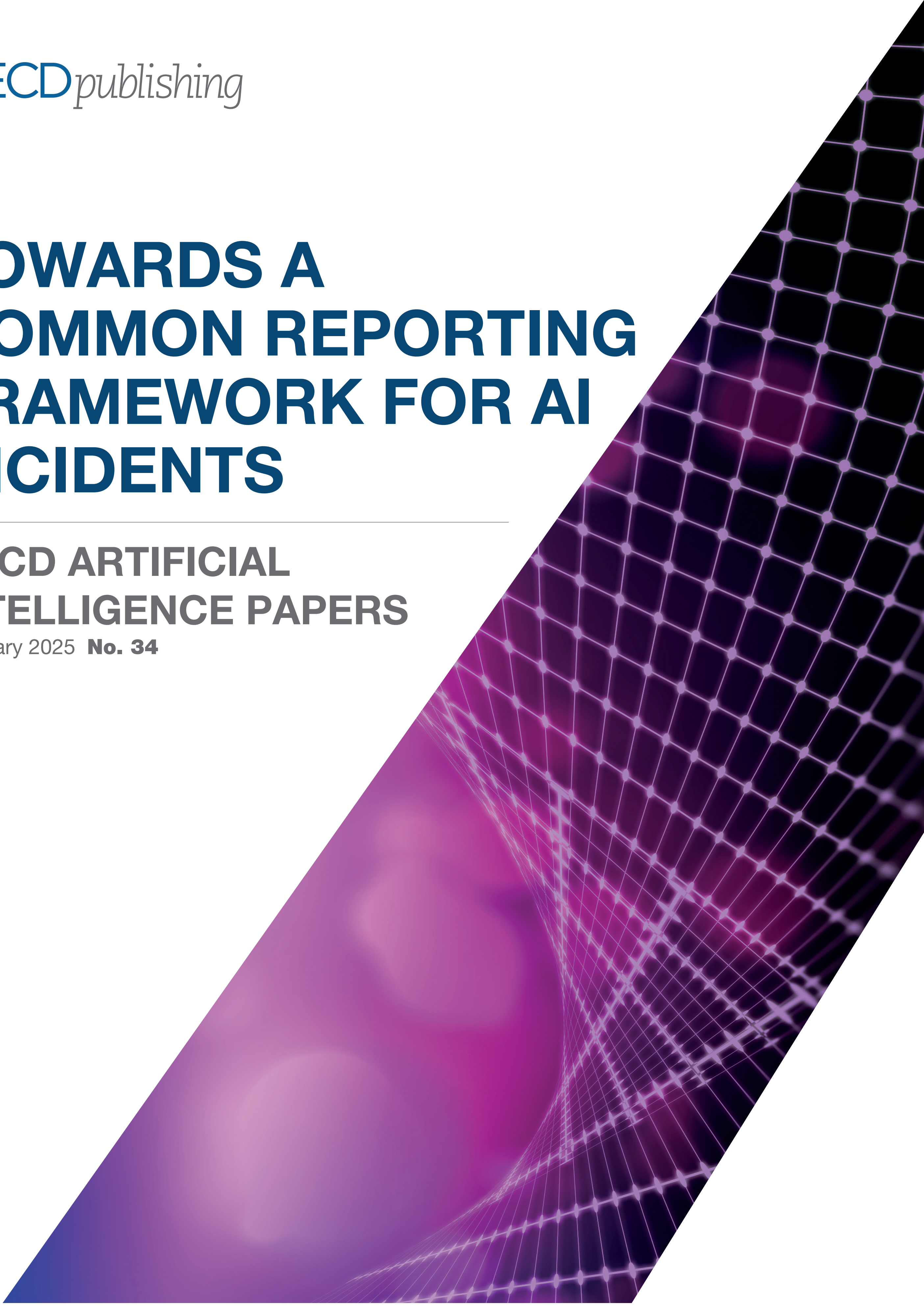
OECD*publishing*

TOWARDS A COMMON REPORTING FRAMEWORK FOR AI INCIDENTS

OECD ARTIFICIAL INTELLIGENCE PAPERS

February 2025 **No. 34**



GPAI

OECD
BETTER POLICIES FOR BETTER LIVES

Foreword

This report presents a common framework to report AI incidents, providing a global benchmark for stakeholders across jurisdictions and sectors. The framework enables countries to adopt a common reporting approach while allowing flexibility in how they respond in accordance with their domestic policies. The framework aims to provide policymakers with a better understanding of AI incidents in diverse contexts, identify high-risk systems, assess their impacts and understand emerging risks.

This report and previous versions of it were discussed and reviewed by members of the OECD.AI Expert Group on AI Incidents at its February, April, June and October 2024 meetings. The OECD Working Party on Artificial Intelligence (AIGO) discussed this report at its June 2024 meeting and the Global Partnership on AI (GPAI) discussed this work during its November 2024 Plenary.

The report was written by Karine Perset, Luis Aranda and Bénédicte Rispal under the supervision of Audrey Plonk, Deputy Director of the OECD Science, Technology and Innovation Directorate. The report also benefitted from the inputs of delegates for the Global Partnership on AI (GPAI), the OECD Working Party on Artificial Intelligence (AIGO), including the Civil Society Information Society Advisory Council (CSISAC), Business at the OECD (BIAC), the Trade Union Advisory Committee (TUAC) and the Internet Technical Advisory Committee (ITAC). John Tarver, Shellie Laffont and Andreia Furtado provided editorial support.

