.

The concept of "harm" is central to technical standards and regulations addressing incidents. Incident definitions often focus on potential harms, actual harms or both. Defining harm and assessing its types, severity levels and other relevant dimensions (e.g. scope, geographic scale, quantifiability, etc.) is key to identifying the incidents that lead or might lead to that harm, and to elaborate an effective framework to address them. Harm definitions and taxonomies are often context-specific.

Defining the type, severity and other dimensions of AI harms is a pre-requisite for developing a definition of AI incidents.

Harms can be of different types (e.g. physical, psychological, economic, etc.) and have different levels of severity (e.g. from inconsequential hazards, to damage to property, to harm to health, to impact to critical infrastructure, to causing human deaths, etc.). Certain aspects of harm may be quantifiable, such as financial loss or number of impacted individuals. Others may be harder to quantify, such as reputational harm. Harm can be tangible, such as physical injury to a person, or damage to property or the environment. Some harms, such as psychological harms, may not be as tangible or readily quantifiable. The scope and scale of harm – whether it affects individuals, organisations, societies, or the environment and whether it

happens at local, national or international level – are also important. Other dimensions of harm include its possible recurrence and reversibility.

Establishing clear terminologies for each dimension of harm is required to enable international alignment.

The criteria by which to categorise AI incidents for each dimension of harm requires a clear mapping of terminology and taxonomies. While some jurisdictions may prefer some terminology over others (e.g. anomaly vs. hazard, accident vs. serious incident, catastrophe vs. emergency, etc.), this framework would aim to facilitate common understanding in incident reporting. Going forward, this work will continue to benefit greatly from the expert input of the OECD.AI network of experts group on AI incidents and from input by communities in adjacent or inter-related areas.

Monitoring actual AI incidents can provide the evidence base to inform this work and AI policy more generally.

To inform the work on developing an AI incident definition and reporting framework, the OECD is developing a global AI incidents monitor (AIM) to collect AI incidents in real time and in a structured manner from available public resources. The AIM will facilitate the identification of AI applications that may need regulatory attention, their impacts and the underlying causes of failure; enable interactive comparison and analysis of AI risks that have materialised or might materialise if no corrective action is taken; help to prevent incidents from re-occurring; and providing evidence for policy-making and regulatory purposes, as appropriate.

1 Introduction

Background

The increasing use of AI systems in real-world contexts around the globe leads to an increase in the number of AI systems that cause real-world harms to people, organisations and the environment. The types and scales of harm posed by AI systems are varied. Among others, an AI incident can vary in the number of people it affects, the number of sectors involved and the degree of the harm it causes.

To foster the development and application of trustworthy AI systems, common terminology is needed to describe the problems or failures of AI systems so that these can be observed, documented, reported and learned from. Broadly, these events can be referred to under the emerging term “AI incidents”.

At the 2nd meeting of the OECD Working Party on Artificial Intelligence Governance (AIGO) in November 2022, the OECD Secretariat presented a concept note on developing a common AI incident reporting framework (OECD, 2022[1]) motivating its timeliness with respect to AI incident reporting mechanisms already being proposed in several jurisdictions and outlining the goals of a successful common framework. The note summarised consultations and preliminary work to date on the topic and presented a preliminary working definition of an “AI incident” considered by the Secretariat and the OECD Expert Group on AI Classification & Risk (Box 1).

In January 2023, the OECD formalised a new *OECD.AI Expert Group on AI Incidents* to further the development of a common AI incident reporting framework and AI Incidents Monitor (AIM). The present Room Document is a preliminary stocktaking of terminology and practices related to defining incidents, both AI-specific incidents and incidents in other fields. It begins to identify commonalities as well as AI-specific adaptations in incident definitions and related terminology.

This work is also expected to inform the work of the *OECD.AI Expert Group on Risk and Accountability* as it seeks to develop a coherent set of due diligence recommendations for AI actors by mapping and analysing AI risk management terminology, actors in the AI value chain, high-priority risks of adverse impacts in the AI sector and gaps in existing risk management frameworks.

Box 1. Preliminary discussions on the definition of AI incidents at the OECD.AI Network of Experts

At the 2nd meeting of the OECD Working Party on Artificial Intelligence Governance (AIGO) in November 2022, the OECD Secretariat presented a concept note on developing a common AI incident reporting framework, motivating its timeliness with respect to AI incident reporting mechanisms already being proposed in various jurisdictions and outlining the goals of a successful common framework. The note also presented a recent working definition of "AI incident" under consideration by the Secretariat and the OECD.AI Network of Experts:

> *AI Incident: an event where the development or use of an AI system:*
>
> *(i) caused harm to person(s), property, or the environment; or*
>
> *(ii) infringed upon human rights, including privacy and non-discrimination.*

In coming to this early working definition, the expert group also discussed the importance of various other terms related to "risk" more generally, such as "hazard," "harm" and other key terms. Review of this work can be found in 4Annex A and 4Annex B. This definition is currently under review by the OECD.AI network of experts.

Source: (OECD, 2022[1])

Timeliness of the project

The potential for AI systems to cause harm is already widely acknowledged. Several initiatives are underway to seek voluntary or mandatory incident reporting in different jurisdictions. In particular, the European Commission's proposed AI Act adopts a risk-based approach to regulate the use of AI systems, where the level of risk reflects the likelihood and severity of the harms they could cause. The proposed EU AI Act classified certain AI systems as high-risk and made them subject to a number of requirements, including post market monitoring obligations and the reporting of serious incidents (European Commission, 2021[2]). Reporting obligations included the malfunctioning of AI systems which results or may result in fundamental rights infringement. An important starting point for developing an AI incident reporting framework will be a widely accepted definition of what constitutes an AI incident and a serious AI incident.

Goals of a successful AI incident reporting framework and AI incident definition

The adoption of a common definition of an AI incident and other key terms can facilitate interoperability between various policy frameworks that seek to monitor AI as a global technology. A successful AI incident reporting framework and definition strive to be:

- **Clear and operational:** Contain clear criteria to allow similar harmful event(s) caused by an AI system to be categorised as an AI incident while facilitating classification into different categories depending on their severity and other characteristics.
- **Actionable and useful:** Help to share best practices and to take action to prevent, mitigate, treat or remedy harm.
- **Modular and flexible:** Provide a common denominator to enable interoperable reporting among different jurisdictions and stakeholder groups and be applicable to both mandatory and voluntary

AI incident reporting frameworks, while still allowing for certain degree of flexibility for jurisdictions to account for their specific need and legal frameworks.

- **Aligned with other incident reporting regimes:** Build on and align with reporting regimes in related fields, such as consumer product safety, cybersecurity incident reporting and aviation accidents. They would also build on lessons learned on AI incidents by stakeholder groups such as academia and civil society, standards-setting organisations and public authorities.
- **Forward-looking**: Be flexible enough to capture new types or incidents in the future that will accompany rapidly evolving technologies.

Complementary project to develop a global AI Incidents Monitor

In parallel, the OECD began a complementary project to develop a global AI Incidents Monitor (AIM) to start tracking actual AI incidents and provide the evidence-base to inform the AI incident reporting framework. The AIM is being informed by the work on the AI incident definition and its taxonomy. In parallel, the AIM seeks to provide a 'reality check' to make sure the definition of an AI incident and reporting framework function with real-world AI incidents. As a starting point, AI incidents reported in reputable international media globally are identified and classified. While recognising the likelihood that they only represent a subset of AI incidents, these publicly reported incidents nonetheless provide a useful starting point in building an evidence base.

Incidents can be composed of one or more news articles covering the same event. To mitigate concerns related to editorial bias and disinformation, each incident's metadata is extracted from the most reputable news outlet reporting on such incident. Additionally, incidents are sorted by the number of articles reporting on them and their relevance to the specific query, as determined by the semantic similarity. Lastly, links to all the articles reporting on a specific incident are provided for completeness.

The data collection and analysis for the AIM is being developed to ensure, to the best extent possible, the reliability, objectivity and quality of the information for AI incidents and that facilitates the classification of AI incidents into different categories depending on their severity, types of harms and other relevant characteristics. This methodology will be made available publicly at the time of launch of the AIM.

