Vol. 2026-01
May 2026

동향 AI 안전 동향 분석 리포트 26-01

AI사고 정의 및 보고체계 국제동향

Global Trends in AI Accident Definition and Reporting Frameworks

송경호, 주지연



ETRI 한국전자통신연구원 www.etri.re.kr
인공지능안전연구소 AI Safety Institute

AI 안전 동향 분석 시리즈 26-01(2026년 5월)

「AI사고 정의 및 보고체계 국제동향」

송경호(선임연구원), 주지연(석사후연수연구원)

한국전자통신연구원 인공지능안전연구소 AI안전정책및대외협력실

본 보고서는 주요국 및 국제기구의 AI사고 정의와 보고체계 현황을 바탕으로 분석·정리한 자료입니다. 보고서 내용은 정보 제공 및 연구 목적을 위한 것이며, 원문 자료와 해석 또는 적용상 차이가 있는 경우에는 원문이 우선합니다. 또한 본 보고서의 분석 및 해설에는 연구진의 견해가 포함될 수 있으며, 소속 기관의 공식 입장을 의미하지 않습니다.

AI Safety Trends Analysis Series 26-01(May 2026)

「Global Trends in AI Incident Definitions and Reporting Frameworks」

Kyungho Song(Senior Researcher), Jiyeon Joo(Research associate)

AI Safety Policy and International Collaboration Section, Korea AI Safety Institute(AISI), Electronics and Telecommunications Research Institute(ETRI)

This report provides an analysis of AI incident definitions and reporting frameworks across major jurisdictions and international organizations. It is intended as a reference material for informational and research purposes, and in the event of any discrepancy between this report and the original source materials, the original sources shall prevail.
The analyses and interpretations presented in this report may reflect the views of the authors and do not necessarily represent the official position of their affiliated institutions.

Contents

Ⅰ. AI사고 정의 및 보고체계의 주요 제도화 동향 ····· 1

1. OECD ····· 1

2. 유럽연합(EU) ····· 4

3. 영국(UK) ····· 6

4. 미국(US) ····· 8

Ⅱ. 기타 국가의 AI사고 정의 및 보고체계 동향 ····· 10

5. 캐나다(Canada) ····· 10

6. 일본(Japan) ····· 12

7. 싱가포르(Singapore) ····· 14

8. 호주(Australia) ····· 15

9. 케냐(Kenya) ····· 17

요 약 문

본 보고서는 주요국 및 국제기구의 AI사고 정의 및 보고체계 현황을 비교·분석하고, 국제적 논의 동향과 제도적 특징을 종합적으로 검토한다.

AI사고 정의: 국제적으로 AI사고 개념은 OECD 프레임워크를 중심으로 형성되고 있다. OECD는 실제 피해 발생 여부를 기준으로 AI사고(AI Incident)와 AI위험원(AI Hazard)을 구분하고, 피해 수준에 따라 중대 AI사고(Serious AI Incident)와 AI재난(AI Disaster)으로 세분화한다. 이는 단순한 기술적 오류를 넘어, AI시스템이 인간의 생명·신체·재산·기본권 및 사회 전반에 미치는 영향을 포괄적으로 고려하려는 접근이라는 점에서 의미를 가진다. 또한 OECD는 국가 간 보고 기준의 정합성과 비교 가능성 확보를 위해 88개 후보 기준을 8개 차원으로 재구성한 후, 29개 공통 보고 기준과 7개 핵심 필수 기준으로 구성된 공통 보고 프레임워크를 구축하고 있다. 이러한 논의는 AI사고를 국제적으로 공유·비교 가능한 위험 관리 대상으로 체계화하려는 흐름으로 볼 수 있다.

AI사고 보고체계: 주요국의 AI사고 보고체계는 크게 통합형 모델과 분산형 모델로 구분된다. EU는 AI사고 전용 의무 보고체계를 법제화한 대표적 사례로, 고위험 AI시스템 제공자에게 중대사고 발생 시 일정 기간 내 시장감시당국(MSA)에 보고 의무를 부과하는 통합적 규제 체계를 운영하고 있다. 특히 사망 사고, 기본권 침해, 중요 인프라 교란 등 중대한 피해 유형에 대해서는 신속 보고 및 감독기관 간 정보 공유 체계를 마련하고 있다. 반면 영국·미국·캐나다 등은 기존 산업별 법체계에 기반한 분산형 접근을 유지하면서 AI사고 보고 의무와 중앙 관리 체계 도입을 검토하고 있다. 일본·싱가포르·호주는 가이드라인 중심의 연성규범(soft law) 접근 아래 자율적 위험 관리와 보완적 거버넌스 체계를 운영하고 있으며, 중앙 모니터링 및 정보 공유 체계 구축 가능성을 함께 논의하고 있다. 케냐는 AI 거버넌스 체계 구축 초기 단계로, 데이터보호 및 사이버보안 관련 기존 법률을 중심으로 제한적 대응 체계를 운영하고 있다.

주요국 및 국제기구는 AI사고를 기존 산업별 규제체계만으로 관리하기 어렵다는 공통된 문제의식을 바탕으로, OECD 중심의 공통 정의 및 보고 기준 마련 논의를 확대하고 있다. 또한 각국은 위험 기반 접근(risk-based approach)을 토대로 의무 보고, 중앙 모니터링, 정보 공유 등 다양한 대응체계 정비를 추진하고 있으며, 이는 향후 국제 AI 거버넌스 체계 논의가 AI사고의 정의 및 보고체계 정립을 중심으로 전개될 가능성을 시사한다.

Executive Summary

This report provides a comparative analysis of AI incident definitions and reporting frameworks across major jurisdictions and international organizations, while examining international regulatory discussions and institutional approaches related to AI incident governance.

Defining AI Incidents: Internationally, the OECD framework has emerged as the primary reference point for defining AI incidents. The OECD distinguishes between AI Incidents, where actual harm has occurred, and AI Hazards, where harm has not yet materialized but could plausibly occur. It further classifies incidents by severity into Serious AI Incidents and AI Disasters. This approach is significant in that it extends beyond technical malfunctions to encompass the broader impacts of AI systems on human life, bodily integrity, property, fundamental rights, and society as a whole. To improve consistency and comparability across jurisdictions, the OECD is also developing a common reporting framework consisting of 29 common reporting criteria and 7 core mandatory criteria refined from an initial set of 88 candidate criteria organized across eight dimensions. These developments reflect broader efforts to establish AI incidents as internationally comparable and shareable objects of risk management and governance.