This paper was approved and declassified by written procedure by the Global Partnership on Artificial Intelligence (GPAI) on 13 December 2024 and prepared for publication by the OECD Secretariat.

Note to Delegations:

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

DSTI/DPC/GPAI(2024)5/FINAL

This document, as well as any data and map included herein, are without prejudice to the status of or sovereignty over any territory, to the delimitation of international frontiers and boundaries and to the name of any territory, city or area.

Cover image: © Kjpargeter/Shutterstock.com

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) and the Global Partnership on AI (GPAI). The co-chairs of OECD.AI Expert Group on AI incidents overseeing this work were Irina Orssich (European Commission), Elham Tabassi (National Institute of Standards and Technology), Mark Latonero (U.S. AI Safety Institute) and Marko Grobelnik (Jožef Stefan Institute). Karine Perset, Luis Aranda and Bénédicte Rispal (OECD Artificial Intelligence and Emerging Digital Technologies Division) led the report development and drafting.

The paper benefitted significantly from the oral and written contributions of AIGO and GPAI delegates as well as experts associated with the OECD.AI Expert Group on AI Incidents, including Abhishek Singh (India); Barry O'Brien (IBM); Carlos Ignacio Gutierrez (Future of Life Institute); Craig Shank (independent expert); Coreene White (United States); Daniel Schwabe (Catholic University in Rio de Janeiro); David Turnbull (United States); Debashis Chakraborty (India); Dunja Mladenić (Jožef Stefan Institute); Elham Tabassi (NIST); Florian Ostmann (The Alan Turing Institute); Gerald Hopster (Autoriteit Persoonsgegevens); Heather Frase (verAItech); Irina Orssich (European Commission); Jesse Dunietz (NIST); Jimena Viveros (IQuilibriumAI); Jimmy Farrell (Pour Demain); Judith Peterka (Germany); Julian Frohnecke (Germany); Kevin Paeth (UL Research Institutes); Larissa Lim (Infocomm Media Development Authority); Luis Ricardo Sánchez Hernández (Mexico); Matthew O'Shaughnessy (U.S. Department of State); Marjoleine Hennis (Netherlands); Mark Latonero (U.S. AI Safety Institute); Marko Grobelnik (Jožef Stefan Institute); Melisa Teleki (Republic of Türkiye); Michaël Reffay (France); Nicolas Miailhe (The Future Society); Nicolas Moës (The Future Society); Nobuhisa Nishigata (Japan); Nozha Boujemaa (Decathlon); Pam Dixon (World Privacy Forum); Patrick Gilroy (TÜV Association); Raja Chatila (IEEE); Rob Procter (University of Warwick); Sarah Box (New Zealand); Sean McGregor (Responsible AI Collaborative); Sebastian Hallensleben (CEN-CENELEC); Sharon Ho (Canada); Theodoros Evgeniou (INSEAD); Thiago Guimarães Moraes (Brazil); Till Klein (AppliedAI); and William Bartholomew (Microsoft).

The Secretariat would also like to thank stakeholder groups at the OECD for their input, including Pam Dixon (Civil Society Information Society Advisory – CSISAC); Nicole Primmer and Maylis Berviller (Business at OECD – BIAC); Sarah Jameson and Aida Ponce (Trade Union Advisory Committee – TUAC); and Jibu Elias (Internet Technical Advisory Committee – ITAC).

Finally, the authors thank all those who have contributed to the report throughout its development. This includes Ashini Bamunuvitharana, Claire Marguerettaz, Lawrence Pacewicz, Magali Viard (Directorate for Legal Affairs) and Michael Donohue (General Secretariat). The authors also thank John Tarver, Shellie Laffont and Andreia Furtado for editorial support, the overall quality of this report benefitted significantly from their engagement.

Table of contents

Tables

Boxes

Abstract

This paper presents a common framework for reporting artificial intelligence (AI) incidents that provides a global benchmark for stakeholders across jurisdictions and sectors. The framework enables countries to adopt a common reporting approach while allowing them to tailor responses to their domestic policies and legal frameworks. Through its 29 criteria, the framework aims to help policymakers understand AI incidents across diverse contexts, identify high-risk systems, assess current and emerging risks, and evaluate the impact of AI on people and the planet.

Résumé

Ce rapport présente un cadre commun pour le signalement des incidents liés à l'intelligence artificielle (IA), offrant une référence internationale pour les parties prenantes à travers divers secteurs et juridictions.
Ce cadre vise à faciliter l’harmonisation des signalements des incidents liés à l’IA à l’échelle internationale tout en laissant au pays la possibilité d’adapter leurs réponses conformément à leurs politiques nationales et cadres juridiques. Grâce à ses 29 critères, le cadre vise à aider les décideurs politiques à comprendre les incidents liés à l'IA dans divers contextes, à identifier les systèmes à haut risque, à évaluer les risques actuels et émergents et à évaluer l'impact de l'IA sur les personnes et la planète.

Executive summary

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

Although AI can provide tremendous benefits, it also poses risks. Some of these risks already materialise into harms to people, organisations and the environment, like discrimination, privacy infringements, and security and safety issues. These harms have been broadly referred to under the emerging term “AI incident”. As AI continues to be deployed rapidly throughout economies and societies, an increase in AI incidents is inevitable.

Countries need a common reporting framework to enable global consistency and interoperability in AI incident reporting now as doing so retroactively would be costly and inefficient

A common and consistent framework to report AI incidents and hazards can provide the necessary information for policymakers and organisations to learn from AI harms identified elsewhere in the world, thereby preventing similar incidents from occurring again. It could align AI incident reporting across jurisdictions before implementing AI incident reporting schemes. Pursuing reporting alignment is urgent, as a retroactive approach would prove costly and inefficient.