In the future, an open submission process will be enabled to complement the AI incidents information from news articles. To ensure consistency in reporting, the existing classification algorithm could be leveraged to process text submissions and provide a pre-selection of tags for a given incident report. Additionally, it is expected that incident information from news articles be complemented by court judgements and decisions of public supervisory authorities wherever they exist.

2 Key features of incident definitions

The concept of "harm" appears as the starting point for nearly every incident definition. It is included in both technical standards and regulations. Once harm is clearly established, in terms of definition, categories and dimensions, the incidents that lead to it and the risk that it materialises may be identified and a framework to address them effectively can be elaborated.

Incident definitions often focus on potential harms, actual harms or both[1]. An AI incident classification system requires clear definitions of the dimensions of potential and actual harm, including types of harm, scope, and severity levels. Facilitating the identification and classification of harms, and determining appropriate responses, is key to the success of any incident reporting framework.

The following sections take stock of potential and actual harm definitions included in key standards and legislation.

Potential harm

Potential harm is often expressed as the risk or likelihood that harm or damage will actually occur. Risk is a function of both the probability of an event occurring and the severity of the consequences that would result. For example, the risk of an explosion in a chemical plant is higher if the plant is located in a densely populated area and the consequences of an explosion would be severe. Per ISO 31000:2018 (Risk management — Guidelines), risk is "usually expressed in terms of risk sources, potential events, their consequences and their likelihood" (International Organization for Standardization, 2018[2]).

Different incident frameworks adopt different definitions of potential harm and risk. These can relate, among others, to the likelihood of causing harm; the severity of the potential harm they may cause; and the nature and origin of the risks (or in a combination of these). Some illustrative definitions of risk are indicated in Table 1. Box 2 provides an illustrative example of a risk-based approach in AI-specific regulation. Identifying and addressing risks and potential harms in AI systems is crucial for risk management and AI incident reporting frameworks.

Table 1. Select definitions of risk

Source	Definition or reference
ISO 31000:2018 Risk management — Guidelines (International Organization for Standardization, 2018[2]) ; ISO/IEC 22989:2022 Information technology — Artificial intelligence (International Organization for Standardization, 2022[3])	"**Risk:** effect of uncertainty on objectives. Note 1: An effect is a deviation from the expected. It can be positive, negative or both, and can address, create or result in opportunities and threats. Note 2: Objectives can have different aspects and categories and can be applied at different levels. Note 3: Risk is usually expressed in terms of risk sources, potential events, their consequences, and their likelihood."
ISO Guide 73:2009 Risk management — Vocabulary (International Organization for Standardization, 2009[4])	"**Risk:** effect of uncertainty on objectives. Note 1 to entry: An effect is a deviation from the expected — positive and/or negative. Note 2 to entry: Objectives can have different aspects (such as financial, health and safety, and environmental goals) and can apply at different levels (such as strategic, organisation-wide, project, product and process). Note 3 to entry: Risk is often characterized by reference to potential events and consequences

	or a combination of these. Note 4 to entry: Risk is often expressed in terms of a combination of the consequences of an event (including changes in circumstances) and the associated likelihood of occurrence. Note 5 to entry: Uncertainty is the state, even partial, of deficiency of information related to, understanding or knowledge of, an event, its consequence, or likelihood." **"Residual risk:** risk remaining after risk treatment. Note 1 to entry: Residual risk can contain unidentified risk. Note 2 to entry: Residual risk can also be known as "retained risk"
ISA-TR84.00.02.2002 Safety Instrumented Functions (SIF) -- Safety Integrity Level (SIL) Evaluation Techniques (The Instrumentation, Systems and Automation Society, 2002[5])	"**Risk**: The combination of the probability of occurrence of harm and the severity of that harm."
NIST Artificial Intelligence Risk Management Framework (RMF) (Tabassi, 2023[6])	"**Risk** refers to the composite measure of an event's probability of occurring and the magnitude or degree of the consequences of the corresponding event. The impacts, or consequences, of AI systems can be positive, negative, or both and can result in opportunities or threats (Adapted from: ISO 31000:2018). When considering the negative impact of a potential event, **risk** is a function of 1) the negative impact, or magnitude of harm, that would arise if the circumstance or event occurs and 2) the likelihood of occurrence (Adapted from: OMB Circular A-130:2016)."
NIST SP 800.30 Rev. 1 Guide for Conducting Risk Assessments (NIST, 2012[7])	**"Risk** is a measure of the extent to which an entity is threatened by a potential circumstance or event, and typically a function of: (i) the adverse impacts that would arise if the circumstance or event occurs; and (ii) the likelihood of occurrence."
ISO/IEC 23894:2023 Information technology — Artificial intelligence — Guidance on risk management (International Organization for Standardization, 2023[8])	Explicitly refers to terms in other standards: "For the purposes of this document, the terms and definitions given in ISO 31000:2018, ISO/IEC 22989:2022 and ISO Guide 73:2009 apply."

Box 2. An example of a risk-based approach in relevant regulations: The proposed EU AI Act

The notion of risk is broadly enshrined in multiple regulations and guidelines on AI. Notably, the proposed EU AI Act takes a "proportionate risk-based approach" to the regulation of AI, wherein AI systems are categorised by the levels of risk they pose: "minimal" or no risk, "limited risk," "high risk" and "unacceptable risk". These designations of AI systems directly affect the obligations of system providers and users under the Act, with many obligations pertaining specifically to "high risk" systems (such as ex post incident reporting requirements) given that "unacceptable risk" systems are prohibited. Additionally, the proposed definition of 'serious incident' in the EU AI Act proposal includes "any incident that directly or indirectly leads, might have led or might lead to any of the following: (a) the death of a person or serious damage to a person's health, to property or the environment, (b) a serious and irreversible disruption of the management and operation of critical infrastructure." Additionally, under the proposed EU AI Act any infringement of fundamental rights is considered a serious incident.

This definition accounts for both serious incidents where harm is latent (potential harm) and where it has materialised (actual harm).

Source: (European Commission, 2021[9]; Council of the European Union, 2022[10]; European Parliament, 2023[11])

Actual harm

The definitions of actual harm in standards and regulation are highly context dependent. They generally focus on physical injury or damage to health, property or the environment. Some standards and regulations, such as the European Union's GDPR, employ the term "damage" to refer to harms (Regulation 2016/679, EU[12]).

Table 2. Select definitions of actual harm

Source	Definition or reference
IEC 61508-1:2010 - Functional safety of electrical/electronic/programmable electronic safety-related systems (International Electrotechnical Commission, 2010[13])	**Harm**: "Death, injury or damage to health, property or the environment."
ISO/IEC Guide 51:2014 - Safety aspects (International Organization for Standardization, 2014[14])	**Harm**: "Injury or damage to the health of people, or damage to property or the environment."
ISO 26262-1:2018 Road vehicles - Functional safety (International Organization for Standardization, 2018[15])	**Harm:** "physical injury or damage to the health of persons."
The General Data Protection Regulation (GDPR) (European Union, 2016[16])	The GDPR uses the term "**damage**" to refer to harms caused by personal data breaches: "Physical, material or non-material damage to natural persons such as loss of control over their personal data or limitation of their rights, discrimination, identity theft or fraud, financial loss, unauthorised reversal of pseudonymisation, damage to reputation, loss of confidentiality of personal data protected by professional secrecy or any other significant economic or social disadvantage to the natural person concerned".
Guide to Protecting the Confidentiality of Personally Identifiable Information (PII) (SP 800-122) (NIST, 2010[17])	NIST Special Publication 800-122 considers a definition of **"harm"** in guidelines on securing information systems with respect to Personally Identifiable Information (PII), defined as "any adverse effects that would be experienced by an individual (i.e., that may be socially, physically, or financially damaging) or an organisation if the confidentiality of PII were breached."

3 Dimensions of harm

The concepts of potential and actual harm described above will be addressed collectively in this Section under the umbrella concept of "harm". In other words, harm in the broad sense, which encompasses both the risks of harm and their materialisation.