AI Incident Reporting Systems: National AI incident reporting systems can broadly be divided into integrated and decentralized models. The EU represents the most developed integrated model, having legislated a mandatory AI-specific reporting regime that requires providers of high-risk AI systems to report serious incidents to Market Surveillance Authorities (MSAs) within specified timeframes. In particular, the EU framework establishes expedited reporting and information-sharing mechanisms for incidents involving death, violations of fundamental rights, or serious disruptions to critical infrastructure. In contrast, the UK, the US, and Canada continue to rely primarily on decentralized approaches based on existing sector-specific legal frameworks while considering the introduction of mandatory reporting obligations and centralized oversight mechanisms. The United States

manages AI-related risks through sectoral reporting systems operated by regulatory agencies in areas such as healthcare, transportation, data protection, and consumer protection. Similarly, the UK addresses AI-related incidents through existing reporting regimes in sectors including healthcare, telecommunications, finance, and data protection. Canada is currently considering the introduction of an AI regulatory and reporting framework centered on the concept of “harm” through the proposed Artificial Intelligence and Data Act(AIDA). Japan, Singapore, and Australia maintain soft law approaches centered on guidelines, voluntary risk management, and supplementary governance mechanisms, while also exploring centralized monitoring and information-sharing systems. Kenya remains at an early stage of developing its AI governance framework and currently relies on existing data protection and cybersecurity laws as limited regulatory mechanisms.

Overall, major jurisdictions and international organizations share the view that AI incidents are difficult to manage solely through existing sector-specific regulatory systems. In response, discussions led by the OECD on common definitions and reporting standards for AI incidents continue to expand internationally. At the same time, jurisdictions are increasingly exploring policy measures such as mandatory reporting obligations, centralized monitoring systems, and information-sharing mechanisms based on risk-based approaches. These developments suggest that future international AI governance discussions are likely to place greater emphasis on establishing common frameworks for defining and reporting AI incidents.

Ⅰ AI사고 정의 및 보고체계의 주요 제도화 동향

1 OECD

1-1. AI사고 정의

■ **OECD는 AI시스템으로 인해 발생하는 피해를 체계적으로 정의하고, AI사고(incident)와 AI위험원(hazard)을 구분하는 개념적 틀을 제시함[1)]**

- AI시스템의 개발·사용·오작동 과정에서 발생한 결과를 기준으로, 피해 발생 여부에 따라 실제 AI사고(incident)와 잠재적 AI위험원(hazard)을 구분함. 이는 피해 발생 여부에 따른 대응 및 관리 방식의 차이를 반영하기 위한 체계임
- 이 체계는 국제적 상호운용성을 확보하는 동시에 각국이 정책 및 법체계에 따라 적용 범위와 판단 기준을 자율적으로 설정할 수 있도록 설계됨

■ **(AI사고, AI Incident) 하나 이상의 AI시스템의 개발·사용·오작동이 직접 또는 간접적으로 피해를 초래하는 사고, 상황 또는 일련의 사건**

- 개인 또는 집단의 건강에 대한 피해를 포함하며, 신체적·심리적·정신적 피해와 핵심 인프라의 관리 및 운영에 대한 교란 등 사회 기능에 대한 영향을 포괄함
- 인권 침해 또는 기본권, 노동권 및 지적재산권 관련 법적 의무 위반과 함께 허위정보, 혐오표현 등 무형적 피해와 재산·공동체·환경에 대한 직·간접적 피해도 포함함
- AI시스템 간 상호작용, 오용, 비의도적 사용 및 개발 단계 등 다양한 경로에서 발생할 수 있는 상황을 포괄하도록 정의됨

■ **(AI위험원, AI Hazard) 하나 이상의 AI시스템의 개발·사용·오작동이 AI사고로 이어질 수 있는 개연성이 있는 상황 또는 상태**

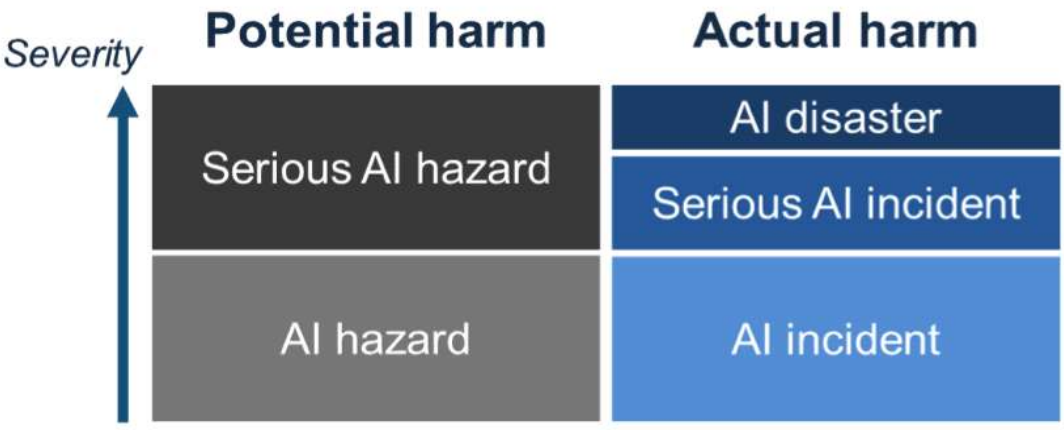
- 피해 발생 가능성과 결과의 심각성을 기준으로 평가되는 잠재적 위험 개념으로, AI 전 수명주기에서 발생할 수 있음

1) OECD, OECD Defining AI Incidents and Related Terms 2024, May 2024

- 실제 피해는 발생하지 않았으나 사고로 이어질 수 있었던 '아차사고(near miss)'를 포함

■ AI사고 및 AI위험원의 분류

| 그림 1. 피해 심각도(severity)에 기반한 AI사고 및 AI위험원의 분류 |



출처 OECD, OECD Defining AI Incidents and Related Terms 2024, May 2024

- AI사고와 AI위험원은 피해의 발생 여부와 피해의 심각도(severity)에 따라 구분되며, 각 범주는 일반 수준과 중대한 수준으로 세분화됨
- 중대 AI사고(Serious AI Incident): 사망 또는 중대한 건강상 피해, 핵심 인프라의 중대하고 비가역적인 교란, 권리(인권, 기본권 등)의 중대한 침해, 재산·공동체·환경에 대한 중대한 피해를 초래하는 사고로서, 실제 피해가 발생한 중대한 수준의 AI사고
- AI재난(AI Disaster): 중대 AI사고 중, 공동체 또는 사회 전반의 핵심 기능에 중대한 영향을 미치고, 자체 대응 능력의 한계를 넘는 수준의 사고로 영향이 광범위하거나 장기적으로 지속될 수 있는 사고
- 중대 AI위험원(Serious AI Hazard): 실제 피해는 발생하지 않았으나, 중대 AI사고 또는 AI재난으로 이어질 가능성이 높고 그 결과의 심각성이 큰 잠재적 위험 상태

1-2. AI사고 보고체계

■ OECD는 AI사고의 체계적 보고 및 국가 간 비교를 용이하게 하기 위해 공통된 사고 보고 기준을 마련하고, 각국의 정책·법체계에 따라 활용할 수 있는 국제 보고체계를 설계함[2)]

- OECD AI시스템 분류 프레임워크, AI사고 데이터베이스(AIID), 글로벌 리콜 포털, AI사고 모니터(AIM) 등 4개의 기존 프레임워크를 분석하여 AI사고 보고를 위한 잠재적 기준 88개를 도출