Defining the most relevant criteria for reporting AI incidents starts with establishing a shared understanding of current reporting systems

Four key resources informed the development of this common reporting framework for AI incidents: the OECD Framework for AI system classification, the AI Incidents Database (AIID), the OECD Global Portal on Product Recalls, and the AI Incidents Monitor (AIM). From these four resources, a total of 88 criteria were identified to evaluate incidents and, in the case of product recalls, faulty products.

The common reporting framework is designed to be concise and comprehensive

Based on these four resources, 29 criteria for a common reporting framework were identified. These criteria, also called *recurrent criteria*, were included if they appeared in at least three of the four analysed frameworks or provided essential details not covered by recurrent criteria, referred to as *complementary criteria*.

Complementing existing policies, the framework informs policymakers on materialised AI risks through adaptable and interoperable AI incident reporting

The framework, partly via its seven mandatory criteria, provides a flexible structure for reporting and monitoring AI incidents. Its implementation will enhance the interoperability of AI incident reporting while complementing domestic policies and regulatory measures. Reporting AI incidents will assist policymakers in identifying high-risk systems across different contexts, understanding current and future risks, and assessing their impact on affected stakeholders. The framework will also facilitate sharing knowledge and information regarding AI incidents among jurisdictions without prejudice to privacy, intellectual property, or security laws.

Allowing open submissions to the AI Incidents Monitor (AIM) will enable the common reporting framework's testing and evaluation

AIM, accessible at *oecd.ai/incidents*, is an important platform for collecting and analysing AI incidents and hazards. By enabling open submissions, AIM will provide a real-world environment for testing and evaluating the common reporting framework and supplying more data about incidents. Thus, it will ultimately contribute to developing and using safe, secure and trustworthy AI.

1 Introduction

With the increasing uptake of AI systems, the frequency of reported incidents and hazards also rises. These "AI incidents" and "AI hazards" may present significant risks and necessitate structured government oversight (OECD, 2023[1]).

The informal OECD.AI expert group on AI incidents, set up in January 2023, has two main work streams: one is a conceptual workstream dedicated to the classification of AI incidents and hazards and the creation of a unified reporting framework; the other is an applied workstream that puts these conceptual definitions and structures into practice to track real-world incidents and hazards using the AI Incidents Monitor (AIM) (OECD, 2023[1]).

The first stream developed definitions for AI incidents, hazards and related terminology (OECD, 2023[1]), which categorise AI harm and facilitate both voluntary and mandatory reporting. These definitions form the basis for a common reporting framework, aiming for a uniform, cross-country incident reporting system.

Box 1.1. Definitions of AI incident and AI hazard

An **AI incident** is an event, circumstance or series of events where the development, use or malfunction of one or more AI systems directly or indirectly leads to any of the following harms:

(a) injury or harm to the health of a person or groups of people;
(b) disruption of the management and operation of critical infrastructure;
(c) violations of human rights or a breach of obligations under the applicable law intended to protect fundamental, labour and intellectual property rights;
(d) harm to property, communities or the environment.

An **AI hazard** is an event, circumstance or series of events where the development, use or malfunction of one or more AI systems could plausibly lead to an AI incident, i.e., any of the following harms:

(a) injury or harm to the health of a person or groups of people;
(b) disruption of the management and operation of critical infrastructure;
(c) violations to human rights or a breach of obligations under applicable law intended to protect fundamental labour and intellectual property rights;
(d) harm to property, communities or the environment.

Source: OECD (2024[2])

The second workstream aims to improve our understanding of AI risks and provide insights on relevant trends and developments. Currently, AIM identifies AI-related incidents and hazards from reputable media outlets, based on the Alexa traffic rank, in real-time (OECD, 2024[3]). In the future, stakeholders will be allowed to submit new incident reports or complement existing ones through an open submission process. This will facilitate the testing and evaluation of the common reporting framework in practice, and will make the data more reflective of real-world patterns.

This report focuses on the reporting of AI incidents and hazards. It does not provide specific guidelines for AI developers, users, operators or policymakers on how to take preventive or corrective actions after an incident or hazard occured. The framework outlined in this report is designed to complement, not replace, existing national incident reporting frameworks. The aim of this framework is to enhance international alignment in incident reporting, while fully respecting and complementing individual countries' legal requirements. Additional reporting guidelines may still be helpful for particular regions or contexts (e.g., to incorporate information about causes, impacts, mitigations, or other elements that policymakers may wish to solicit from specific actors with privileged access).

2 Developing a common reporting framework for AI incidents

Tracking AI incidents and hazards globally calls for a consistent and interoperable reporting framework across jurisdictions. Such a reporting framework should be concise yet flexible and comprehensive. It should allow anyone to report incidents, while ensuring that incident reports meet certain quality standards. The framework is intended to be used by governments, national authorities and other stakeholders to facilitate interoperable reporting of AI incidents. It enables countries to adopt a common approach to reporting while allowing flexibility in how they respond.

The common reporting framework for AI incidents discussed in this report is designed to support the international monitoring of AI incidents and hazards. It aims to provide the basis for both mandatory and voluntary incident reporting across jurisdictions (OECD, 2023[1]). It is envisioned that the common reporting framework will be incorporated into the AI Incidents Monitor, establishing a system for reporting and monitoring AI incidents in practice. This will help gather evidence to inform AI governance and prevent future incidents. International adoption of the framework could serve as the foundation for global AI incident reporting.