Available definitions deal with specific sets of harms, applicable to certain types of harm, with different scopes and scales, among other. Certain aspects of harm may be quantifiable, such as financial loss or number of impacted individuals. Others may be harder to quantify, such as reputational harm. Harm can be physically tangible, such as physical injury to a person, or damage to property or the environment. For example, injury to pedestrians due to the malfunction of an autonomous vehicle AI system would be tangible physical harm. Some harms, however (such as psychological harms) may not be as tangible or quantifiable.

Other aspects of harm that may be intangible or difficult to directly observe include bias or discrimination that may disproportionately and negatively impact particular communities but be difficult to observe at the level of a single individual. Violations of the fundamental right to privacy may also be intangible, such as the non-transparent use of an employee monitoring AI system.

The scope and scale of harm are also important. For example, EUROPOL highlights the possibility of large language models and AI applications like ChatGPT being used to "facilitate the perpetration of disinformation, hate speech and terrorist content online", in addition to providing false objectivity to the messages, and at significantly expanded scale (EUROPOL, 2023[18]). This indicates the possibility of harmful impacts not only to individuals, but at scale: harm can be inflicted to certain groups or to society collectively, as AI may exponentially enhance existing issues as already observed, for example, in the use of algorithms by social media. Among others, AI may provide scale for the dissemination of adverse mental health effects; corrosion of ethical and cultural values; societal polarisation and electoral sway.

Table 3 illustrates some of the possible dimensions of harm that could provide a baseline for further development and discussion on the specificities that an AI incident definition should cover.

Table 3. Illustrative dimensions of harm

Dimensions	Potential criteria for classification
Type	Physical, psychological, reputational, economic/financial (including harm to property), environmental, public interest (e.g. protection of critical infrastructure and democratic institutions), human rights and fundamental rights
Level of severity	Hazard, incident, serious incident, accident, catastrophe; low, medium, high; minor, major, critical; numeric or alphabetical scale
Scope (type of harmed entity)	Individual, group, organisation, institution, society, environment, property
Geographic scale	Single entity, local, national, regional, global
Tangibility	Tangible, intangible
Quantifiability	Quantifiable, unquantifiable
Materialisation	Potential harm (not materialised e.g. hazard), actual harm (materialised e.g. serious incident)
Reversibility	Harm is reversible/irreversible
Recurrence	One-off harms, cumulative effects
Impact	Direct (to individuals), indirect (e.g., to a group, society, environment or public interest; externalities)

Timeframe	Short, medium, long term; within a certain period

In summary, the above discussion highlights the existence of multiple possible *dimensions of harm*, including alternative, competing, overlapping, or complementary dimensions. Different regulations and standards may consider distinct *dimensions of harm*, depending on the specific contexts, goals, areas of impact, and regulatory frameworks in place. Further analysis and discussion about the dimensions of AI harms is key as the work on defining AI incidents unfolds.

The following section provides an initial exploration of the current landscape for two dimensions of harm: *types* of harm and *severity* of harm.

Types of harm

One of the most relevant dimensions of harm when defining an incident is its type. Specific types of harm are included in sectoral and horizontal technical standards and regulations, depending on the goals, context, and industry (Table 4). An AI incident may result in one or multiple of the following types of harm:

- ***Physical harm***: In standards related to product safety or functional safety, physical injury can be categorised according to the type or severity of the injury. For example, the IEC 60950-1 standard for information technology equipment defines physical injury categories as "slight," "moderate," and "severe" (International Electrotechnical Commission, 2010[13]).
- ***Environmental harm***: Some standards categorise harm based on the type of environmental damage caused, such as soil contamination, air pollution, or water pollution. For example, the ISO 14001 standard for environmental management systems includes categories for "minor environmental impact" and "major environmental impact" (International Organization for Standardization, 2015[19]).
- ***Economic or financial harm, including harm to property***: In standards related to financial or economic risk, harm can be categorised based on the magnitude of financial loss or damage. For example, the Basel Framework provides standardised approaches to risk management in the banking sector, addressing risks to credit, market, and operation. (Basel Committee on Banking Supervision, 2017[20]).
- ***Reputational harm***: In standards related to business or organisational risk, harm can be categorised based on the potential impact to an organisation's reputation or public trust in that an organisation. For example, the ISO 26000 standard for social responsibility includes categories for "minor," "moderate," and "major" negative impacts on reputation (International Organization for Standardization, 2010[21]). Individuals may also be affected by reputational harm.
- ***Harm to public interest:*** the International Society of Automation provides the ISA/IEC 62443 Series of Standards, which account for cybersecurity risks that may cause harm to critical infrastructure. It defines levels of security, reliability and integrity (International Society of Automation, 2009[22]). Harm to public interest includes harms to critical infrastructure and functions such as the political system and the rule of law.
- ***Harm to human rights and to fundamental rights:*** these rights are established in domestic and international law. The EU General Data Protection Regulation (GDPR) is a well-known example of a regulation requiring that certain companies carry out impact assessments to identify and manage risks that may cause harm to privacy rights and other fundamental rights and freedoms of natural persons (Regulation 2016/679, EU[12]). In relation specifically to fundamental rights, the Institute of Electrical and Electronics Engineers' – IEEE 7000-2021 standard integrates value-based engineering into the design of AI systems to address risks to human and social values (IEEE, 2021[23]).

- ***Psychological harm*:** Increasing inclusion of psychological harm in standards and product safety legislations reflects a growing recognition of the need to consider the full range of potential impacts of products, services, and business operations on individuals and communities (Box 3). For example, the EU (European Union, 2016[16]) Safety Regulation takes into account the risk to mental health among the health risk posed by digitally connected products. It should be noted that the concept of psychological harm can be more difficult to assess and quantify than physical harm. Further work is needed to specify psychological harm and ways to identify it.

Table 4. Illustrative standards and regulations for different types of harm

Category of harm	Illustrative standard/regulation	Illustrative categories
Physical harm	IEC 60950-1	Categories for "slight," "moderate," and "severe"' physical injury
Environmental harm	ISO 14001	Categories for "minor environmental impact" and "major environmental impact"
Economic harm	Basel Framework	Categories based on the magnitude of financial loss or damage
Reputational harm	ISO 26000	Categories based on the potential impact to reputation or public trust
Harm to public interest	ISA/IEC 62443	Levels of security, reliability, security and integrity for control systems to protect critical architecture
Harm to fundamental rights	IEEE 7000-2021, proposed EU AI Act (European Commission, 2021[2])	N/A
Psychological harm	EU General Product Safety Regulation (European Union, 2023[24])	N/A

Box 3. Illustrative examples of psychological harms included in legal instruments

Psychological harm in the proposed EU AI Act

The notion of harm can also be explicitly extended to psychological health. For example, the European Commission's proposed Artificial Intelligence Act (EU AI Act) would forbid placing on the market an AI system with characteristics that make it more likely to alter a person's behaviour in effect of "psychological harm". Per Article 5 of the currently proposed text, "placing on the market, putting into service or use of an AI system that deploys subliminal techniques beyond a person's consciousness in order to materially distort a person's behaviour in a manner that causes or is likely to cause that person or another person physical or psychological harm" would be prohibited.

Harms to a person's development from child protection legislation

The psychological development of children and young people may be disproportionately affected by AI risks (e.g. manipulation and recommendations of illegal or malicious content) given their developmental vulnerabilities (5Rights Foundation, 2021[25]). Child protection legislation has traditionally included the concept of "development" in its definition of harm: "ill-treatment or the impairment of health or development". The level of harm can be assessed by "comparing a child's health and development with what might be reasonably expected of a similar child".

Risks to the right to form opinions and take decisions independently of automated systems

AI systems have the potential to influence people's opinions and polarise societies at scale. The Council of Europe Declaration by the Committee of Ministers on the manipulative capabilities of algorithmic processes draws attention to "the growing threat to the right of human beings to form opinions and take decisions independently of automated systems, which emanates from advanced digital technologies. Attention should be paid particularly to their capacity to use personal and non-personal data to sort and

micro-target people, to identify individual vulnerabilities and exploit accurate predictive knowledge, and to reconfigure social environments in order to meet specific goals and vested interests" (Council of Europe, 2019[26]).

