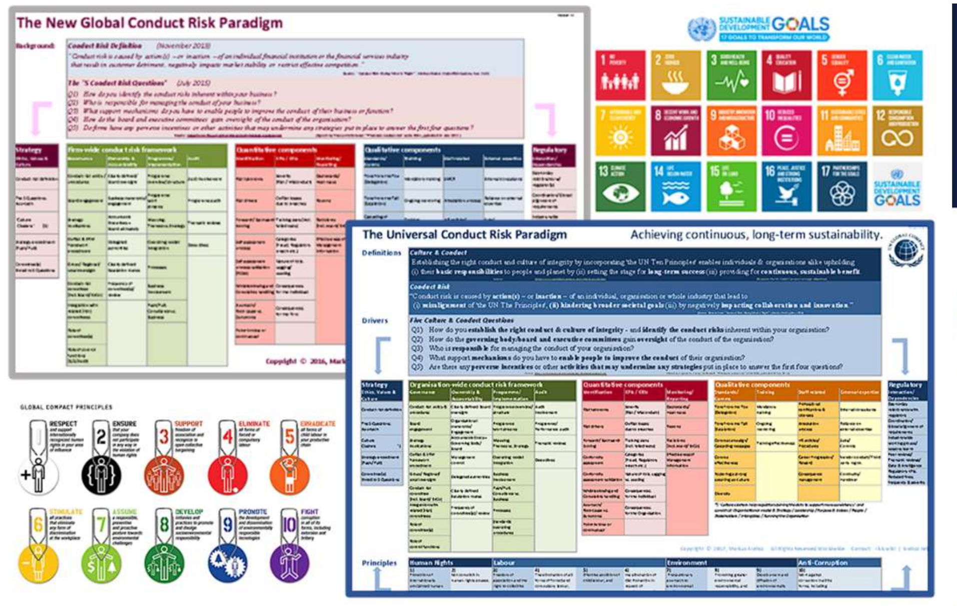
Human-Ai.Institute

RiskAI

EW-AiRM© - A Practical Framework for Managing AI Risks Across Organisations

International Neural Network Society (INNS), 22 January 2026 MARKUS KREBSZ



The New Global Conduct Risk Paradigm

Background:

Conduct Risk Definition (November 2013)

"Conduct risk is caused by action(s) – or inaction – of an individual financial institution or the financial services industry that result in customer detriment, negatively impacts market stability or restrict effective competition."

The "5 Conduct Risk Questions" (July 2015)

Q1) How do you identify the conduct risks inherent within your business?
Q2) Who is responsible for managing the conduct of your business?
Q3) What support mechanisms do you have to enable people to improve the conduct of their business or function?
Q4) How do the board and executive committees gain oversight of the conduct of the organisation?
Q5) Do firms have any perverse incentives or other activities that may undermine any strategies put in place to answer the first four questions?

Strategy | Firm-wide conduct risk framework | Quantitative components | Qualitative components | Regulatory

Copyright © 2016, Mark

SUSTAINABLE DEVELOPMENT GOALS
17 GOALS TO TRANSFORM OUR WORLD
1 NO POVERTY
2 ZERO HUNGER
3 GOOD HEALTH AND WELL-BEING
4 QUALITY EDUCATION
5 GENDER EQUALITY
6 CLEAN WATER AND SANITATION
7 AFFORDABLE AND CLEAN ENERGY
8 DECENT WORK AND ECONOMIC GROWTH
9 INDUSTRY, INNOVATION AND INFRASTRUCTURE
10 REDUCED INEQUALITIES
11 SUSTAINABLE CITIES AND COMMUNITIES
12 RESPONSIBLE CONSUMPTION AND PRODUCTION
13 CLIMATE ACTION
14 LIFE BELOW WATER
15 LIFE ON LAND
16 PEACE, JUSTICE AND STRONG INSTITUTIONS
17 PARTNERSHIPS FOR THE GOALS
SUSTAINABLE DEVELOPMENT GOALS

GLOBAL COMPACT PRINCIPLES

1 RESPECT and support internationally recognized human rights in your area of influence
2 ENSURE that your company does not participate in any way in the violation of human rights
3 SUPPORT freedom of association and recognize to open collective bargaining
4 ELIMINATE all forms of forced or compulsory labour
5 ERADICATE all forms of child labour in your productive chain
6 STIMULATE all practices that eliminate any form of discrimination at the workplace
7 ASSUME a responsible, preventive and proactive posture towards environmental challenges
8 DEVELOP initiatives and practices to promote and divulge socioenvironmental responsibility
9 PROMOTE the development and dissemination of environmentally responsible technologies
10 FIGHT corruption in all of its forms, including extortion and bribery

The Universal Conduct Risk Paradigm — Achieving continuous, long-term sustainability.