The methodology for creating this common reporting framework, which includes 29 criteria, is outlined in the following subsections.

Methodology

Stocktaking of existing frameworks

Four existing resources and frameworks informed the development of the common reporting framework: the OECD Framework for the Classification of AI systems (OECD, 2022[4]), the Responsible AI Collaborative AI Incidents Database (AIID, 2024[5]), the OECD Global Portal on Product Recalls (OECD, 2024[6]) and the OECD AI Incidents Monitor (AIM) (OECD, 2024[3]). Together, these resources offer a comprehensive understanding of various relevant criteria to characterise AI systems, AI incidents and faulty products more generally (Box 2.1).

Box 2.1. Brief description and scope of relevant frameworks for AI incident reporting

- **OECD Framework for the Classification of AI systems:** User-friendly framework that allows policymakers to classify different types of applied AI systems. It helps distinguish AI applications according to their potential impact on individuals, society and the planet.
- **Responsible AI Collaborative's AI Incidents Database**: Repository indexing harms or near harms realised by the deployment of artificial intelligence systems. AI incidents are human-reviewed for inclusion in the database and are subsequently annotated in more detail using different taxonomies.
- **OECD Global Portal on Product Recalls**: Collects mandatory and voluntary information on product recalls notified and issued by governmental authorities.
- **OECD AI Incidents Monitor**: Documents AI incidents to help policymakers, AI practitioners, and all stakeholders worldwide gain insights into the incidents and hazards that concretise AI risks. Over time, and with a broader reporting process, AIM will help to show risk patterns and establish a collective understanding of AI incidents and their multifaceted nature.

Note: For more information on each of these frameworks, please see Annex A.
Source: OECD (2022[4]; 2024[6]; 2024[3]) and AIID (2024[5]).

In total, 88 criteria to characterise an AI system, incident or product were identified from these four frameworks. These criteria were then grouped into eight dimensions, five of which were based on the OECD Framework for the Classification of AI systems (namely, criteria related to people and planet; economic context; data and input; AI model; and task and output). The three remaining dimensions correspond to incident metadata, harm details and complementary information about the incident (Table 2.1).

Table 2.1 A total of 88 potential criteria for a common AI incident reporting framework were grouped into 8 dimensions

Dimensions (8)	Description	Number of criteria (88)
Incident metadata	Metadata such as date of occurrence, title and description for each incident.	14
Harm details	Exploration of the harm, focusing on its severity, type and impact.	17
People and planet	Includes impacted stakeholders and associated AI principles.	10
Economic context	Study of the economic and environment sectors where the AI system was deployed.	11
Data and input	Description of the data and inputs selected to train the AI-system.	10
AI model	Information related to the model type, including its capacity to evolve before or after deployment and the associated usage rights.	15
Task and output	Description of the AI system tasks, action autonomy level, and outputs.	5
Other information about this incident	Set of actions and complementary information reported by actors with respect to an incident.	6

Source: OECD-compiled database of possible criteria for a common AI incident reporting framework.

Criteria selection

A two-step process was used to determine the criteria for inclusion in the common reporting framework. Criteria meeting either of the following two conditions were selected:

- **Recurrent criteria**: These are 10 criteria that appear in at least three of the four frameworks analysed. Being the minimum common denominator between most frameworks, these criteria are deemed relevant to incident reporting. Examples of recurrent criteria include affected stakeholders, sector of deployment and country in which the incident occurred.
- **Complementary criteria**: These are criteria providing relevant and complementary information to characterise AI incidents not addressed in the recurrent criteria. There are a total of 19 complementary criteria: they provide essential information to ensure that the common reporting framework captures important details about AI incidents, potentially including technical information on data and input, the AI model, and the tasks and outputs of the related AI system. Other examples of complementary criteria include details about the individual or organisation submitting the incident and, where applicable, the quantification of harm.

This process led to the selection of a total of 29 criteria as the basis for a common AI incident reporting framework.

Common reporting framework

The resulting common reporting framework includes the following features:

- **Optionality**: Drawing inspiration from the frameworks analysed, only a subset of the 29 criteria within the common reporting framework will be mandatory. These seven mandatory criteria will include fundamental information necessary to understand the incident, its impacts, and its links with the AI system. Making only some criteria mandatory streamlines the reporting process. Meanwhile, optional criteria facilitate the inclusion of supplementary information where available. Mandatory criteria are denoted by an asterisk in Table 2.2.
- **Answer format:** Inputs to the common reporting framework vary in format, encompassing binary input (e.g., yes/no), multi-selection (e.g., allowing the reporting entity to select one or multiple options), and open text. Binary input and multi-selection criteria promote consistency in reporting and comparability by offering predefined responses.
- **Dimensions:** Consistent with the categorisation presented in Table 2.1, the 29 criteria chosen for the common reporting framework are organised into 8 dimensions, primarily aligned with the OECD Framework for the Classification of AI systems.

This framework describes the data required for each incident report. While the AIM seeks to provide an interface for reporting incidents and hazards in line with the framework's format, the framework itself primarily focuses on defining the data format rather than the interface. Consequently, alternative reporting interfaces for AI incidents may wish to make further adjustments – such as making additional criteria mandatory – to better align with specific reporting contexts.