Mental health risks posed by digitally connected products

In 2023, the European Union enacted the General Product Safety Regulation (GPSR) which repealed the previously used General Product Safety Directive. The regulation highlights the importance of protecting consumers from mental health risks, especially when it comes to the most vulnerable groups such as children. According to the GSPR, the manufacturers of digitally connected products that can affect children will need to meet the highest standards of safety, security and privacy by design to protect them from potential psychological harms (European Union, 2023[24]).

Source: (European Commission, 2021[9]; Children Act 1989, UK[27]; The Children Order 1995, Northern Ireland[28]; Scottish Government, 2021[29]; European Union, 2023[24])

Severity of harm

The OECD.AI Expert Group on AI Incidents has been working to develop a classification of incidents that comprises various harms caused by AI systems, according to their nature and impact. It developed a preliminary classification ranging from the least severe to the most severe harm, in the following order: hazard, near miss, incident and severe incident (Annex B). Over time, experts agreed that "near misses" should be part of "hazards", due to the difficulty of determining in an objective manner when the latter is close to causing an incident but does not actually cause one. As a result, the concepts that are being considered – subject to further analysis – are hazard, serious hazard, incident and serious incident. Each of these concepts are addressed below with illustrative examples of their uses and definitions from existing standards and legal instruments. "Accident" was added to the stocktaking for comparison and discussion. Box 4 provides an illustrative example of complementary efforts to classify AI harms.

Box 4. Illustrative example of ongoing efforts to classify AI harms: The CSET Taxonomy

At present, there is no widely accepted and overarching system of classification to account for all the dimensions of harm caused by AI systems (that is, in relation to the scope, scale, severity, types and impact, to name a few). A step in this direction is the taxonomy developed by Georgetown University's Center for Security and Emerging Technology (CSET). It proposes five levels of severity of harm caused by AI systems:

- '*Negligible'* harm means minor inconvenience or expense, easily remedied.
- '*Minor'* harm means limited damage to property, social stability, the political system, or civil liberties occurred or nearly occurred (e.g. events that do not result in harm or damage but had the potential to do so are also referred to as "near misses"; see 4Annex B).
- '*Moderate'* harm means that humans were injured (but not killed) or nearly injured, or that financial, property, social, or political interests or civil liberties were materially affected (or "nearly" so affected).
- '*Severe'* harm means that a small number of humans were or were almost gravely injured or killed, or that financial, property, social, or political interests or civil liberties were significantly disrupted at least at regional or national scale (or "nearly" so disrupted).

- '*Critical'* harm means that many humans were or were almost killed, or that financial, property, social, or political interests were seriously disrupted at a national or global scale (or nearly so disrupted)."

The CSET taxonomy further classifies harm according to the following categories (or types) of harm: psychological, to physical health/safety, to civil liberties, financial, and to physical property, among others. Additional dimensions of harm include the social group inflicted with harm (e.g. by age, religion, race, sex, national origin, geography and ideology, etc.) and the identification of the companies/ system developers of the AI that caused harm.

Source: (AI Incident Database, 2023[30])

If not quantifiable, incident severity terminology risks being vague and subjective. If quantifiable, it could lead to arbitrary choices in threshold selection, which would also pose international comparability issues (e.g. severity thresholds in US dollars may need to be adjusted depending on the country). Different counties may have diverging considerations when assessing the severity of harms.

Hazard

Hazard commonly refers to something that has the *potential* to cause harm or damage. For example, a hazardous chemical is one that has the potential to cause harm if it is released or comes into contact with living organisms. While there is no AI-specific definition of "hazard," multiple horizontal standards define "hazard" and related terms, such as "hazardous event" or "hazardous situation" (Table 5).

The notion of "hazard" extends beyond harmful materials or self-contained objects. In the context of AI, this means hazards are not simply related to the trained AI model itself, but also to elements of the design, training, and operating context of the AI system. Hazards related to AI systems may be present in any stage of the AI system lifecycle, as modelled in the *OECD Framework for the Classification of AI Systems* (OECD, 2022[31]). For example, the poor quality of data used to train a model may be considered a potential hazard, possibly creating an ineffective or biased AI system that could cause harm in the future. Or the procedural ability of an AI system to make inferences without human review (or a "human-in-the-loop") may be considered a hazard in the operating context of the system. Hazards can also be associated with "anomalies" in some industries and jurisdictions; for example, ISO 26262-1:2018 (Road vehicles - Functional safety) defines a "safety anomaly" as "conditions that deviate from expectations and that can lead to harm" (International Organization for Standardization, 2018[15]), similar to broader definitions of hazard.

Table 5. Select definitions of "hazard"

Source	Definition or reference
ISO/IEC Guide 51:2014 Safety aspects (International Organization for Standardization, 2014[14])	"**Hazard**: potential source of harm"; "**Hazardous event**: event that can cause harm"; "**Hazardous situation**: circumstance in which people, property or the environment is/are exposed to one or more hazards";
ISO Guide 73:2009 Risk management — Vocabulary (International Organization for Standardization, 2009[4])	"**Hazard**: source of potential harm"
ISO 26262-1:2018 Road vehicles - Functional safety (International Organization for Standardization, 2018[15])	"**Hazard:** potential source of harm caused by malfunctioning behaviour of the item." "**Safety anomaly**: conditions that deviate from expectations and that can lead to harm."
IPCC Climate Change 2022: Impacts, Adaptation, and Vulnerability (IPCC, 2022[32])	"**Hazard**: The potential occurrence of a natural or human-induced physical event or trend that may cause loss of life, injury, or other health impacts, as well as damage and loss to property, infrastructure, livelihoods, service provision, ecosystems and environmental resources."

Serious hazard

Serious hazard could be introduced as a further category where there is a high probability that a hazard might lead or might have lead to a serious AI incident. As for hazards, the potential to cause harm or damage has not materialised in serious hazards. In the proposed EU AI Act text, a serious hazard could be considered to be an event that "might have led or might lead to any of the following: (a) the death of a person or serious damage to a person's health, to property or the environment, (b) a serious and irreversible disruption of the management and operation of critical infrastructure" (European Commission, 2021[2]).

Incident

In technical standards, incidents usually refer to *unexpected* or *unplanned* events that have the potential to cause harm or disrupt normal operations of a system, organisation, or process. Incidents may or may not result in harm, but they have the potential to do so. Examples of incidents include equipment malfunctions, power outages, or data breaches. In the AI context, an incident could refer to a wide range of events, including system failures, inaccuracies, and ethical or legal violations. The OECD.AI Network of Experts has held several workshops to discuss the term "incident" (Annex B).

In some incident reporting disciplines, the term "incident" indicates or connotes a lesser severity than the term "accident," or is used to capture all anomalous events other than accidents. Additionally, the term "incident" may also be preferred in contexts where harm may come from *intentional* malicious action by an adversary or other actors, as is often the case in cybersecurity. In such instances, using the term "accident" may be misleading about the nature of harm (i.e. unintentional).

As an information technology, AI systems may be the targets of malicious intention leading to harm (for example, as the target of data breaches or poisoned training data). In addition, AI systems themselves may cause harm based on ethical expectations not shared by all affected parties or be used in outright malicious or illegal purposes. For example, a facial recognition technology deployed by an employer to monitor employees without their consent or knowledge could constitute a potential violation of principles and regulatory frameworks related to the lawful and transparent use of AI systems. The use of AI-generated "deepfake" images and videos could be maliciously intended to harm individuals, organisations, or other parties. These examples would all fall under the "AI incidents" umbrella.

The term "incident" has started to be used in proposed AI-specific regulations and in academic/civil society efforts to document AI harms. These uses of the term "AI incident" resemble the use of the term "incident" in some other disciplines like information security/cybersecurity, but also aviation, business management guidelines, and other fields (Table 6).

Table 6. Select definitions of "incident"

Source	Definition or reference
ISO/IEC/IEE 15288:2015 Systems and software engineering — System lifecycle processes (International Organization for Standardization, 2015[33])	**Incident**: "Anomalous or unexpected event, set of events, condition, or situation at any time during the life cycle of a project, product, service, or system".
OECD Recommendation on Digital Security Risk Management (OECD/LEGAL/0479, 2022)	**Digital security incident:** "Uncertainties which are dynamic in nature and can affect the digital and physical environments, damaging stakeholders' objectives, reputation, human rights and fundamental values such as freedom of expression and privacy, protection of data, trust, economic interests, business operations, physical assets and safety, and affecting competitiveness, well-being, and public welfare".
ISO/IEC 27035-1:2016 Information security, cybersecurity and privacy protection — Information security controls (International Organization for Standardization, 2022[34]); ISO/IEC 27002:2022 Guidelines for Information Security	**Information security incident:** "One or multiple related and identified information security events that can harm an organization's assets or compromise its operations".