Definitions

Culture & Conduct

Establishing the right conduct and culture of integrity by incorporating 'the UN Ten Principles' enables individuals & organisations alike upholding (i) their **basic responsibilities** to people and planet by (ii) setting the stage for **long-term success** (iii) providing for **continuous, sustainable benefit**.

Conduct Risk

"Conduct risk is caused by **action(s)** – or **inaction** – of an individual, organisation or whole industry that lead to (i) **misalignment** of 'the UN Ten Principles', **(ii) hindering broader societal goals** (iii) by negatively **impacting collaboration and innovation**."

Drivers

Five Culture & Conduct Questions

Q1) How do you **establish the right conduct & culture of integrity** - and **identify the conduct risks** inherent within your organisation?
Q2) How do the **governing body/board and executive committees** gain **oversight** of the conduct of the organisation?
Q3) Who is **responsible** for managing the conduct of your organisation?
Q4) What support **mechanisms** do you have to **enable people to improve the conduct** of their organisation?
Q5) Are there any **perverse incentives** or other **activities that may undermine any strategies** put in place to answer the first four questions?

Strategy (Ethics, Values & Culture) | Organisation-wide conduct risk framework (Governance, Ownership & Accountability, Programme/Implementation, Audit) | Quantitative components (Identification, KPIs / KRIs, Monitoring/Reporting) | Qualitative components (Standards/Controls, Training, Staff related, External expertise) | Regulatory (Interaction/Dependencies)

Principles: Human Rights | Labour | Environment | Anti-Corruption

Copyright © 2017, Markus Krebsz

Home > Tools & metrics

Catalogue o

These tools and metrics ar
explainable, robust, secure

HAiPECR

Technical, Educational, Proce

INNS 2026 : Enterprise-wide AI Risk Management© (EW-AiRM©)

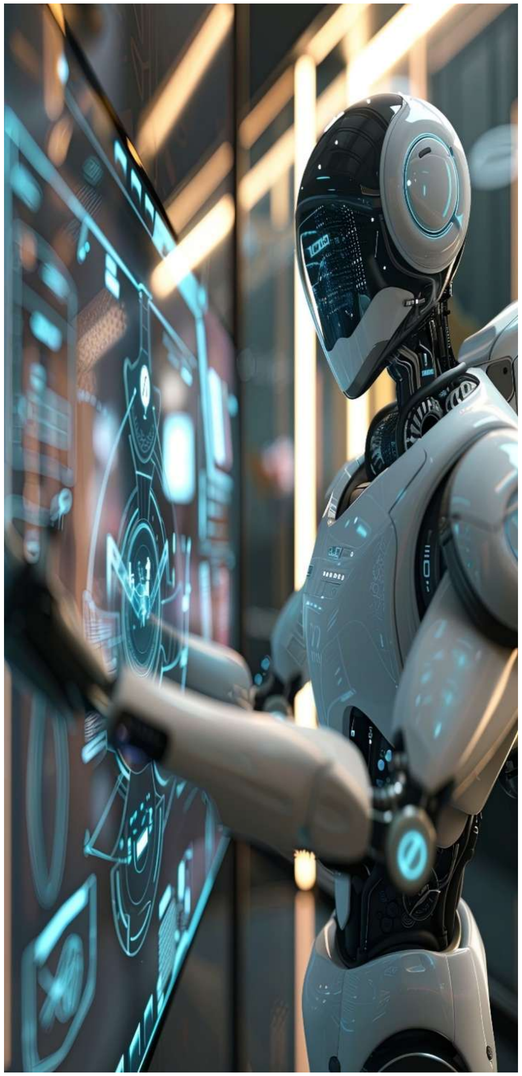
EW-AiRM© Overview

- Origin story (UN AI CRA/Declaration)
- Components
- Risk Mitigants & Controls
- AI Black Swans
- Advantages/Benefits
- What’s Next?