Box 2.2. Reporting AI incidents and hazards through eight dimensions

- **Metadata dimension (9 criteria):** Includes the incident's title, description, and supporting material.
- **Harm details dimension (4 criteria):** Describes the severity of the incident and the type of harm caused.
- **People and planet dimension (3 criteria):** Covers affected stakeholders, associated AI principles, and violations of human rights.
- **Economic context dimension (4 criteria):** Encompasses factors such as industry, business function, and impact on critical infrastructure.
- **Data and input dimension (1 criterion):** Relates to the AI system's training data.
- **AI model dimension (3 criteria):** Indicates whether the incident is linked to the AI model or the interaction of multiple models.
- **Task and output dimension (2 criteria):** Provides information on the task and autonomy level of the AI system.
- **Other information dimension (3 criteria):** Allows submitters to provide additional incident details. Submitters affiliated to the organisation that developed or deployed the AI system can describe actions taken to cease, prevent or mitigate risks.

The 29 criteria presented in Table 2.2 summarise the information needed to understand an AI incident, at the same time allowing for additional details to provide more nuanced insights to policymakers and regulators. A more detailed table is available in Annex B (Table B.1).

Table 2.2. Criteria for the common reporting framework

Incidents reporting framework criteria	Sub-criteria
1. *Title**	N/A
2. *Description of the incident**	N/A
3. How is the AI system(s) related to the incident*	Direct cause; contributing factor; failure to act; overreliance and intentional misuse; human error; failure to comply with legal frameworks; other (specify for all) (Annex C)
4. Submitter information (role, affiliation, etc.)*	Role; email; affiliation; stakeholder group or source type; relation to the incident: "I represent a government or regulatory body", "I work or am affiliated to a public interest body or NGO", "I work in or am affiliated to the organisation that developed or provided the related AI system", "I am a user of the related AI system", "I am an affected stakeholder", "None of the above, but have partial or substantial knowledge of the incident (e.g. first-hand knowledge, research etc.)", "Other (specify)"
5. *Date of first known occurrence*	N/A
6. *Country(ies) where incident occurred*	List of countries
7. *Supporting material(s) about the incident**	N/A
8. Name and version of the AI system(s)/product(s)	N/A
9. Organisation(s) that developed and/or deployed the AI system	N/A
10. *Severity**	Hazard; serious hazard; incident; serious incident; disaster; other (specify) (OECD[1])
11. *Harm type**	Physical; psychological; reputational; economic/property; environmental; public interest/critical infrastructure; human or fundamental rights; other (specify) (OECD[1])
12. If applicable, quantification of harm	Economic losses; death; injury; number of affected stakeholders; compensation; other (specify)
13. Incident linked to use of AI system(s) in unintended/wrongful way (and how)	If selected, please specify (short answer, limited characters)
14. *Affected stakeholder(s)*	Consumer; children; workers; trade unions; business; government; civil society; general public; other (specify) (OECD[4])
15. Adverse impacts on human rights or fundamental rights	If selected, please specify (short answer, limited characters)
16. *Associated AI Principles*	Accountability; fairness; inclusive growth; privacy; data governance; respect of human rights; robustness; digital security; safety; environmental sustainability; transparency; explainability; democracy; human autonomy (OECD[7])
17. *Industry(ies)*	Classification from the International Standard Industrial Classification of All Economic Activities (ISIC) (ILOSTAT[8])
18. Business function(s) where the AI incident occurred	Human resource management; sales; ICT management and information security; marketing and advertisement; logistics; citizen/customer service; procurement; maintenance; accounting; monitoring and quality control; production; planning and budgeting; research and development; compliance and justice; other (specify) (OECD[4])

19. Incident linked to the functioning of critical functions/infrastructure	Energy, including oil and gas; water supply and wastewater management; healthcare and public health; transportation and logistics; telecommunications and ICT infrastructure; food and agriculture; financial services; public safety and emergency services; government operations and public administration, including electoral systems; manufacturing and industry; education and research; housing and urban infrastructure; public utilities and environmental protection; supply chain and distribution networks; national defense and border security; other (specify). (CISA[9]; EU[10])
20. Breadth of deployment	Pilot project (e.g. team/small group); narrow deployment (e.g. company/city); broad deployment (e.g. sector/country); widespread deployment (e.g. sectors/countries); other (specify) (OECD[4])
21. Incident linked to the training data of AI system(s) (and how)	If selected, please specify (short answer, limited characters)
22. Incident linked to the AI model (and how)	If selected, please specify (short answer, limited characters)
23. Usage rights	One-time license; fee-based; research purposes only; non-commercial; restricted access; free of charge; creative commons; open source/permissive; copyleft/share alike; other (specify) (OECD[11])
24. Incident linked to interaction of multiple AI systems	If selected, please specify (short answer, limited characters)
25. Task(s) of AI system(s)	Recognition/object detection; organisation/recommenders; event/anomaly detection; forecasting/prediction; interaction support/chatbots; goal-driven organisation; reasoning with knowledge structures/planning; content generation; other (specify) (OECD[4])
26. Maximum autonomy level of AI system(s)	No-action autonomy (human support); low-action autonomy (human-in-the-loop); medium-action autonomy (human-on-the-loop); high-action autonomy (human-out-of-the-loop); other (specify) (OECD[4])
27. If applicable, action(s) taken	Prevention; mitigation; ceasing; remediation; other (specify for all) (OECD)
28. If applicable, steps to reproduce the incident	If selected, please specify (open text)
29. Additional information	N/A

Note: Criteria in italics are included in at least three frameworks. The asterisks denote mandatory criteria. A core list of critical functions and infrastructure, commonly included across jurisdictions, is provided to enhance usability.
Source: OECD

3 Conclusion and next steps

The proposed common reporting framework aims to facilitate alignment in international AI incident reporting while allowing national authorities to monitor incidents according to their domestic policies and legal frameworks. This flexibility enables variations in reporting, and studying these differences will help policymakers understand perceptions of incidents in different contexts.