Management (International Organization for Standardization, 2022[34])	
NIST Cybersecurity Framework, Version 1.1 (National Institute of Standards and Technology, 2018[35])	**Cybersecurity incident:** "A cybersecurity event that has been determined to have an impact on the organization prompting the need for response and recovery".
ISO 22301:2019 Security and resilience — Business continuity management systems — Requirements (International Organization for Standardization, 2019[36])	**Incident:** "Event that can be, or could lead to, a disruption, loss, emergency or crisis".
United States Code of Federal Regulations, "Transportation" 49 CFR "Incident" (Code of Federal Regulations, 2023[37]); Regulation (EU) No 996/2010, "on the investigation and prevention of accidents and incidents in civil aviation" (Regulation 996/2010, EU[38])	"'**Incident'** means an occurrence other than an accident, associated with the operation of an aircraft, which affects or could affect the safety of operations".
ISA/IEC 62443 Guidelines for Industrial Automation and Control Systems Security (International Society of Automation, 2009[22])	**Incident:** "An event that occurs within or affects an industrial automation and control system that violates the security policy of that system".
The International Nuclear and Radiological Event Scale (International Atomic Energy Agency, 2008[39])	**Incident:** "In the context of the reporting and analysis of events, the word incident is used to describe events that are less severe than accidents. For communicating the significance of events to the public, INES rates events at one of seven levels and uses the term incident to describe events up to and including Level 3. Events of greater significance are termed accidents".
Medical Device Regulation (EU) 2017/745 (Regulation 2017/745, EU[40])	**Incident:** "Any malfunction or deterioration in the characteristics or performance of a device made available on the market, including use-error due to ergonomic features, as well as any inadequacy in the information supplied by the manufacturer and any undesirable side-effect."

Accident

Accident is a term applied across a wide range of contexts and situations. It is sometimes employed as a synonym to "incident". Most of the time, however, it is used as an equivalent to "serious incident".

In disciplines where the term "accident" is used, it often indicates a higher severity of harm than an incident. As such, "accident" is commonly used to refer to an incident that results in harm or damage to people, property, or the environment. Definitions of "accident" oftentimes include injuries, fatalities, or significant property or environmental damage (Table 7). Examples could include car crashes, workplace injuries, or environmental disasters. An accident in the context of AI might refer to a situation where an AI-powered system caused harm to individuals, property or the environment.

The term "accident" may be preferred when the objective is to mitigate the risk of tangible catastrophic failure that is always or almost-always present in a system's standard operation, such as in aviation or nuclear systems.

Table 7. Select definitions of "accident"

Source	Definition or reference
United States Code of Federal Regulations, "Transportation" 49 CFR § "Aircraft Accident" (Code of Federal Regulations, 2023[37])	"**Aircraft accident** means an occurrence associated with the operation of an aircraft which takes place between the time any person boards the aircraft with the intention of flight and all such persons have disembarked, and in which any person **suffers death or serious injury**, or in which the aircraft receives substantial damage".
Regulation (EU) No 996/2010, "on the investigation and prevention of accidents and incidents in civil aviation" (Regulation 996/2010, EU[38])	**Accident: "**An occurrence associated with the operation of an aircraft which, in the case of a manned aircraft, takes place between the time any person boards the aircraft with the intention of flight until such time as all such persons have disembarked, or in the case of an unmanned aircraft, takes place between the time the aircraft is ready to move with the purpose of flight until such time it comes to rest at the end of the flight and the primary propulsion system is shut down, in which: (a) a person is fatally or seriously injured as a result of: […] (b) the aircraft sustains damage or structural failure which [...] the aircraft is missing or is completely inaccessible".
The International Nuclear and Radiological Event Scale	**Accident**: "In the context of the reporting and analysis of events, an accident is an event that has led to significant consequences to people, the environment or the facility. Examples include lethal effects to

(International Atomic Energy Agency, 2008[39])	individuals, large radio-activity release to the environment, reactor core melt. For communicating the significance of events to the public, INES rates events at one of seven levels and uses the term accident to describe events at Level 4 or above. Events of lesser significance are termed incidents."
International Electronical Commission Guide (International Electrotechnical Commission, 2013[41])	**Accident:** "An incident that has occurred and resulted in harm can be referred to as an accident. Whereas an incident that has occurred and that did not result in harm can be referred to as a near miss occurrence."

Serious Incident

Serious incidents are commonly associated with events that cause significant harm, including death and serious damage to property and the environment. Table 8 presents illustrative definitions of "serious incident" from select instruments.

"Serious incident" is not the only term employed to specify a severe degree of harm. Some industries also employ the term "catastrophe" or "major accident" (International Atomic Energy Agency, 1990[42]). In addition, the term "'disaster'' is used by the United Nations Office for Disaster Risk Reduction and the International Federation of Red Cross and Red Crescent Societies, and is defined as "the disruption of the functioning of a community or a society and which may test or exceed its capacity to cope using its own resources'' (UNDRR, 2023[43]; IFRC, 2023[44]).

An alternative term found in the studies conducted is "accident", as mentioned in the previous section. Due to its use in different domains, this term's definitions in standards are briefly covered below.

Table 8. Select definitions of "serious incident"

Source	Definition or reference
Regulation (EU) No 996/2010, "on the investigation and prevention of accidents and incidents in civil aviation" (Regulation 996/2010, EU[38])	"**'Serious incident'** means an incident involving circumstances indicating that there was a high probability of an accident and is associated with the operation of an aircraft, which in the case of a manned aircraft, takes place between the time any person boards the aircraft with the intention of flight until such time as all such persons have disembarked, or in the case of an unmanned aircraft, takes place between the time the aircraft is ready to move with the purpose of flight until such time it comes to rest at the end of the flight and the primary propulsion system is shut down."
Medical Device Regulation (Regulation 2017/745, EU[40])	"**Serious incident**" means any incident that directly or indirectly led, might have led or might lead to any of the following: (a) the death of a patient, user or other person, (b) the temporary or permanent serious deterioration of a patient's, user's or other person's state of health, (c) a serious public health threat; "**Serious public health threat**" means an event which could result in imminent risk of death, serious deterioration in a person's state of health, or serious illness, that may require prompt remedial action, and that may cause significant morbidity or mortality in humans, or that is unusual or unexpected for the given place and time.
European Commission's proposed AI Act (under discussion) (European Commission, 2021[9])	"'**Serious incident'** means any incident that directly or indirectly leads, might have led or might lead to any of the following: (a) the death of a person or serious damage to a person's health, to property or the environment, (b) a serious and irreversible disruption of the management and operation of critical infrastructure." Additionally, under the proposed EU AI Act any infringement of fundamental rights is considered a serious incident.

4 Assessment of findings and next steps

This report takes stock of various incident definitions and related terminology. In doing so, it captures aspects related to potential and actual harms that could result from the development or use of AI systems as opposed to AI-specific harms or hazards only. Each incident framework analysed adopts its own definitions and covers different dimensions of harm, usually with a common focus on the type and severity of harm. For instance, the international classification for nuclear and radiological accidents and incidents adopts seven levels of severity, ranging from anomaly to major accident, while also considering other dimensions of harm, notably areas of impact - people and the environment; radiological barriers and control; and defence-in-depth (International Atomic Energy Agency, 1990[42]).