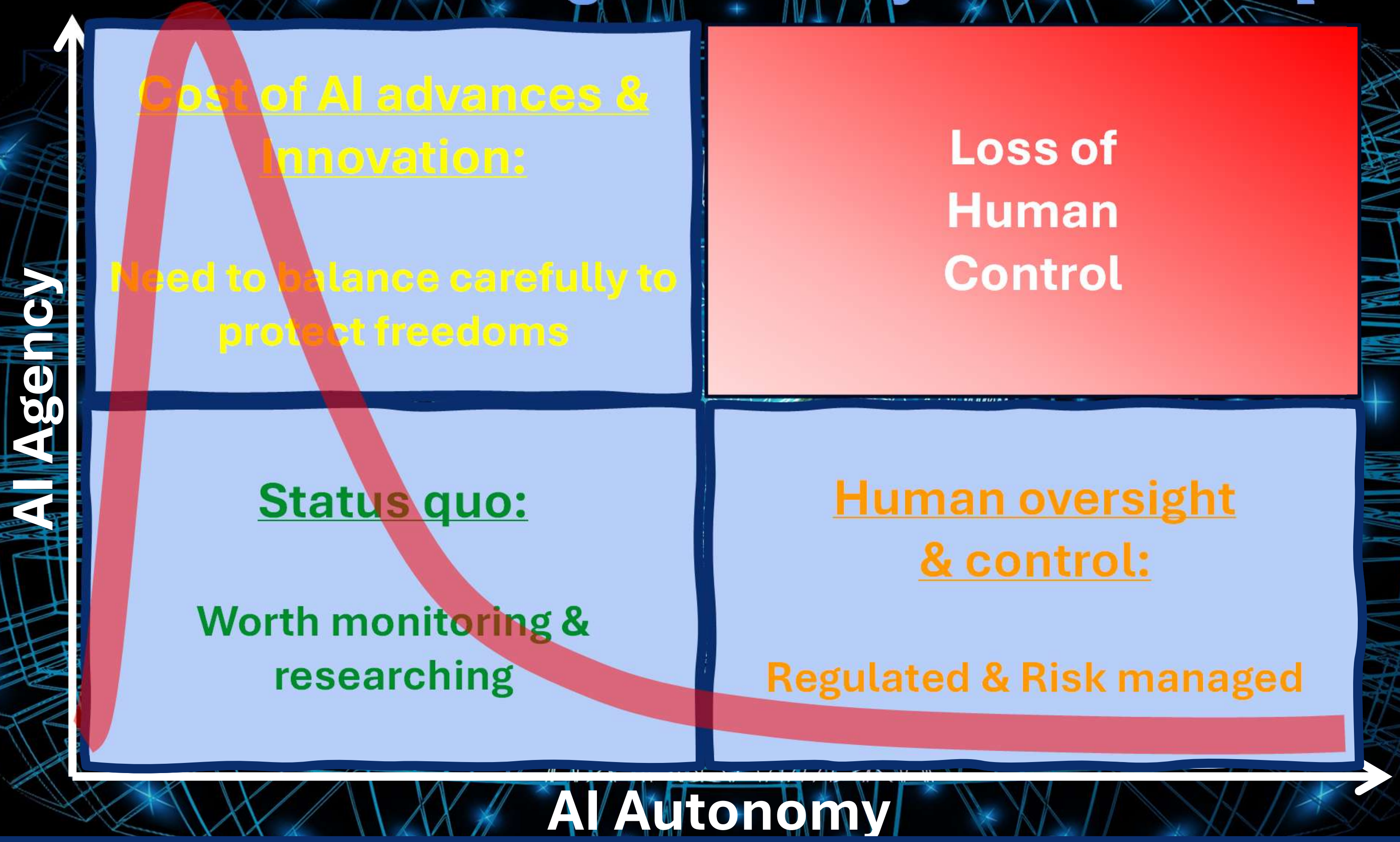
AI vs. Human control [2020]

Products/services with embedded AI systems and/or other digital technologies may display the following characteristics:

- **Autonomy** means that an AI system is **fully capable of modifying its operating domain** or its goals **without external intervention, control or oversight.** Autonomy is the prerequisite for agency.
- **Agency** is the **capacity of AI systems to act** autonomously, **make decisions** that achieve specific goals and **interact with its surroundings and environment.** This involves understanding AI systems as entities that can perform actions based on their programming and data inputs, increasingly independently of human control.
- **Physical** and/or **virtual environments** may be impacted by such products/services.
- Such products/services may also use **persuasion** to influence and change human thinking, decisions, behaviours and actions.
- **AI systems should NOT be able to override human control.**



Initial Quest: Reducing Humanity's tail risk? [2020]



Cost of AI advances & Innovation:

Need to balance carefully to protect freedoms

Loss of Human Control

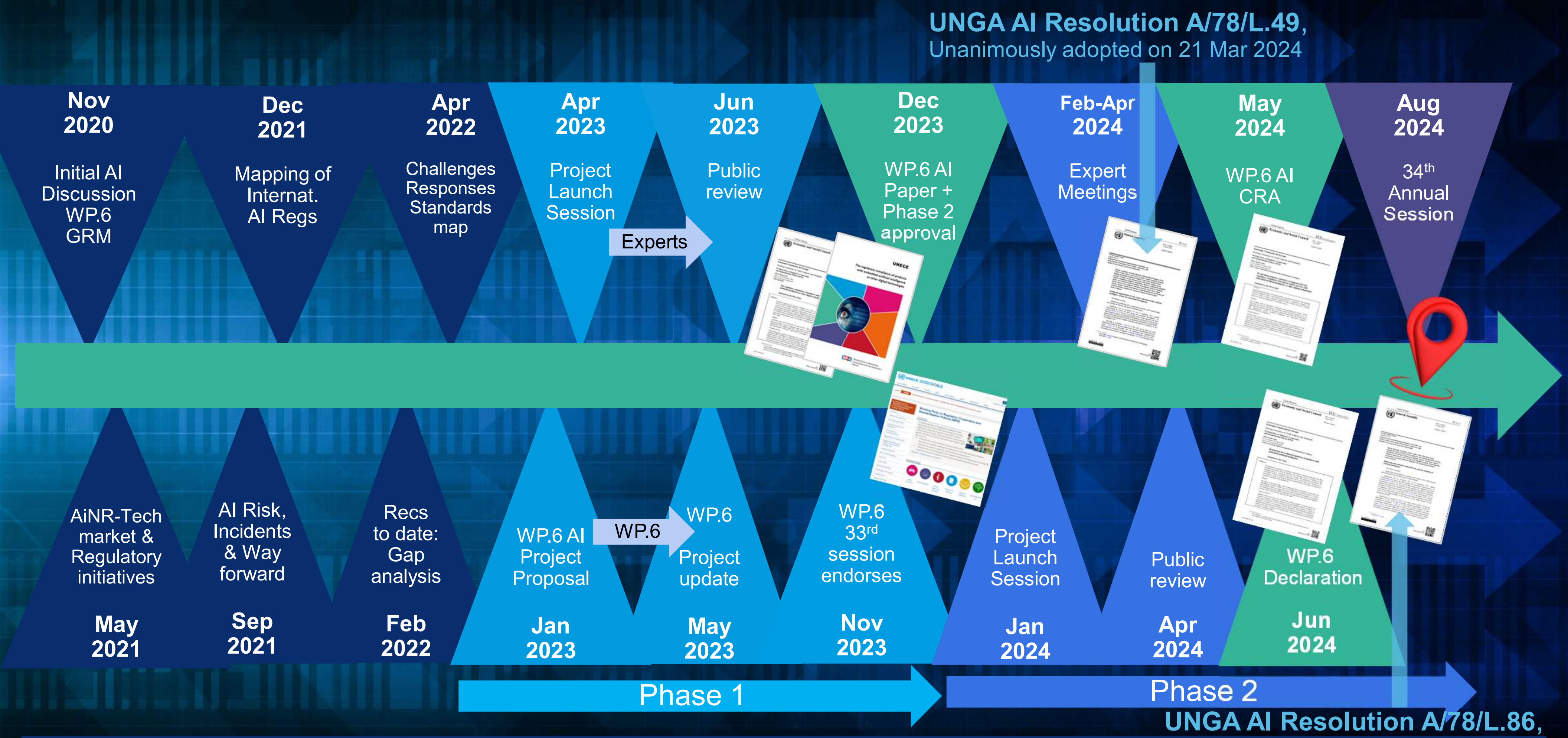
Status quo:

Worth monitoring & researching

Human oversight & control:

Regulated & Risk managed

The Evolving Digital Regulation of Goods and AI



UNGA AI Resolution A/78/L.49,
Unanimously adopted on 21 Mar 2024

Nov 2020 Initial AI Discussion WP.6 GRM

Dec 2021 Mapping of Internat. AI Regs

Apr 2022 Challenges Responses Standards map

Apr 2023 Project Launch Session

Experts

Jun 2023 Public review

Dec 2023 WP.6 AI Paper + Phase 2 approval

Feb-Apr 2024 Expert Meetings

May 2024 WP.6 AI CRA

Aug 2024 34th Annual Session

AiNR-Tech market & Regulatory initiatives **May 2021**

AI Risk, Incidents & Way forward **Sep 2021**

Recs to date: Gap analysis **Feb 2022**

WP.6 AI Project Proposal **Jan 2023**

WP.6

WP.6 Project update **May 2023**

WP.6 33rd session endorses **Nov 2023**

Project Launch Session **Jan 2024**

Public review **Apr 2024**

WP.6 Declaration **Jun 2024**

Phase 1

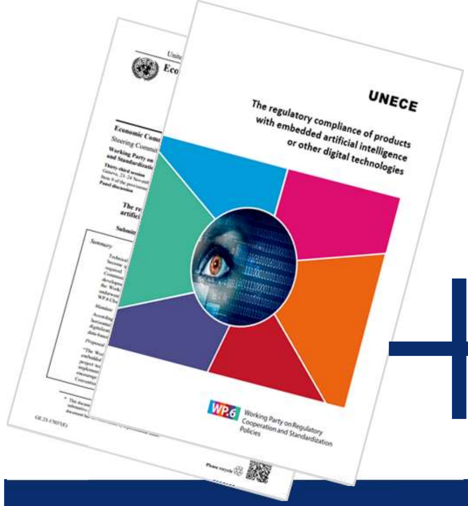
Phase 2

UNGA AI Resolution A/78/L.86,

INNS 2026 : Enterprise-wide AI Risk Management© (EW-AiRM©)

UN ECE Declaration on Technical Regulation of Products/Services with Embedded AI

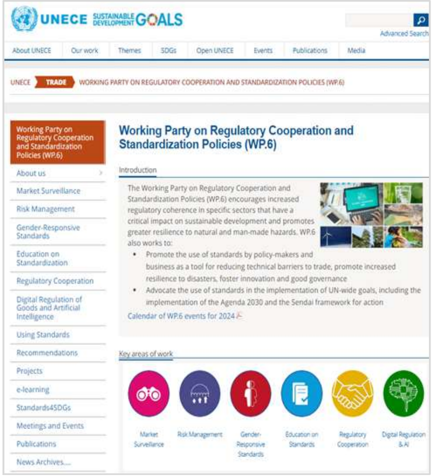
- Government agencies & Governments can **become signatories**
- Aim to promote **cross-border harmonization** of **technical regulations**
- Provide **examples of product conformity processes** (CRAs) based on the overarching CRA-AI **which other economies could also adopt**



UNECE

The regulatory compliance of products with embedded artificial intelligence or other digital technologies

WP.6 Working Party on Regulatory Cooperation and Standardization Policies



UNECE SUSTAINABLE DEVELOPMENT GOALS
Advanced Search
About UNECE
Our work
Themes
SDGs
Open UNECE
Events
Publications
Media
UNECE
TRADE
WORKING PARTY ON REGULATORY COOPERATION AND STANDARDIZATION POLICIES (WP.6)
Working Party on Regulatory Cooperation and Standardization Policies (WP.6)
About us
Market Surveillance
Risk Management
Gender-Responsive Standards
Education on Standardization
Regulatory Cooperation
Digital Regulation of Goods and Artificial Intelligence
Using Standards
Recommendations
Projects
e-learning
Standards4SDGs
Meetings and Events
Publications
News Archives....