National authorities monitoring AI incidents are encouraged to test the common reporting framework in practice. The evidence gathered will facilitate in-depth analyses of AI incidents and their underlying risks, enabling deeper investigations of serious incidents. This will help policymakers to identify high-risk AI systems, assess their impacts and understand current and future risks. The implementation of this framework would also provide policymakers with valuable insights into preventative and mitigation measures, especially for common incidents, helping to inform future policy recommendations.

Forging partnerships across international organisations and jurisdictions is essential to expand the common reporting framework's reach and effectiveness. Collaborating with various organisations and experts in incident reporting, including standard-setting organisations, will promote knowledge sharing and good practices for managing AI incidents. Understanding the alignment of the framework with other reporting mechanisms – such as the reporting framework for the G7 code of conduct on advanced AI development –would further encourage uptake and interoperability in reporting.

Moving forward, it is essential for the AI Incidents Monitor (AIM) to align closely with the common reporting framework. This alignment can be achieved by integrating open submissions into AIM in accordance with the framework and by ensuring that AI incidents and hazards from the media are tagged using the criteria defined within the framework.

Annex A. Analysis of existing frameworks to inform AI incident reporting

OECD Framework for the Classification of AI systems

The OECD Framework for the Classification of AI systems, published in 2022, is a user-friendly framework that allows policymakers to classify different types of applied AI systems. It helps distinguish AI applications according to their potential impact on individuals, society and the planet (OECD, 2022[4]).

The framework links the technical characteristics of AI systems with the policy implications set out in the OECD AI Principles. It classifies AI systems along five dimensions and a total of 37 criteria (Table **A**.1).

Table A.1. Dimensions of the OECD Framework for the Classification of AI systems

Dimension	Description	Number of criteria
People and Planet	List of criteria applicable to promote human-centric and trustworthy AI for the benefit of people and the planet. Includes impacted stakeholders, users and environmental impacts.	6
Economic Context	Study of the context where the AI system was deployed. Highlights the need for sector-specific policies and includes criteria such as the industry and breadth of deployment.	6
Data & Input	Description of the data and inputs selected to train the AI system. It includes the provenance of the data, collection methods and data properties, necessary to ensure privacy, inclusiveness, and fairness.	9
AI Model	Description of model characteristics, as well as model building and inferencing methods.	11
Task & Output	Includes the tasks of the system, its action autonomy, evaluation methods and core application areas.	5

Source: Adapted from OECD (2022[4]).

The framework allows users to zoom in on specific risks that are typical of AI, such as bias, explainability and robustness, yet it is generic in nature. It facilitates nuanced and precise policy debate. The framework can also help develop policies and regulations, since AI system characteristics influence the technical and procedural measures they need for implementation (OECD, 2022[4]).

AI Incidents Database (AIID)

The AIID, a project of the Responsible AI Collaborative, is a database of AI harms and near harms (AIID, 2024[5]). The AIID uses two taxonomies to classify AI incidents: the Center for Security and Emerging Technology (CSETv1) AI Harm Taxonomy for AIID; and the Goals, Methods and Failures (GMF) taxonomy (Table **A**.2).

Table A.2. Description of the taxonomies used by the AI Incidents Database (AIID)

Taxonomy	Description	Number of criteria
CSETv1	Taxonomy of harm characteristics linked to AI incidents. Presents a structure for extracting AI harm information, which can be used to track trends, prevent incidents, and identify the various types of AI harms. Includes details on the AI system, sector, environment, entities, locations, dates and types of harms.	70
GMF	Taxonomy built on three factors: AI system goals; AI methods and technologies; and AI failure causes. The factors study the taxonomic relationship, records of incidents and technological knowledge to create its GMF annotation.	18

Source: Hoffmann et al. (2023[12]), Pittaras and McGregor (2022[13]).

The AIID's submission and vetting process can provide valuable lessons to the proposed open reporting system of the AI Incidents Monitor (AIM). The OECD and the AIID collaborate on AI incident monitoring and reporting, with a focus on identifying synergies and complementarities between the two platforms.

Global Recalls Portal

The OECD Global Portal on Product Recalls, developed by the OECD Working Party on Consumer Product Safety promotes information sharing and co-operation for product safety amongst multiple players. This is achieved thanks to the identification of safety issues early on, sharing of information and practices and addressing safety concerns in a consistent way, all while supporting international dialogue (OECD, 2024[6]). Similar goals are expected to be drawn from the monitoring of AI incidents.

The portal contains mandatory and voluntary information of product recalls which have been made publicly available and have been notified and issued by governmental authorities. Accessible by consumers and businesses, the portal contains product recall information from 47 jurisdictions. Each product recall has its own page of details, where 16 criteria, as described in Table A.3, are presented to describe one recall.

Table A.3. Global Recalls Portal table of criteria

Dimension	Description	Number of criteria
Recall detail	Information on the overall alert, details on the date of the alert and economies involved.	6
Product details	Description of the hazard, possible injuries and action chosen to respond to the alert.	8
Categorisation	Segment detail.	1
Tags	Tag for description of the recall.	1

Source: OECD (2024[6]).

AI Incidents Monitor (AIM)

Currently, AIM tracks AI incidents from reputable media globally and in real time to help policymakers, AI practitioners, and all stakeholders worldwide gain valuable insights into the incidents and hazards that concretise AI risks. Over time, and with the possible addition of an open reporting system based on this common reporting framework, AIM will help to show risk patterns and establish a collective understanding of AI incidents and their multifaceted nature and serve as an important tool for trustworthy AI.