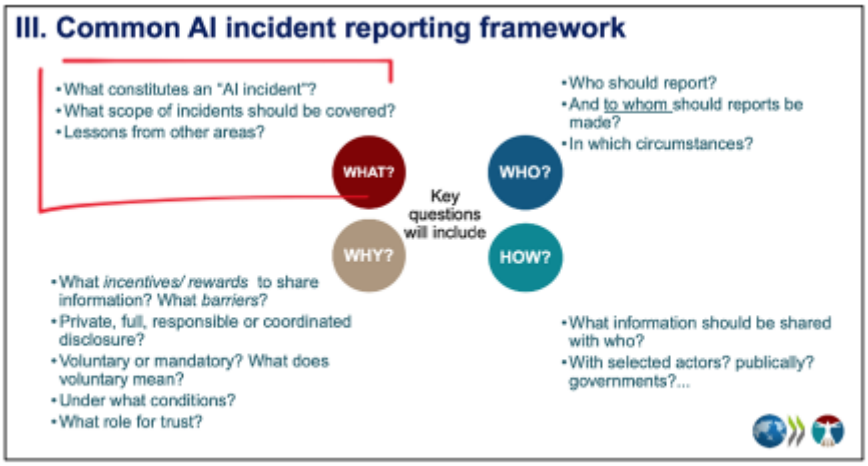
It is proposed that the OECD.AI Expert Group on AI Incidents leverages the findings from this stocktake to develop a definition of AI incidents and related terminology. Additionally, the expert group could conduct further analysis to allow for an appropriate assessment of risks and harms in the AI context. A possible next step in this direction could be to identify the key dimensions of harm that an AI incident definition should cover, such as 'type' of harm, level of severity, scope, geographic scale, tangibility, quantifiability, materialisation, reversibility, recurrence and timeframe. It would also be important to establish clear taxonomies by which to categorise incidents for each of these dimensions and defining appropriate thresholds wherever different levels of severity exist (e.g. serious incidents and serious hazards).

Annex A. Preliminary findings and notes from past workshops

The OECD.AI Expert Group on Classification & Risk began its work on AI risk and incidents in 2020. In addition to consulting member experts of the OECD Network of Experts on Artificial Intelligence, the OECD has sought participation and feedback from specific stakeholders and other parties that are already engaged in work related to AI incident tracking, reporting, and research. One such party is the Responsible AI Collaborative, a non-profit organisation with a founding mission to "identify, define, and catalog artificial intelligence incidents" (Responsible AI Collaborative, 2022[45]) that has created an open-source database of human-reviewed and indexed articles about AI incidents known as the AI Incident Database ("AIID") (AI Incident Database, 2022[46]). Stakeholders involved also include officials from the European Commission (EC), the US National Institute of Standards and Technology (NIST), and the Singapore InfoCom Media Development Authority (IMDA), among others (4Annex C).

Key work started with broader consideration of the challenges and open questions regarding a common incident-reporting framework, including, but not limited to: understanding what constitutes an AI incident, who are the responsible parties to report incidents, to whom reports would be made, incentives to report, and mandatory vs. voluntary reporting frameworks (Figure 1).

Figure 1. Common AI incident reporting framework



Source: OECD, presentation at the 1st meeting of AIGO (25 May 2022)

Discussions prioritised the need to define what constitutes an AI incident. Key workshops and consultations to date have included:

- Singapore AI Industry Dialogue on the OECD Global AI Incident Tracker (29 March 2022)

- Presentation of the strawman definition to OECD.AI Classification & Risk Expert Group (22nd Meeting, 6 July 2022)
- Feedback from the 1st OECD.AI informal workshop on the AI incident strawman definition (19 July 2022)
- Feedback from the 2nd OECD.AI informal workshop on the AI incident strawman definition (11 August 2022)
- Feedback from the OECD.AI Expert Groups on Classification & Risk and Tools & Accountability (23rd Meeting, 22 September 2022)

On voluntary reporting of AI incidents

Workshop participants emphasised that the causal chain of AI incidents may be hard to prove, and clear reporting guidelines distinguishing between what "causes", "nearly causes" or "contributing to" an AI incident each mean. Participants agreed that the goal of voluntary incident reporting is to foster socially beneficial AI while minimising the legal burden on businesses. They put forward that mandatory reporting could discourage industry self-reporting and that therefore "near misses" may not be captured as well in mandatory reporting. For a voluntary system of reporting to work, strong trust among parties is critical given the risks of reputational harm.

Experts also noted that regardless of the reporting regime being voluntary or mandatory in nature, reporting mechanisms should also respect established data governance principles; the responsible parties must establish and maintain administrative, physical and technical security measures for the protection of personal data involved in the reporting of AI incidents, which allow protecting them against damage, loss, alteration, destruction or their unauthorized use, access or processing, as well as guaranteeing their confidentiality, integrity and availability. Analysis of AI incidents allows to identify possible issues and implement corrective and preventive measures. Finally, as the majority of reporting is expected to be generated by those harmed in AI incidents rather than by the industry, it is relevant to establish whistleblowing mechanisms to facilitate reporting.

Possible tiers and reporting scopes of an AI incident definition

Workshop participants agreed that the ability to express and capture events where an AI system "nearly caused" harm in a definition is important to be able to monitor and learn from experience with AI systems. However, they thought that identifying instances where an AI system "nearly caused" harm would require a multiple-tier approach to defining an AI incident. Such events would belong in a broader category of "candidate" incidents. In addition, several partners and experts engaging in AI incident reporting developed or are developing additional "tiers" of events. Additional "tiers" discussed included both:

- "Near miss" events where an AI system "nearly caused" harm, but harm was avoided as a matter of circumstance rather than safety measures; as well as,
- "Hazards", i.e. known AI system deployments that present particular risk of causing AI incidents or are reasonably suspected of having already caused unidentified incidents.

The semantics of such secondary and tertiary tiers is hard to capture in a few words. Many terms have been proposed or commented on by experts (e.g., "candidates," "issues," "hazards," "potential incidents"). Any terms chosen should be consistent with those used in related AI risk management terminology (such as those in ISO/IEC 23894 and ISO 31000) as well as learn from incident reporting in other fields as applicable.

When considering AI incidents, strict or exclusive adherence to a narrow definition of an AI system could neglect other intelligent systems to which decision-making authority is delegated and that can cause similar harms. For example, consider the relatively new advent of "smart contracts" in cryptocurrency systems. These programs do not necessarily use recognised AI techniques but can act autonomously to engage in non-trivial financial trades, resulting in high levels of latent risk that may or may not be transparent to

affected parties. Limiting the scope of the definition to incidents that require regulatory attention might be a desirable approach.

Working with "harm" in an AI incident definition

Workshop participants suggested a focus on defining a process involving impacted stakeholders and responsible parties to assess the presence and/or impact of harm on a case-by-case basis because broad consensus on a succinct meaning of "harm" may be impractical.

Participants also suggested further considerations of how the severity of harm could be treated – either as a component of a core AI incident definition or as part of subsequent classification work for incidents –, noting that severity should likely be used as a variable to determine reporting requirements.

Human rights infringements

Participants highlighted that "harm" should not be limited to "tangible" manifestations (e.g. bodily harm, financial loss) but also include infringements of human rights, including privacy and non-discrimination. However, they acknowledged practical difficulties in providing evidence of intangible (or less tangible) harms.

Several participants suggested a focus on infringements of human rights with legal precedent. In addition, referring to human rights in the definition could cover group-level values-based harms caused by AI systems. However, others cautioned that identifying AI incidents based on the violation of rights may be problematic since it can be difficult to demonstrate tangible harm.

Piloting and evaluation of a definition

Stakeholders emphasized that consideration is given to how to pilot any proposed OECD definition of an AI incident, especially in coordination with parallel work on the AI Incidents Monitor (AIM) project, and to allow for a subsequent opportunity to re-evaluate the definition based on such a pilot. In this manner, an understanding of AI incidents should be based "from the bottom up."