- 도출된 88개 기준을 8개 차원으로 재구성한 후, 2단계 절차를 거쳐 29개 기준을 선정
  - 1단계에서는 4개 프레임워크 중 최소 3개 이상에 공통적으로 확인되는 반복 기준 10개를 선정하고, 2단계에서는 사고 분석을 보완하기 위해 관련성이 있고 필수적인 정보 기준 19개를 추가 선정함

- 29개 기준 중 7개는 사고의 내용, 영향 및 AI시스템과의 연관성 파악에 필요한 핵심 정보로서, 보고의 일관성과 비교 가능성을 확보하기 위한 최소 요건으로 설정됨
  - 제목, 사고 설명, AI시스템과 사고 간 연계 방식, 제출자 정보, 증빙 자료, 사고의 심각성, 피해 유형 등을 포함하며, 사고의 기본적 특성과 영향 범위를 파악할 수 있도록 구성됨

| 표 1. AI사고 보고 기준의 구성요소 및 체계 |

주요 구성요소	정의된 기준 (88)	선정된 기준 (29)	주요 내용
사고 메타데이터 (Incident Metadata)	14개	**9개**	각 사고에 대한 발생 일자, 제목 및 설명 등 메타데이터
피해 상세 정보 (Harm details)	17개	**4개**	피해의 심각성, 유형 및 영향에 중점을 둔 분석
인간 및 환경 (People and planet)	10개	**3개**	영향받은 이해관계자 및 관련 AI 원칙을 포함
경제적 맥락 (Economic context)	11개	**4개**	AI시스템이 적용된 산업 및 경제·환경 분야에 대한 분석
데이터 및 입력 (Data and input)	10개	**1개**	AI시스템 학습에 사용된 데이터 및 입력에 대한 설명
AI 모델(AI model)	15개	**3개**	모델 유형, 배포 전·후 진화 가능성 및 관련 사용 권한에 관한 정보
작업 및 출력 (Task and output)	5개	**2개**	AI시스템의 작업, 자율성 수준 및 출력에 대한 설명
기타 정보 (Other information)	6개	**3개**	제출자가 사고에 대한 추가 세부사항을 제공함. AI시스템을 개발하거나 배포한 조직과 관련된 제출자는 위험을 중단, 예방 또는 완화하기 위해 취한 조치를 설명함

출처 OECD, Towards A Common Reporting Framework for AI Incidents, Feb 2025

■ AIM(AI Incidents Monitor) 연계를 통한 AI사고 및 위험의 보고·모니터링 체계 구축 추진

- 현재 AIM은 신뢰 가능한 글로벌 미디어를 기반으로 AI사고 및 위험을 실시간 식별·수집하고 있으며, 향후 개방형 제출 체계(open submission) 도입을 통해 이해관계자의 직접 보고 및 기존 사례 보완이 가능하도록 운영될 예정

2) OECD, Towards A Common Reporting Framework for AI Incidents, Feb 2025

2 유럽연합(EU)

2-1. AI사고 정의

■ **EU AI Act는 중대 사고(serious incident)를 AI시스템의 사고 또는 오작동으로 인해 직·간접적으로 발생하는 사망, 중대한 건강 피해, 중요 인프라의 중대하고 비가역적인 교란, 기본권 침해, 중대한 재산·환경 피해를 포함하는 사건으로 정의[3)]**

- OECD 규정에 충실하면서도, AI사고 및 위험에 대한 포괄적 정의를 두지 않으며, 경미한 사고나 잠재적 위험은 법적으로 다루지 않음
- 사고 개념을 '중대 사고'로 한정한 것은 AI 규제의 과도한 행정 부담을 피하고 위험 기반 접근(risk-based approach)*을 유지하기 위한 규제 설계로 해석됨

  * EU AI Act의 risk-based approach는 AI 기술 자체가 아니라 사용 목적과 맥락에서 발생할 수 있는 위험 수준을 기준으로 규제를 차등 적용하는 방식으로, 위험이 클수록 규제를 강화하고 위험이 낮을수록 규제를 최소화함[4)]

- OECD는 EU AI Act의 '중대 사고' 정의를 차용하면서, 실질적·물질적 피해를 포함하도록 AI사고 개념을 확장함[5)]

2-2. AI사고 보고체계

■ **중대 사고 발생 시, 고위험 AI시스템(High-risk AI system)* 제공자는 사고가 발생한 회원국의 시장감시당국(MSA, Market Surveillance Authority)에 보고할 의무를 부담함[6)]**

  * 고위험 AI시스템은 인간의 건강, 안전, 기본권에 중대한 영향을 미칠 수 있는 AI로, (1) 다른 EU 안전 관련 법령과 연계된 제품에 포함된 경우, 또는 (2) Annex III에 열거된 특정 분야(예: 생체인식, 교육, 고용, 법 집행 등)에 사용되는 경우를 의미[7)]

- 보고는 제공자가 AI시스템과 사고 간 인과관계 또는 합리적 관련성을 확인한 즉시 이루어져야 하며, 늦어도 사고 인지 후 15일 이내 제출되어야 함[8)]

3) Regulation (EU) 2024/1689 (Artificial Intelligence Act), Chapter I, Article 3 — Definitions
4) European Commission, Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act), COM/2021/206 final, 21 April 2021
5) OECD, OECD Defining AI Incidents and Related Terms 2024, May 2024
6) Regulation (EU) 2024/1689 (Artificial Intelligence Act), Chapter I, Article 73(1)
7) Regulation (EU) 2024/1689 (Artificial Intelligence Act), Chapter I, Article 6
8) Regulation (EU) 2024/1689 (Artificial Intelligence Act), Chapter I, Article 73(2)

- 사고가 광범위한 침해에 해당하거나 중요 인프라의 중대하고 비가역적인 교란을 초래한 경우에는 2일 이내의 신속 보고 의무가 적용됨[9)]
- 사망이 발생한 경우, 인과관계가 확인되거나 합리적으로 의심되는 즉시 보고해야 하며, 늦어도 사고 인지 후 10일 이내 보고해야 함[10)]

■ **기본권 침해에 해당하는 중대 사고 발생 시, 시장감시당국(MSA)은 해당 사실을 기본권 보호 관련 국가 공공기관에 통보해야 함[11)]**

- 기본권 보호 공공기관은 고위험 AI시스템의 기본권 보호 의무 준수 여부를 감독하기 위해, 제공자(providers) 및 배포자(deployers)에 대해 관련 규제 문서(documentation, technical documentation 등)의 제출을 요구하고 이에 대한 접근 권한을 행사할 수 있음[12)]
  - 해당 문서는 접근 가능한 언어 및 형식(accessible language and format)으로 제공되어야 하며, 공공기관은 문서 요청 사실을 시장감시당국(MSA)에 통보해야 함
- 기본권 침해 관련 사고에 대해 국가 공공기관에 제출되는 표준화된 보고 양식(template)은 별도로 규정되어 있지 않음

■ **고위험 AI시스템 배포자는 시스템 운영을 모니터링하는 과정에서 제79조(1)에서 정의된 위험이 발생할 가능성이 있다고 합리적으로 판단되는 경우, 지체 없이 제공자, 유통업자(distributor) 및 시장감시당국(MSA)에 통보하고 해당 시스템의 사용을 즉시 중단해야 함[13)]**