AIM contains incidents characteristics that mirror the OECD's definition of an AI incident and related terminology OECD (2024[2]). AIM contains 27 criteria including harm type, severity, affected stakeholders, country, industry and other incident metadata (Table A.4).

Table A.4. Summary table of information present in AIM

Dimension	Description
Harm type	Includes psychological, physical, environmental, etc.
Severity	Mostly related to physical harm, includes hazard, injury and death
Affected stakeholders	Ranging from consumers to businesses and the general public
AI Principles	AI principle most closely related to the incident
Industry	20+ industries
Future threats	Hazards that could materialise into incidents
Tags	Key topics related to the incident
Incident metadata	ID number, date of occurrence, country, link to news article
Description	Incident summary, "why is this an AI incident" section

Source: OECD (2024[3]).

Annex B. Detailed criteria of the common reporting framework

Table B.1. Criteria for the common reporting framework

Dimension	Incidents reporting framework criteria	Answer format	Sub-criteria
Incident metadata	1. *Title**	Open text	N/A
Incident metadata	2. *Description of the incident**	Open text	N/A
Incident metadata	3. How is the AI system(s) related to the incident*	Multi-selection with open text	Direct cause; contributing factor; failure to act; overreliance and intentional misuse; human error; failure to comply with legal frameworks; other (specify for all) (Annex C)
Incident metadata	4. Submitter information (role, affiliation, etc.)*	Open text and multi-selection	Role; email; affiliation; stakeholder group or source type; relation to the incident: “I represent a government or regulatory body”, "I work or am affiliated to a public interest body or NGO", “I work in or am affiliated to the organisation that developed or provided the related AI system”, “I am a user of the related AI system”, “I am an affected stakeholder”, “None of the above, but have partial or substantial knowledge of the incident (e.g. first-hand knowledge, research etc.)”, “Other (specify)”
Incident metadata	5. *Date of first known occurrence*	Date	N/A
Incident metadata	6. *Country(ies) where incident occurred*	Multi-selection	List of countries
Incident metadata	7. *Supporting material(s) about the incident**	Open text, URLs and upload button	N/A
Incident metadata	8. Name and version of the AI system(s)/product(s)	Open text	N/A
Incident metadata	9. Organisation(s) that developed and/or deployed the AI system	Open text	N/A
Harm details	10. *Severity**	Multi-selection	Hazard; serious hazard; incident; serious incident; disaster; other (specify) (OECD[1])
Harm details	11. *Harm type**	Multi-selection	Physical; psychological; reputational; economic/property; environmental; public interest/critical infrastructure; human or fundamental rights; other (specify) (OECD[1])
Harm details	12. If applicable, quantification of harm	Multi-selection	Economic losses; death; injury; number of affected stakeholders; compensation; other (specify)

Harm details	13. Incident linked to use of AI system(s) in unintended/wrongful way (and how)	Checkbox	If selected, please specify (short answer, limited characters)
People & planet	14. *Affected stakeholder(s)*	Multi-selection	Consumer; children; workers; trade unions; business; government; civil society; general public; other (specify) (OECD[4])
People & planet	15. Adverse impacts on human rights or fundamental rights	Checkbox	If selected, please specify (short answer, limited characters)
People & planet	16. *Associated AI Principles*	Multi-selection	Accountability; fairness; inclusive growth; privacy; data governance; respect of human rights; robustness; digital security; safety; environmental sustainability; transparency; explainability; democracy; human autonomy (OECD[7])
Economic context	17. *Industry(ies)*	Multi-selection	Classification from the International Standard Industrial Classification of All Economic Activities (ISIC) (ILOSTAT[8])
Economic context	18. Business function(s) where the AI incident occurred	Multi-selection	Human resource management; sales; ICT management and information security; marketing and advertisement; logistics; citizen/customer service; procurement; maintenance; accounting; monitoring and quality control; production; planning and budgeting; research and development; compliance and justice; other (specify) (OECD[4])
Economic context	19. Incident linked to the functioning of critical functions/infrastructure	Checkbox	Energy, including oil and gas; water supply and wastewater management; healthcare and public health; transportation and logistics; telecommunications and ICT infrastructure; food and agriculture; financial services; public safety and emergency services; government operations and public administration, including electoral systems; manufacturing and industry; education and research; housing and urban infrastructure; public utilities and environmental protection; supply chain and distribution networks; national defense and border security; other (specify). (CISA[9]; EU[10])
Economic context	20. Breadth of deployment	Single choice	Pilot project (e.g. team/small group); narrow deployment (e.g. company/city); broad deployment (e.g. sector/country); widespread deployment (e.g. sectors/countries); other (specify) (OECD[4])
Data & input	21. Incident linked to the training data of AI system(s) (and how)	Checkbox	If selected, please specify (short answer, limited characters)
AI model	22. Incident linked to the AI model (and how)	Checkbox	If selected, please specify (short answer, limited characters)
AI model	23. Usage rights	Multi-selection	One-time license; fee-based; research purposes only; non-commercial; restricted access; free of charge; creative commons; open source/permissive; copyleft/share alike; other (specify) (OECD[11])
AI model	24. Incident linked to interaction of multiple AI systems	Checkbox	If selected, please specify (short answer, limited characters)
Task & output	25. Task(s) of AI system(s)	Multi-selection	Recognition/object detection; organisation/recommenders; event/anomaly detection; forecasting/prediction; interaction support/chatbots; goal-driven organisation; reasoning with knowledge structures/planning; content generation; other (specify) (OECD[4])
Task & output	26. Maximum autonomy level of AI system(s)	Single choice	No-action autonomy (human support); low-action autonomy (human-in-the-loop); medium-action autonomy (human-on-the-loop); high-action autonomy (human-out-of-the-loop); other (specify) (OECD[4])
Other	27. If applicable, action(s) taken	Open text, multi-selection	Prevention; mitigation; ceasing; remediation; other (specify for all) (OECD)
Other	28. If applicable, steps to reproduce the incident	Open text	If selected, please specify (open text)
Other	29. Additional information	Open text	N/A

Note: Criteria in italics are included in at least three frameworks. The asterisks denote mandatory criteria.
Source: OECD.