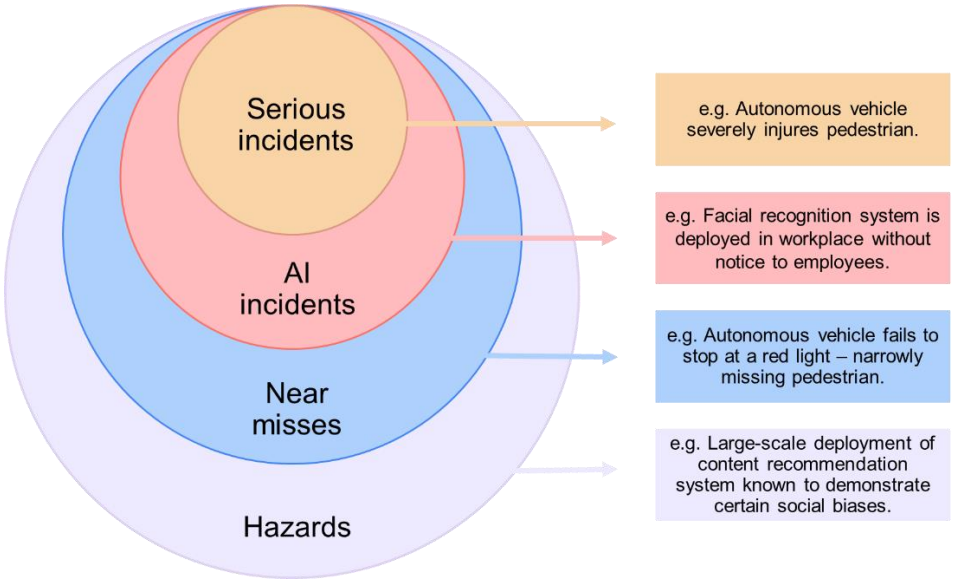
Annex B. Preliminary working definition of "AI incident"

The OECD-developed working proposed definition of an "AI incident" and an "AI hazard" evolved through informal expert workshops and consultations. The working proposals are expected to further evolve and should be read in the context of experts' feedback. The current working proposal of an AI incident discussed by experts is as follows:

> ***AI Incident**: an event where the development or use of an AI system [allegedly]:*
> *(i) caused harm to person(s), property, or the environment;*
> *(ii) including by infringing upon human rights, such as privacy and non-discrimination.*

If the harm involves bodily injury or death, it could be considered to be a **"serious incident"** (Figure 2), such as in the European Commission's proposed AI Act

Figure 2. Illustration of AI incident concepts as tiers



Source: OECD.

In addition to this core definition, different categories of AI-related events could also be considered. Less stringent categories capture more information about AI system deployments and events and could serve as preliminary categorisations for events that may later become AI incidents. Secondary tiers of risks have also been discussed.

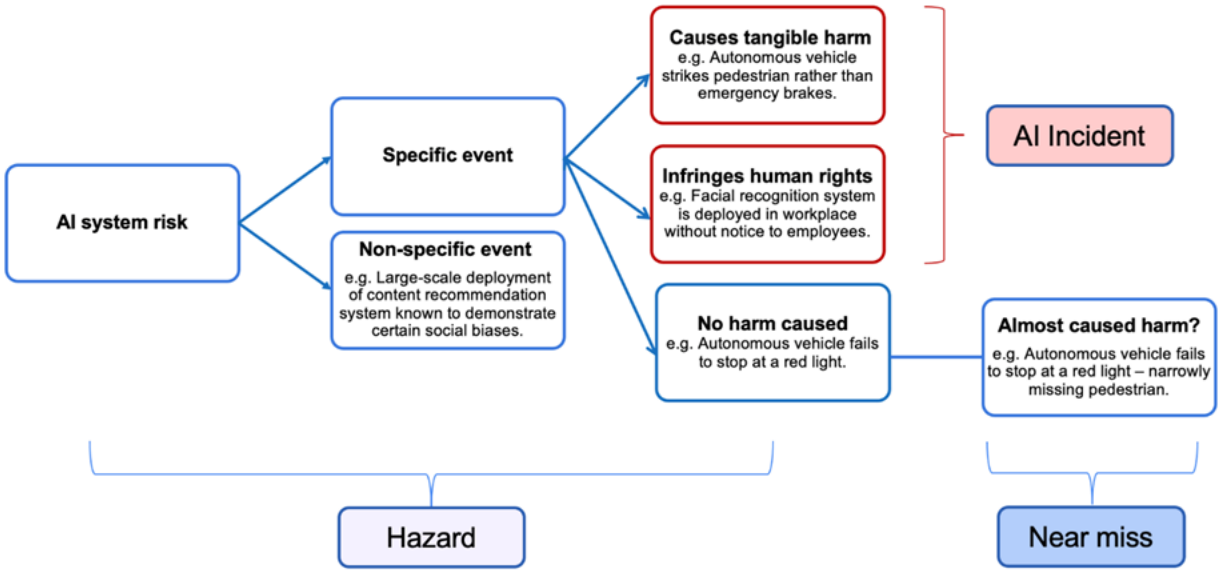
A "***near miss***" is an event where an AI system almost caused harm, but harm was avoided as a matter of circumstance and not due to safety measures.

Secondary tiers of risks also include "AI incident hazards," defined tentatively as follows:

> ***AI incident hazard:*** *a situation where the risk posed by an AI system, or the perception thereof, is relevant to:*
> *(i) a possible harm to person(s), property, or the environment that has yet to occur;*
> *(ii) including an infringement upon human rights, such as privacy and non-discrimination.*

By way of example, autonomous vehicle systems for sale that are known to have defects could be considered **"hazards"** per the working definition above. An event where such an autonomous vehicle fails to stop at a traffic intersection, but does not cause identified harm could be considered a **"near miss"** event. If harm is identified, it could be considered as an **AI incident**. Multiple variables must be considered when classifying AI incidents, for example, whether an applied AI system resulted in a specific identifiable event, whether there was tangible harm or whether harm was avoided altogether (Figure 2. and Figure 3).

Figure 3. Illustrating key differences between AI incidents, hazards and "near misses"



Source: OECD.

Annex C. Participants at OECD informal workshops and consultations on defining AI incidents

Table 9: Participation in Singapore's AI Industry Dialogue on the OECD Global AI Incident Tracker on 29 March 2022

Organisation	Industry
OECD	Intergovernmental Organisation
Co-Chairs of the OECD Working Party on the Classification of AI Systems	Intergovernmental Organisation
IMDA	Government
Alibaba/Lazada	Tech/E-Commerce
X0PA	Tech – HR
Truera	Tech – AI Model Development
Grab	Tech
Data Robot	Tech
Microsoft	Tech
Google	Tech
Singtel	Telecommunications
AT&T/WarnerMedia	Telecommunications/Media
MSD	Healthcare
Changi Airport Group	Transport
Development Bank of Singapore	Banking
Overseas Chinese Banking Corporation	Banking
Standard Chartered Bank	Banking
Union Digital Bank	Banking
MasterCard	Finance
Temasek	Investment
National University of Singapore	Academia
Nanyang Technological University	Academia

Table 10: Participation in the OECD.AI informal workshops on the AI incident strawman definition on 19 July and 11 August 2022

Organisation	Industry
OECD	Intergovernmental Organisation
European Commission	Intergovernmental Organisation
CEN-CENELEC	International Standards Organisation
CEPS	Think Tank
NIST	Government

CSET	Academia
AIID	Tech
XPrize	Tech
Jožef Stefan Institute	Research Institute

References

5Rights Foundation (2021), *Pathways: How digital design puts children at risk*, https://5rightsfoundation.com/uploads/Pathways-how-digital-design-puts-children-at-risk.pdf. [25]

AI Incident Database (2023), *CSET Taxonomy AI Incident Database*, https://incidentdatabase.ai/taxonomy/cset/ (accessed on March 2023). [30]

AI Incident Database (2022), *Artificial Intelligence Incident Database*, https://incidentdatabase.ai/ (accessed on 10 October 2022). [46]

Basel Committee on Banking Supervision (2017), *Bank for International Settlements*, https://www.bis.org/bcbs/basel3.htm. [20]

Children Act 1989 (UK), *Children Act 1989 (UK), c. 41*, https://www.legislation.gov.uk/ukpga/1989/41/section/31A. [27]

Code of Federal Regulations (2023), *830.2 Definitions.*, https://www.ecfr.gov/current/title-49/subtitle-B/chapter-VIII/part-830/subpart-A/section-830.2#p-830.2(Incident) (accessed on March 2023). [37]

Council of Europe (2019), *Declaration by the Committee of Ministers on the manipulative capabilities of algorithmic processes - Decl(13/02/2019)1*, https://search.coe.int/cm/pages/result_details.aspx?ObjectId=090000168092dd4b. [26]

Council of the European Union (2022), *Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts - General approach*, https://artificialintelligenceact.eu/wp-content/uploads/2022/12/AIA-%E2%80%93-CZ-%E2%80%93-General-Approach-25-Nov-22.pdf. [10]

European Commission (2021), *Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts, COM/2021/206 final*, https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A52021PC0206. [9]

European Parliament (2023), *Amendments adopted by the European Parliament on 14 June 2023 on the proposal for a regulation of the European Parliament and of the Council on laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Unio*, https://artificialintelligenceact.eu/wp-content/uploads/2023/06/AIA-%E2%80%93-IMCO-LIBE-Draft-Compromise-Amendments-14-June-2023.pdf. [11]

European Union (2023), *Regulation (EU) 2023/988 of the European Parliament and of the* [24]

Council of 10 May 2023 on general product safety, amending Regulation (EU) No 1025/2012 of the European Parliament and of the Council and Directive (EU) 2020/1828 of the European Parliament an, https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=uriserv%3AOJ.L_.2023.135.01.0001.01.ENG&toc=OJ%3AL%3A2023%3A135%3ATOC.