- 여기에서 '위험'은 AI시스템이 개인의 건강, 안전 또는 기본권에 위해를 초래할 가능성을 의미[14)]
- 이와 같은 위험 개념은 건강·안전 및 기본권에 대한 위해 가능성을 중심으로 정의된다는 점에서 OECD의 AI 위험 개념과도 정합성을 갖는 것으로 평가됨

■ **한편, EU AI Act가 회원국 공통의 규율 체계로 기능함에 따라, 프랑스 역시 별도의 포괄적 AI 규제나 독자적인 AI사고 정의를 두고 있지 않음[15)]**

9) Regulation (EU) 2024/1689 (Artificial Intelligence Act), Chapter I, Article 73(3)
10) Regulation (EU) 2024/1689 (Artificial Intelligence Act), Chapter I, Article 73(4)
11) Regulation (EU) 2024/1689 (Artificial Intelligence Act), Chapter I, Article 73(7)
12) Regulation (EU) 2024/1689 (Artificial Intelligence Act), Chapter I, Article 77(1)
13) Regulation (EU) 2024/1689 (Artificial Intelligence Act), Chapter I, Article 26(5)
14) Regulation (EU) 2024/1689 (Artificial Intelligence Act), Chapter I, Article 79(1)
15) White&Case, AI Watch: Global regulatory tracker – France, 16 July 2024

3 영국(UK)

3-1. AI사고 정의

■ AI사고의 포괄적 정의 및 의무 보고를 규정하는 단일 법률 또는 통합 프레임워크 부재[16)]

- 영국의 AI시스템 규제는 기존 산업별 법적 체계를 기반으로 한 혁신 지향적 접근을 취함
  - 2023년, 영국 정부는 AI 규제를 위한 포괄적 입법을 마련하는 대신, 보건안전청(HSE), 평등·인권위원회(EHRC), 경쟁시장청(CMA) 등 기존 규제 당국이 소관 분야별로 AI를 규제하는 구조를 도입[17)]

- 이러한 접근은 AI시스템의 개발 및 활용을 규율하기 위한 교차 부문적 원칙(cross-sectoral principles)을 중심으로 설계됨
  - 영국 정부는 규제 당국이 안전·보안·강건성(safety·security·robustness), 투명성·설명가능성(appropriate transparency·explainability), 공정성(fairness), 책임·거버넌스(accountability·governance), 이의제기 및 구제 가능성(contestability·redress) 등 5대 원칙을 고려하여 적용하도록 함
  - 해당 원칙은 비구속적(non-statutory) 지침으로 분야별 특성과 위험 수준에 따라 유연하게 적용되도록 설계됨

3-2. AI사고 보고체계

■ AI사고에 대한 국가 차원의 통합 보고체계는 부재하며, 규제기관별로 분산된 체계로 인해 AI사고를 구조적으로 수집·관리하지 못하는 한계가 있음[18)]

- 산업별 법적 프레임워크에 기반한 규제로 인해 AI사고 보고체계가 파편화된 구조를 보임

- 일부 규제기관은 각 분야에서의 사고 보고체계를 운영하고 있으나, 산업별 규제 목적에 한정됨
  - 의약품·의료기기(MHRA)[19)], 통신(Ofcom)[20)], 개인정보 보호(ICO)[21)], 금융(FCA)[22)] 등에서 사고 보고·통지 의무가 존재하나, 해당 산업에 한정되어 AI사고를 통합적으로 수집·관리하기에는 한계가 있음

16) Regulating Artificial Intelligence: U.S. and International Approaches and Considerations for Congress June 4, 2025
17) UK Government, A pro-innovation approach to AI regulation, 3 Aug 2023
18) The Centre for Long-Term Resilience, AI incident reporting: Addressing a gap in the UK's regulation of AI, June 2024
19) UK, MHRA, Yellow Card: how to report a problem with a medicine or medical device, 2023
20) UK, Ofcom, General Conditions of Entitlement, Condition A3 – Network functioning and resilience, 2022
21) UK, ICO, Personal data breaches – Guidance, 2023
22) UK, FCA, Supervision manual (SUP 15: Notifications to the FCA), 2024

■ **이러한 구조로 인해 기존 규제·보고체계의 적용 범위를 벗어난 AI사고에 대해 규제 공백(regulatory gap)이 발생**

- 영국 정부는 규제공백 보완을 위해 기존의 분산적 규제체계를 유지하면서도, AI위험을 중앙에서 종합적으로 파악·분석할 필요성을 제기[23)]
  - 영국 AI안전연구소(UK AISI)는 안전 평가 고도화를 위해 사고 보고로부터 도출되는 신호* 활용을 제안하였으나, 정부 차원의 의무 보고체계로 제도화되지는 않음[24)]

    * AI사고 보고를 통해 수집되는 실제 환경 기반 데이터와 패턴을 의미하며, UK AISI와 중앙AI위험관리조직(CAIRF, Central AI Risk Function)이 AI위험 평가·대응을 위한 조기 경보 지표로 활용[25)]

■ **싱크탱크 CLTR(Centre for Long-Term Resilience)은 AI 오용·오작동을 기록·관리하는 공식 보고체계의 부재를 지적하고, 공공서비스 영역을 중심으로 사고 기록 체계를 구축하여 중앙 허브로 집계·통합한 데이터베이스를 마련할 것을 권고[26)]**

- 기존 규제기관이 수집한 사고 보고를 중앙으로 집계하여 교차 부문 학습(cross-sectoral learning)을 가능하게 하고, 사고 기록을 통해 위험 평가 및 향후 예방·완화 조치에 필요한 정책적 통찰을 도출하는데 있음

- 정부의 ATRS(Algorithmic Transparency Recording Standard)는 공공서비스 AI 활용의 투명성 확보를 위한 제도이나 현재 기록 범위가 제한적임
  - 이에 AI 도구의 사용 맥락, 위험, 완화 조치까지 포함하도록 기록 범위를 확대할 필요성 제기됨
  - CLTR은 영국 정부가 ATRS를 공공부문 AI사고 보고 프레임워크로 확장할 것을 권고함

■ **영국 정부는 고성능 범용 AI 개발사에 대한 타깃형 규제 도입을 논의하고 있으나, AI사고 보고 의무 혹은 강제적 입법을 즉시 도입할 계획은 없음[27)]**

- 2025년에는 내부 보안 사고 또는 위협 발생 시, 식별 및 내부 보고·대응을 하도록 규정하였으나 법적 구속력이 없는 가이드라인에 해당하여 실효성에는 한계가 있음[28)]

- 사이버보안 법안에서는 기업이 중대한 사이버 침해를 일정 기간 내 보고하도록 규정하는 방안이 제안되었으나, 일반 사이버 사고·침해에 한정된 규율에만 해당[29)]