Annex C. Taxonomy of possible links between an AI system and an incident

Table C.1 proposes a categorisation of the different relationships an AI system can have with a given incident. Descriptions and examples are provided for each category.

The below categories do not intend to provide an exhaustive list of the different relationships between an AI system and an incident. These links are not mutually exclusive and multiple ones may occur per incident. Submitters are invited to select all applicable links and specify any additional ones not included in these 6 categories.

Table C.1. Taxonomy of possible links between an AI system and an incident

Type of involvement	Category	Description	Example
Direct	**Direct cause**	The AI system is the primary reason for the incident due to a malfunction or erroneous output.	A self-driving car causes a collision due to a misinterpretation of road signals.
Direct	**Contributing factor**	The AI system played a supportive or secondary role in causing the incident.	An AI-based traffic management system incorrectly optimises traffic flow, contributing to congestion during an emergency response.
Direct	**Failure to act**	The AI system did not detect or respond to an issue that it was expected to handle.	An AI-powered fraud detection system fails to flag suspicious transactions, leading to financial loss.
Indirect	**Overreliance and intentional misuse**	*Overreliance*: The incident occurs because the user intentionally misuses the AI or overly depends on it, disregarding proper oversight.	An AI-assisted medical diagnosis tool suggests a wrong treatment, which is accepted by the physician.
		Intentional misuse, including malicious use: The AI system may function as intended, but its malicious use causes an incident.	An AI facial recognition system developed for security purposes is used to surveil individuals without their consent.
Indirect	**Human error**	*Developer error*: The AI itself may function as intended, but human errors in its development lead to unintended outcomes.	A data scientist trains an AI model on flawed data, leading to incorrect predictions.
		Operator error: The AI system may function as intended, but unintended outcomes arise from the operator's lack of skills, incorrect system configuration, inadequate monitoring, or inappropriate application.	An operator misinterprets the AI's recommendations, causing inappropriate actions to be taken.
		User error: Mistakes made by users when interacting with an AI system, often due to misunderstanding of outputs, improper inputs, or lack of training.	An autonomous vehicle switches lanes because it has detected a collision, but the driver, unaware of the collision, prevents the vehicle from changing lanes.
Indirect	**Failure to comply with legal frameworks**	The AI system functions as intended but fails to comply with existing legal frameworks.	An AI system does not comply with current data protection laws and policies, thereby violating user privacy rights.

Source: OECD.

References

AIID (2024), *AI Incident Database (AIID)*, https://incidentdatabase.ai/ (accessed on 22 January 2024). [5]

CISA (2024), *National Critical Functions Set*, https://www.cisa.gov/national-critical-functions-set (accessed on 30 October 2024). [9]

EU (2022), *Directive (EU) 2022/2557 of the European Parliament and of the Council of 14 December 2022 on the resilience of critical entities and repealing Council Directive 2008/114/EC (Text with EEA relevance)*, http://data.europa.eu/eli/dir/2022/2557/oj. [10]

Hoffmann, M. et al. (2023), *CSET AI Harm Taxonomy for AIID and Annotation Guide*, https://github.com/georgetown-cset/CSET-AIID-harm-taxonomy/blob/main/CSET%20V1%20AI%20Annotation%20Guide%20(with%20Schema%20and%20Field%20Descriptions)%2025Jul2023.pdf. [12]

ILOSTAT (2024), *International Standard Industrial Classification of All Economic Activities (ISIC)*, https://ilostat.ilo.org/resources/concepts-and-definitions/classification-economic-activities/ (accessed on 12 March 2024). [8]

OECD (2024), *AI Incidents Monitor (AIM)*, https://oecd.ai/incidents (accessed on 18 January 2024). [3]

OECD (2024), "Defining AI incidents and related terms"*, OECD Artificial Intelligence Papers*, No. 16, OECD Publishing, Paris, https://doi.org/10.1787/d1a8d965-en. [2]

OECD (2024), *Global portal on product recalls*, https://globalrecalls.oecd.org/#/ (accessed on 22 January 2024). [6]

OECD (2023), "Stocktaking for the development of an AI incident definition"*, OECD Artificial Intelligence Papers*, No. 4, OECD Publishing, Paris, https://doi.org/10.1787/c323ac71-en. [1]

OECD (2022), "OECD Framework for the Classification of AI systems", *OECD Digital Economy Papers*, https://doi.org/10.1787/cb6d9eca-en. [4]

OECD (2021), "Tools for trustworthy AI: A framework to compare implementation tools for trustworthy AI systems", *OECD Digital Economy Papers*, No. 312, OECD Publishing, Paris, https://doi.org/10.1787/008232ec-en. [11]

OECD (2019), *Recommendation of the Council on Artificial Intelligence*, https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449. [7]

Pittaras, N. and S. McGregor (2022), *A taxonomic system for failure cause analysis of open source AI incidents*, https://doi.org/10.48550/arXiv.2211.07280. [13]