European Union (2016), *Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data (General Data Protection Regulation)*, http://data.europa.eu/eli/reg/2016/679/oj. [16]

EUROPOL (2023), *ChatGPT - the impact of Large Language Models on Law Enforcement*, Europol Public Information. [18]

IEEE (2021), *IEEE 7000™-2021 Standard - Addressing Ethical Concerns During Systems Design*, https://engagestandards.ieee.org/ieee-7000-2021-for-systems-design-ethical-concerns.html. [23]

IFRC (2023), *What is a disaster?*, https://www.ifrc.org/our-work/disasters-climate-and-crises/what-disaster (accessed on 11 August 2023). [44]

International Atomic Energy Agency (2008), *The International Nuclear and Radiological Event Scale*, https://www-pub.iaea.org/mtcd/publications/pdf/ines2013web.pdf. [39]

International Atomic Energy Agency (1990), *International Nuclear and Radiological Event Scale (INES)*, https://www.iaea.org/resources/databases/international-nuclear-and-radiological-event-scale. [42]

International Electrotechnical Commission (2013), *Electropedia*, https://www.electropedia.org/iev/iev.nsf/display?openform&ievref=903-01-06. [41]

International Electrotechnical Commission (2010), *IEC 61508:2010 CMV - Functional safety of electrical/electronic/programmable electronic safety-related systems*, https://webstore.iec.ch/publication/22273. [13]

International Organization for Standardization (2023), *ISO/IEC 23894:2023 Information technology — Artificial intelligence — Guidance on risk management*, https://www.iso.org/standard/77304.html. [8]

International Organization for Standardization (2022), *ISO/IEC 23894:2022 Information technology — Artificial intelligence — Guidance on risk management*, https://www.iso.org/obp/ui/#iso:std:iso-iec:22989:ed-1:v1:en. [3]

International Organization for Standardization (2022), *ISO/IEC 27002:2022 Information security, cybersecurity and privacy protection — Information security controls*, https://doi.org/10.6028/NIST.CSWP.04162018. [34]

International Organization for Standardization (2019), *ISO 22301:2019(en) Security and resilience — Business continuity management systems — Requirements*, https://www.iso.org/standard/75106.html. [36]

International Organization for Standardization (2018), *ISO 26262-1:2018 Road vehicles — Functional safety — Part 1: Vocabulary*, https://www.iso.org/standard/68383.html. [15]

International Organization for Standardization (2018), *ISO 31000:2018 Risk management — Guidelines*, https://www.iso.org/standard/65694.html. [2]

International Organization for Standardization (2015), *ISO 14001:2015 Environmental management systems — Requirements with guidance for use*, https://www.iso.org/standard/60857.html. [19]

International Organization for Standardization (2015), *ISO/IEC/IEEE 15288:2015 Systems and software engineering — System life cycle processes*, https://www.iso.org/standard/63711.html. [33]

International Organization for Standardization (2014), *ISO/IEC Guide 51:2014 Safety aspects — Guidelines for their inclusion in standards*, International Organization for Standardization, https://www.iso.org/standard/53940.html. [14]

International Organization for Standardization (2010), *ISO 26000 Guidance on social responsibility*, https://www.iso.org/standard/42546.html. [21]

International Organization for Standardization (2009), *ISO Guide 73:2009 Risk management — Vocabulary*, https://www.iso.org/standard/44651.html. [4]

International Society of Automation (2009), *ISA/IEC 62443 Series of Standards*, https://www.isa.org/standards-and-publications/isa-standards/isa-iec-62443-series-of-standards. [22]

IPCC (2022), *Climate Change 2022: Impacts, Adaptation, and Vulnerability,*, Cambridge University Press, https://doi.org/10.1017/9781009325844. [32]

National Institute of Standards and Technology (2018), *Framework for Improving Critical Infrastructure Cybersecurity*, https://www.nist.gov/cyberframework/framework. [35]

NIST (2012), *SP 800-30 Rev. 1 Guide for Conducting Risk Assessments*, https://nvlpubs.nist.gov/nistpubs/Legacy/SP/nistspecialpublication800-30r1.pdf. [7]

NIST (2010), *SP 800-122 Guide to Protecting the Confidentiality of Personally Identifiable Information (PII)*, https://nvlpubs.nist.gov/nistpubs/Legacy/SP/nistspecialpublication800-122.pdf. [17]

OECD (2022), *DSTI-CDEP-AIGO(2022)11 - Concept note on developing a common framework for AI incident reporting & AI incidents monitor*. [1]

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. [47]

OECD (2022), *OECD Framework for the Classification of AI Systems*, https://www.oecd-ilibrary.org/docserver/cb6d9eca-en.pdf. [31]

Regulation 2016/679 (EU), *Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Da*, https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32016R0679&from=EN#d1e1374-1-1. [12]

Regulation 2017/745 (EU), *REGULATION (EU) 2017/745 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL on medical devices, amending Directive 2001/83/EC, Regulation* [40]

(EC) No 178/2002 and, https://www.medical-device-regulation.eu/wp-content/uploads/2019/05/CELEX_32017R0745_EN_TXT.pdf.

Regulation 996/2010 (EU), *REGULATION (EU) No 996/2010 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 20 October 2010 on the investigation and prevention of accidents and incidents in civil aviation and repealing Directive 94/56/EC*, https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A02010R0996-20180911. [38]

Responsible AI Collaborative (2022), *Founding Report*, https://docsend.com/view/47z2whznatd39vf9. [45]

Scottish Government (2021), *National Guidance for Child Protection in Scotland*, https://www.gov.scot/binaries/content/documents/govscot/publications/advice-and-guidance/2021/09/national-guidance-child-protection-scotland-2021/documents/national-guidance-child-protection-scotland-2021/national-guidance-child-protection-scotland-2021/g. [29]

Tabassi, E. (2023), *AI Risk Management Framework*, National Institute of Standards and Technology, Gaithersburg, MD, https://doi.org/10.6028/nist.ai.100-1. [6]

The Children Order 1995 (Northern Ireland), *The Children (Northern Ireland) Order 1995 No. 755 (N.I. 2)*, https://www.legislation.gov.uk/nisi/1995/755/article/2/made. [28]

The Instrumentation, Systems and Automation Society (2002), *ISA-TR84.00.02.2002 Safety Instrumented Functions (SIF) -- Safety Integrity Level (SIL) Evaluation Techniques*, https://www.yumpu.com/en/document/read/11683507/safety-instrumented-functions-sif-safety-integrity-level-isa. [5]

UNDRR (2023), *Disaster: Sendai Framework Terminology On Disaster Risk Reduction*, https://www.undrr.org/terminology/disaster (accessed on 11 August 2023). [43]

Notes

[1] For example, the International Nuclear and Radiological Event Scale (INES) refers to situations that could potentially lead to harm by using terminology like “anomaly”, “incident” and “serious incident”. For actual harm, the INES leverages “accident” and further scales them according to the level of severity as “accidents with local consequences”, “accidents with wider consequences”, “serious accident” and “major accident”.