Regulating Artificial Intelligence: U.S. and International Approaches and Considerations for Congress June 4, 2025
24) UK Government, Introducing the AI Safety Institute, 17 Jan 2024
25) The Centre for Long-Term Resilience, AI incident reporting: Addressing a gap in the UK's regulation of AI, June 2024
26) The Centre for Long-Term Resilience, AI incident reporting: Addressing a gap in the UK's regulation of AI, June 2024
27) House of Lords, Artificial Intelligence (Regulation) BILL 18 Mar 2024
28) UK Government, Code of Practice for Cyber Security of AI, 31 Jan 2025

4 미국(US)

4-1. AI사고 정의

■ **AI사고에 대한 연방 차원의 법정 정의를 두고 있지 않으며, 국립표준기술연구소(NIST)의 AI RMF(Risk Management Framework)를 중심으로 AI를 개별 사고가 아닌 '위험관리 대상'으로 다루는 자율적·비구속적 접근을 취함[30]**

- NIST는 OECD 정의를 준용하여 'AI사고'를 AI시스템의 개발·사용·오작동과 관련해 개인과 집단, 사회에 직·간접적 피해를 발생시키는 사고, 상황 또는 일련의 사건으로 정의함
- OECD는 AI위험원과 AI재난까지 포함하는 확장적 개념을 사용하는 반면, NIST는 위험관리의 실무적 관리 단위로 한정하여 사전적 위험 식별·평가·완화 중심의 접근을 취함
- 이와 같이, NIST AI RMF는 AI사고를 독립적으로 규율하기보다는 AI 수명주기 전반의 위험을 통합적으로 관리하는 프레임워크로 기능하며, 신뢰성 확보와 책임 있는 AI 개발·운영을 유도하는 기준으로 활용됨

4-2. AI사고 보고체계

■ **연방 차원의 단일하고 의무화된 AI사고 보고체계는 존재하지 않으며, 사고 보고는 산업별 규제체계에 따라 분산적으로 운영되는 구조[31]**

- 뉴욕, 캘리포니아, 콜로라도 등 일부 주에서는 AI사고 보고 의무를 규정하고 있으나 특정 위험 영역 또는 사례에 한정된 것으로, AI사고 전반을 포괄하는 통합적 보고체계로 기능하지는 않음
- 규제기관별로 특정 AI사고에 대한 보고 의무를 규정
  - 식품의약국(FDA): SaMD(Software as a Medical Device) 및 AI/ML 기반 의료기기에 대해 사망 또는 중대한 부작용 발생 시 의료기기 보고제도(MDR)에 따라 의무 보고 적용[32]
  - 고속도로교통안전국(NHTSA): 자율주행 시스템(ADS) 및 레벨2 ADAS 관련 사고에 대해 충돌 발생 시 일정 요건 하에서 보고 의무를 부과하는 상시명령(Standing General Order) 운영[33]
  - 연방거래위원회(FTC): AI 기반 헬스 앱·웨어러블 등에서 발생한 개인 건강정보 유출 사건에 대해

29) UKAI, UK to expand AI oversight with new cyber resilience law, 23 Jun 2025
30) NIST, Artificial Intelligence Risk Management Framework (AI RMF 1.0), Jan 2023
31) NIST, Artificial Intelligence Risk Management Framework (AI RMF 1.0), Jan 2023
32) 21 CFR part 803(MDR: Medical Device Reporting
33) NHTSA, Standing General Order on Crash Reporting

통지 및 보고 의무를 규정[34)]

■ 학계·의회 차원의 제도 도입 논의

- 미국의 학계 및 싱크탱크는 자발적 AI사고·오류 보고 시스템인 AIIRH(AI Incident Reporting Hub) 구축을 제안[35)]
  - 항공안전보고시스템(ASRS, Aviation Safety Reporting System)·사이버 보안 CVE(Common Vulnerabilites and Exposures)모델을 벤치마킹하여, 연방 차원의 중앙 허브에서 AI사고를 수집·분석·공유하는 체계 구축을 구상
  - ASRS는 자발적 참여, 기밀성 보장, 제재 면제 원칙을 기반으로 운영되는 안전사건 보고체계로, 규제기관인 연방항공국(FAA, Federal Aviation Administration)이 아닌 NASA가 운영을 담당하며 수집된 정보는 익명화된 분석·교훈·권고 및 절차 개선으로 환류되는 구조를 가짐
  - CVE(Common Vulnerabilities and Exposures)는 취약점에 고유 식별자(identifier)를 부여하여 표준화·공개·추적할 수 있도록 하는 체계로, 관련 취약점 정보는 NVD(National Vulnerability Database)의 공개 DB에 메타데이터, 심각도 평가 등으로 저장됨[36)]
  - AI사고에 관한 보편적 정의가 확립되지 않은 상황에서, AIIRH는 초기 보고체계 구축을 위해 OECD의 AI사고 및 AI위험원 정의를 준용할 필요성을 제시[37)]

- AI시스템에 특화된 취약점 관리를 연방 차원에서 공식적으로 다루는 법안(H.R. 9720 「AI Incident Reporting and Security Enhancement Act」) 발의[38)]
  - AI 안전(safety)·보안(security) 사고의 자발적 보고 체계 필요성을 검토하고, 관련 프로세스를 의회에 보고하는 구조를 전제로 설계됨
  - 안전 사고는 ASRS, 보안 사고는 CVE 모델을 각각 참고하여, 두 체계를 결합한 하이브리드형 보고·관리 방식 도입을 지향
  - 주요 입법목적은 AI사고 보고 허브 운영을 위한 법적 권한 부여 및 연방 예산 배정 근거 확보에 있음

34) 16 CFR Part 318 (Health Breach Notification Rule)
35) ARI, ARI Participates in Federation of American Scientists' AI Legislation Policy Sprint; Proposes Establishing an AI Incident Reporting System, 25 June 2024
36) NIST, National Vulnerability Database
37) FAS, Message Incoming: Establish an AI Incident Reporting System, 25 Jun 2024
38) H.R. 9720, AI Incident Reporting and Security Enhancement Act, 118th Cong. 2024

Ⅱ 기타 국가의 AI사고 정의 및 보고체계 동향

5 캐나다(Canada)

5-1. AI사고 정의

■ **캐나다는 연방 차원에서 AI사고에 대한 명시적 정의를 두고 있지 않으며, 이를 통합적으로 규율하는 전담 기관도 부재함**

- 현재 추진 중인 AIDA(Artificial Intelligence and Data Act) 초안은 AI사고를 별도로 정의하지 않고, 대신 '피해(harm)' 개념을 중심으로 규율 구조를 설계[39)]
  - '피해(harm)'는 개인에 대한 신체적·심리적 피해, 재산 피해 및 경제적 손실을 포함하는 개념으로 정의됨

- 다만, 해당 정의는 개인의 피해에 초점을 두고 있어, 집단적 피해(collective harm)나 구조적 불평등과 같은 문제를 충분히 반영하지 못한다는 비판이 제기됨[40)]