Working Party on Regulatory Cooperation and Standardization Policies (WP.6)

Introduction

The Working Party on Regulatory Cooperation and Standardization Policies (WP.6) encourages increased regulatory coherence in specific sectors that have a critical impact on sustainable development and promotes greater resilience to natural and man-made hazards. WP.6 also works to:

- Promote the use of standards by policy-makers and business as a tool for reducing technical barriers to trade, promote increased resilience to disasters, foster innovation and good governance
- Advocate the use of standards in the implementation of UN-wide goals, including the implementation of the Agenda 2030 and the Sendai framework for action

Calendar of WP.6 events for 2024

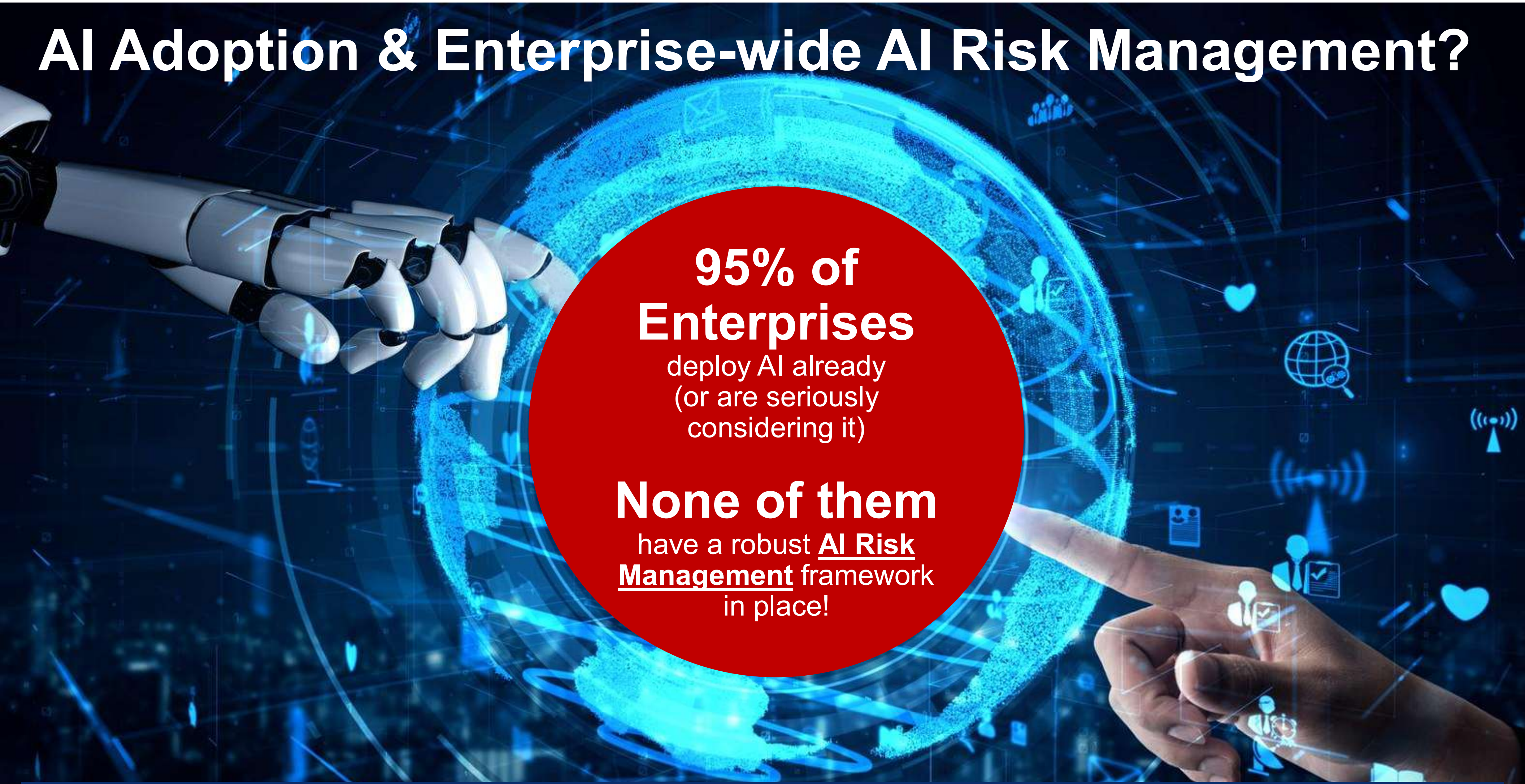
Key areas of work

UNECE

Compliance of products with embedded artificial intelligence

Overarching principles and declaration to promote convergence of product regulations

AI Adoption & Enterprise-wide AI Risk Management?