5-2. AI사고 보고체계

■ **AI사고를 일괄적으로 규율하는 단일 보고체계를 두고 있지 않으며, 사고 대응 및 보고는 기존 분야별 법률에 따라 분산적으로 이루어짐**

- 개인정보 유출은 개인정보보호 및 전자문서법(PIPEDA, Personal Information Protection and Electronic Documents Act)에 따른 보호조치 위반으로 간주되어 법령에 따라 보고 의무가 발생함[41)]

- 차별·편향 문제는 인권법상 금지된 차별 사유(인종, 민족 또는 출신, 종교, 연령, 성별, 성적 지향, 성별 정체성 및 표현, 혼인·가족 상태, 유전적 특성, 장애 등)에 해당되는 경우 제소 대상이 될 수 있음[42)]

- 보안 침해는 사이버보안 사건(Cyber security incident)으로 분류되어, 정부의 사이버보안 사건 관리체계(CSEMP, Cyber Security Event Management Plan)에 따라 탐지·평가·보고 및 대응 절차를

39) Artificial Intelligence and Data Act(AIDA), Bill C-27, Part 3
40) the dais, Submission on the Proposed Artificial Intelligence and Data Act, Nov 2023
41) PIPEDA, s.2(1), Personal Information Protection and Electronic Documents Act, S.C. 2000, c.5 (Canada)
42) Canadian Human Rights Act, R.S.C., 1985, c. H-6, s.3

통해 처리됨*[43]

* 캐나다 연방정부의 CSEMP는 무단 접근, 악성코드 감염, 데이터 유출 등 보안 침해를 'Cyber security incident'로 정의하고 있으며, AI시스템 자체의 사고에 대해서는 별도의 정의를 두고 있지 않음

■ **이러한 분산적 보고체계를 보완하기 위해, 캐나다는 「AIDA(Artificial Intelligence and Data Act)」를 통해 AI 관련 '피해(harm)'를 기준으로 보고 의무와 정부의 개입 권한을 규정하는 방향의 입법을 추진 중**[44]

- 고영향(high-impact) AI시스템에 대해 피해(harm) 및 편향(bias) 위험의 식별·평가·완화를 의무화하는 위험 기반(risk-based) 규제 체계를 도입하는 것을 주요 내용으로 함
- 불법적으로 취득한 개인정보를 활용한 AI 개발, 또는 중대한 피해를 초래할 수 있는 AI의 무분별한 배포 등에 대해 형사처벌 및 행정제재 규정 마련
- 실질적 피해(material harm)가 발생하였거나 발생할 가능성이 있는 경우, 규제 대상자는 이를 연방 산업부 장관에 보고해야 함[45]

  * 산업부 장관은 AI 규제 집행을 담당하는 ISED(Innovation, Science and Economic Development Canada)의 수장을 겸함

- 또한, AIDA는 장관에게 위험 상황에 대응하기 위한 사전적 개입(preventive intervention) ·사후적 개입(remedial intervention) 권한을 부여[46]
  - (사용·제공 중단 명령) AI시스템이 임박하고 중대한 피해 위험을 초래할 수 있다고 믿을 만한 합리적 근거가 있는 경우, 장관은 해당 시스템의 사용 또는 제공 중단을 명령할 수 있음
  - (정보 제출 및 공개 명령) 장관은 필요한 경우 관련자에게 문서 및 데이터 제출을 요구할 수 있고, 중대한 피해 방지를 위해 해당 시스템의 관련 정보를 대중에 공개하도록 명령할 수 있음

43) Government of Canada, Cyber Security Event Management Plan(CSEMP), Treasury Board of Canada Secretariat, 2019
44) Artificial Intelligence and Data Act (AIDA), Bill C-27, Part 3
45) Artificial Intelligence and Data Act (AIDA), Bill C-27, Part 3 s28(1)
46) Artificial Intelligence and Data Act (AIDA), Bill C-27, Part a3, ss.29-30

6 일본(Japan)

6-1. AI사고 정의

■ AI사고에 대한 법적 정의를 두지 않고, 총무성(MIC)·경제산업성(METI)의 「AI Guidelines for Business(2024)」를 중심으로 위험 기반 접근(risk-based approach)에 따른 AI 위험 관리 체계를 운영[47)]

- AI Guidelines for Business는 생성형AI 확산 등 최근 기술 변화에 대응하기 위해 기존 AI 관련 가이드라인을 통합·개정한 것으로, AI 사업자 전반을 대상으로 AI의 안전하고 책임 있는 개발·활용을 위한 기준을 제시
  - 기존 「AI R&D Guidelines(2017)」, 「AI Utilization Guidelines(2019)」, 「Governance Guidelines for Implementation of AI Principles(2022)」를 통합하여 마련됨

- 일본은 AI가 활용 방식 및 적용 분야에 따라 사회에 중대한 영향을 미칠 수 있다고 보고, 사전에 위험 수준(영향 및 발생 가능성)을 평가하고 이에 비례하는 대응조치를 취하는 것을 핵심 원칙으로 제시
  - 위험에는 신체·재산 피해뿐 아니라 사회적 혼란, 허위정보 확산, 개인정보 및 프라이버시 침해, 공정성 저해 등이 포함되며, 생성형AI 확산에 따라 허위정보 생성·유포, 지식재산권 침해 등 새로운 유형의 위험 가능성에도 주목
  - AI시스템의 오작동 또는 편향된 데이터 학습으로 인해 차별적·불공정한 결과가 발생할 수 있고, 자동화 편향(automation bias) 및 필터버블(filter bubble) 등으로 인간의 의사결정이 왜곡될 가능성을 경고하면서, AI 출력의 한계와 오류 가능성을 전제로 인간의 개입(human oversight) 필요성을 강조

- 일본은 '사고(incident)'를 독립적으로 정의하기보다, AI 전 수명주기(lifecycle)에 걸친 위험 식별·평가·완화 및 거버넌스 구축을 중심으로 규율하는 접근을 채택
  - 이는 OECD가 AI사고, AI위험원, AI재난을 구분하는 체계를 제시한 것과 달리, 사고 자체보다 위험 관리와 거버넌스에 중점을 둔 접근이라는 점에서 차이가 있음

6-2. AI사고 보고체계

■ AI사고에 대한 의무화된 법적 보고체계를 두지 않고, 가이드라인에 따라 기업과 기관이 위

47) Japan, AI Guidelines for Business Ver.1.0, Ministry of Internal Affairs and Communications Ministry of Economy, Trade and Industry, Apr. 19, 2024.

험을 자율적으로 관리하는 연성규범(soft law) 기반 접근을 채택

- 사고 발생 시에는 의료[48], 자동차[49], 개인정보[50] 등 분야별 기존 법령에 따른 사고 보고 체계를 통해 보고 및 대응이 이루어지는 구조

■ 일본 정부는 AI 규제에 대해 라이트 터치(light touch) 접근을 취하고 있으며, 단일 AI 법체계보다는 기존 분야별 법률과 기업의 자율적 위험 완화 조치를 중심으로 대응[51]

- AIDA와 같은 포괄적인 AI 규제 도입보다는 기술 중립성 원칙에 따라 기존 분야별 법률에 기반한 규제 방식을 선호

- 다만 이러한 자율 규제 중심 접근의 한계를 보완하기 위해, 정부는 전략적 리더십 기구를 중심으로 보완적 관리 체계를 구축할 계획
  - 기업의 자율적 사고·위험 보고를 유도하고, 정부 차원의 위험 모니터링 및 정보 수집 체계를 구축하여, 수집된 정보를 정책 개선 및 국제 협력에 활용하는 방향을 제시

■ 일본 정부는 중대한 AI사고 발생 시, 조사 및 정보 공유 메커니즘 도입을 검토중[52]

- 자율 보고 기조를 유지하면서도, 정부의 상시적으로 위험을 모니터링하고 필요 시 개입할 수 있는 보완적 관리 체계 구축 방안을 논의

- 산업계는 증대한 사고 보고의 필요성에는 공감하면서도, 보고 주체·범위, 인과관계 판단 기준, 행정 부담 등을 명확히 할 필요가 있으며, OECD 등 국제 기준과의 정합성을 고려한 단계적 도입이 필요하다는 입장[53]

■ 이러한 논의를 구체화하기 위한 방안으로, 일본 정부는 전략적 리더십 기구의 기능 강화 및 관련 제도 정비 필요성을 제기[54]

- 관계 행정기관에 대한 협력 요청 권한을 포함하여, 전략적 리더십 기구의 역할과 권한을 명확히 하기 위한 입법적 정비 필요성이 제기됨

- 일본 AI안전연구소(AISI)는 AI 안전성 관련 지식 허브로서, 해당 기구의 위험 평가·정보 공유 및 정책 지원 기능을 지속적으로 뒷받침할 것으로 예상

48) Japan, Pharmaceuticals and Medical Devices Act(PMD Act), Act No.145 of 1960, Art.68-10.
49) Japan, Road Transport Vehicle Act, Act No.185 of 1951, Art.63.
50) Japan, Act on the Protection of Personal Information (APPI / PIPA), Act No.57 of 2003, Arts.22-2, 26
51) CSIS, New Government Policy Shows Japan Favors a Light Touch for AI Regulation, 25 Feb 2025
52) Government of Japan, Interim Report of the AI Policy Study Group, Cabinet Secretariat, Feb. 2025
53) BSA, Comments on the Draft Legal Framework for AI Systems in Japan (re: Interim Report), Jan. 25, 2025.
54) Government of Japan, Interim Report of the AI Policy Study Group, Cabinet Secretariat, Feb. 2025

7 싱가포르(Singapore)

7-1. AI사고 정의

■ **'중대 AI사고'의 정의 및 의무 보고 대상 범위 설정 필요성을 인식하고 있으나, 관련 기준은 구체적으로 확립되지 않은 상태[55]**

- AI를 직접적으로 규율하는 단일 법체계는 존재하지 않으며, 개인정보 보호, 사이버보안, 통신, 금융 등 기존 법률이 AI 활용에 간접적으로 적용되는 구조

7-2. AI사고 보고체계

■ **AI사고에 대한 독립적 보고체계는 없으며, 사고 유형에 따라 기존 분야별 체계를 통해 보고 및 대응이 이루어짐**

- 데이터 관련 사고의 경우, 개인정보보호위원회(PDPC)가 운영하는 데이터 침해 신고 채널을 통해 보고가 이루어지며, 조직은 개인정보 유출 등 침해 사실을 감독기관 및 정보주체에 통지해야 함[56]
- 사이버보안 사고의 경우, SingCERT(Cyber Security Agency of Singapore)를 통해 침해 사고를 신고 및 대응 지원이 이루어짐[57]
- 통신·디지털 서비스 관련 취약점의 경우, IMDA(Infocomm Media Development Authority)의 취약점 신고 채널(CSVR)을 통해 보고됨[58]
- 금융 분야의 경우, MAS(Monetary Authority of Singapore)가 정한 내부통제 및 위험 관리 가이드라인에 따라 사고 대응 및 보고 절차가 운영됨[59]

■ **「Model AI Governance Framework for Generative AI(2024)」는 AI사고 보고를 영향 수준에 따라 비례적으로 적용할 필요성을 제시하고, EU AI Act를 참고 모델로 언급함으로써 향후 보고 및 규율 체계 도입 가능성을 시사함[60]**

- AI 관련 사고는 영향 수준에 따라 대중 및 정부 당국에 대한 통지 여부가 결정되는 구조

55) White&Case, AI Watch: Global regulatory tracker – Singapore, 13 May 2024
56) PDPC, Report Your Organisation's Data Breach
57) SingCERT, About Singapore Cyber Emergency Response Team (SingCERT)
58) IMDA, Cyber Security Vulnerability Reporting (CSVR) Guide
59) MAS, Internal Controls, July 2014
60) White&Case, AI Watch: Global regulatory tracker – Singapore, 13 May 2024

8 호주(Australia)

8-1. AI사고 정의

■ **AI를 직접 규율하는 기술특정 법률 또는 규정을 두고 있지 않으며, 이에 따라 법령 차원의 공식적인 AI 정의도 별도로 마련되어 있지 않음**[61]

- 다만, 정책 및 지침에서는 OECD 기반 AI 정의를 참고하여 개념을 설정하고 있으며, 법적 정의 대신 가이드라인 중심으로 개념이 형성되고 있는 것이 특징

■ **이러한 접근에 따라, 호주는 「자발적 AI 안전표준(VAISS, Voluntary AI Safety Standard)」을 통해 위험 기반 접근법(risk-based approach)에 따른 AI 위험 및 피해 관리 원칙을 제시**[62]

- 개인에 대한 피해 : 시민적 자유 및 기본권 침해, 신체적·정신적 안전에 대한 위험 등 개인 수준에서 발생하는 직접적 피해
- 집단 및 커뮤니티에 대한 피해 : AI시스템의 편향 또는 차별적 결과로 인해 특정 집단에 대한 불리한 영향 확대 및 사회적 불평등 심화
- 사회 구조에 대한 피해 : 민주적 의사결정, 정보 접근, 교육 기회 등 사회 전반의 구조적 시스템에 미치는 광범위한 영향
- 이와 같이 VAISS는 기술적 위험을 넘어 권리, 형평성, 민주주의 등 사회적 가치에 대한 영향을 포함하여 피해 개념을 포괄적으로 규정하고 있음

■ **VAISS는 AI의 안전하고 책임 있는 개발·배포를 위한 기준을 제시하기 위해 10개의 자율적 가드레일(guardrails)을 핵심 구성요소로 제시**[63]

- 가드레일은 책임성(accountability), 위험관리(risk management), 데이터 거버넌스 및 시스템 보호(data governance & protection), 테스트 및 모니터링(testing & monitoring), 인간 감독 및 개입(human oversight), 이용자 고지(user disclosure), 이의제기 및 대응절차(challenge & contestability), 공급망 투명성(supply chain transparency), 기록관리(record keeping), 이해관계자 참여(stakeholder engagement) 등으로 구성되며, AI 전 수명주기에 걸친 안전·책임성 확보 기준을 제시함

61) White&Case, AI Watch: Global regulatory tracker – Australia, 25 Sep 2025
62) Australian Government, Voluntary AI Safety Standard, Aug 2024
63) Australian Government, Voluntary AI Safety Standard, Aug 2024

| 표 2. VAISS의 10대 가드레일 |

번호	제목	주요 내용
1	책임성(Accountability)	거버넌스, 내부 역량, 규제 준수 전략을 포함한 책임 프로세스를 확립하고 구현함
2	위험관리 (Risk Management)	AI 영향과 위험을 평가하고 잠재적 위해를 식별 및 완화하는 위험 관리 프로세스를 구축함
3	데이터 거버넌스 및 시스템 보호 (Data Governance & Protection)	데이터 품질, 출처, 사이버 취약성을 관리하기 위한 데이터 거버넌스, 개인정보 보호, 사이버 보안 조치를 구현함
4	테스트 및 모니터링 (Testing & Monitoring)	배포 전에 AI시스템을 철저히 테스트하고, 배포 후 잠재적 행동 변화를 지속적으로 모니터링
5	인간 감독 및 개입 (Human Oversight)	의미 있는 인간 감독을 달성하기 위해 AI시스템 전체 수명주기에 걸쳐 인간 통제 또는 개입 메커니즘을 확보
6	이용자 고지 (User Disclosure)	AI 활용 여부, 역할 및 생성 콘텐츠 여부를 이용자에게 명확히 고지하여 AI 사용에 대한 인지와 이해를 보장
7	이의제기 및 대응절차 (Challenge & Contestability)	AI 결정·출력에 대한 이의제기 및 검토 절차를 마련
8	공급망 투명성 (Supply Chain Transparency)	데이터·모델·시스템 정보를 공급망 내 공유
9	기록관리 (Record Keeping)	AI시스템 목록 및 문서화 기록을 유지
10	이해관계자 참여 (Stakeholder Engagement)	안전, 다양성, 포용성, 공정성에 중점을 두고 영향받는 이해관계자 참여

출처 Australian Government, Voluntary AI Safety Standard, Aug 2024

8-2. AI사고 보고체계

■ AI사고에 대한 별도의 법정 보고체계를 두고 있지 않으며, 사고 대응 및 보고는 기존 법률 체계에 따라 이루어짐[64]

- 개인정보 보호법에 따른 데이터 유출 신고, 소비자법에 따른 제품 안전 문제 보고, 온라인 안전법에 따른 유해 콘텐츠 관련 보고 등 사고 유형별로 개별 법률에 따른 보고 의무가 적용

■ 다만, 호주 정부는 자율적 기준을 넘어 고위험 AI에 대한 규제 체계(2024 「Mandatory guardrails for safe and responsible AI」) 도입 가능성을 검토하고 있음[65]

- 고위험 AI의 정의와 함께 10개 의무적 가드레일 도입 방안을 제시하였으며, 이는 기존 자율적 AI 안전표준(VAISS)을 기반으로 향후 의무 규제로 전환될 가능성을 시사하는 조치로 평가됨

64) Australian Government, Voluntary AI Safety Standard, Aug 2024
65) Australian Government, Mandatory guardrails for safe and responsible AI: have your say, 5 Sep 2024

9 케냐(Kenya)

9-1. AI사고 정의

■ **AI를 직접적으로 규율하는 법률이 존재하지 않아 AI사고에 대한 법적 정의도 마련되어 있지 않으며, 관련 규제 체계는 초안 단계에 머물러 있음[66)]**

- 현재 케냐에는 AI를 직접적으로 규율하는 독립적 법률이나 하위 규정이 존재하지 않으며, 정보기술 인공지능 실무규정(Draft Code of Practice), 로보틱스 및 AI 법안(「Robotics and Artificial Intelligence Bill」) 등이 논의되고 있으나, 아직 입법 단계에는 이르지 않음
- 다만 「Kenya Artificial Intelligence Strategy 2025-2030」에서는 AI를 머신러닝, 데이터 처리 및 알고리즘 기반 기술로 정의하는 등 정책적 수준에서의 개념 정립은 이루어지고 있음[67)]

9-2. AI사고 보고체계

■ **AI사고에 대한 독립적 보고체계는 존재하지 않으며, 관련 사고 대응 및 보고는 기존 법률 체계에 따라 이루어짐**

- 데이터보호법(2019)은 자동화된 처리(프로파일링 포함)에 기반한 결정과 관련하여 데이터 주체의 권리를 보장하고 있으며, AI 활용에 따른 피해 대응의 법적 근거로 기능[68)]
- 사이버보안 규정에 따라 특정 산업(예: 결제서비스 분야)에서는 중대한 사이버 사고 발생 시 감독기관(CBK)에 대한 보고 의무가 부과됨[69)]
  * 결제서비스 제공업체(PSP, Payment Service Provider)는 고객 서비스 제공 능력, 평판 또는 재정 상태에 중대한 영향을 미칠 수 있는 사이버보안 사고를 중앙은행 CBK(Central Bank of Kenya)에 24시간 이내 보고하도록 요구됨

■ **케냐 정부는 2025년 「Kenya Artificial Intelligence Strategy 2025-2030[70)]」을 통해 포괄적 AI 거버넌스 프레임워크 구축을 추진하고 있으며, 이는 향후 AI사고 정의 및 보고체계 도입으로 이어질 것으로 예상됨**

66) White&Case, AI Watch: Global regulatory tracker – Kenya, 28 Apr 2025
67) White&Case, AI Watch: Global regulatory tracker – Kenya, 28 Apr 2025
68) Kenya, Data Protection Act, 2019 (No. 24 of 2019)
69) CMS, Data protection and cybersecurity laws in Kenya, 11 Mar 2022
70) Kenya Government, Kenya Artificial Intelligence Strategy 2025-2030, Mar 2025

AI 안전 동향 분석 시리즈 26-01(2026년 5월)
「AI사고 정의 및 보고체계 국제동향」

발 행 일 2026년 5월 19일
발 행 처 한국전자통신연구원 인공지능안전연구소 AI안전정책및대외협력실
발 행 인 최 민 석 (AI안전정책및대외협력실 실장)
연 구 진 송 경 호 (AI안전정책및대외협력실 선임연구원)
주 지 연 (AI안전정책및대외협력실 석사후연수연구원)
전 화 번 호 031-739-7651
주 소 (13488) 경기도 성남시 분당구 대왕판교로712번길 22, 글로벌 R&D센터 A동 4층
편집·디자인 ※ 본 자료의 표지 디자인에는 생성형 AI 기반 이미지 제작 도구가 활용됨.

본 자료의 저작권은 한국전자통신연구원 인공지능안전연구소에 있으며, 무단 복제·배포·상업적 이용을 금합니다.

AI 안전 동향 분석 리포트 26-01

AI사고 정의 및 보고체계 국제동향

Global Trends in AI Accident Definition and Reporting Frameworks



ETRI 한국전자통신연구원 www.etri.re.kr
인공지능안전연구소 AI Safety Institute

95560
9 788955 193626