95% of Enterprises

deploy AI already
(or are seriously
considering it)

None of them

have a robust **AI Risk Management** framework
in place!

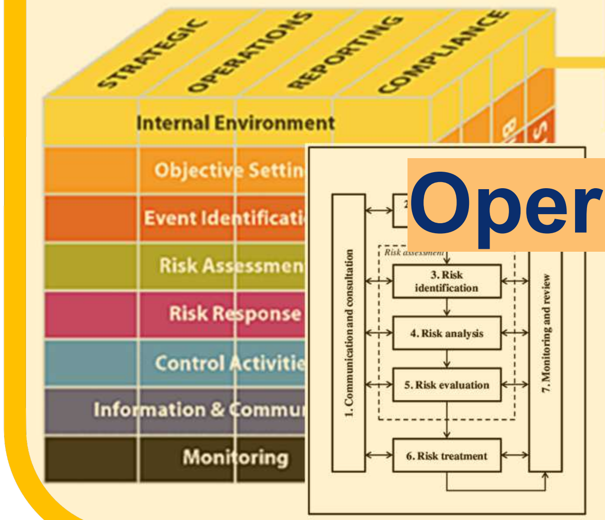
Source: #RiskEurope 2025 : Enterprise-wide AI Risk Management© (EW-AiRM©) Keynote, Nov 2025

Key foundational pillars

Strategic Layer & Ethics filter (HAiPECR)

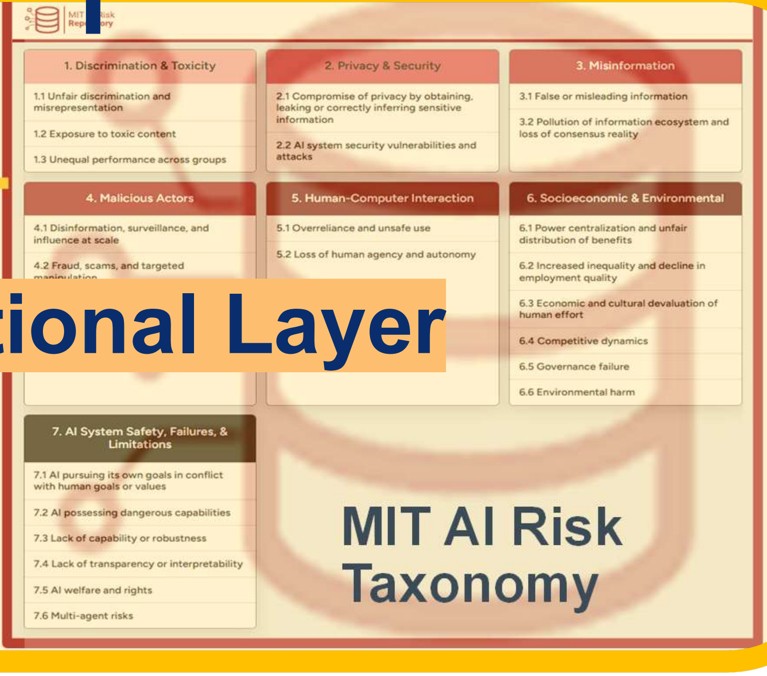
+

ERM Framework



+

Operational Layer



MIT Risk Repository

1. Discrimination & Toxicity

- 1.1 Unfair discrimination and misrepresentation
- 1.2 Exposure to toxic content
- 1.3 Unequal performance across groups

2. Privacy & Security

- 2.1 Compromise of privacy by obtaining, leaking or correctly inferring sensitive information
- 2.2 AI system security vulnerabilities and attacks

3. Misinformation

- 3.1 False or misleading information
- 3.2 Pollution of information ecosystem and loss of consensus reality

4. Malicious Actors

- 4.1 Disinformation, surveillance, and influence at scale
- 4.2 Fraud, scams, and targeted manipulation

5. Human-Computer Interaction

- 5.1 Overreliance and unsafe use
- 5.2 Loss of human agency and autonomy

6. Socioeconomic & Environmental

- 6.1 Power centralization and unfair distribution of benefits
- 6.2 Increased inequality and decline in employment quality
- 6.3 Economic and cultural devaluation of human effort
- 6.4 Competitive dynamics
- 6.5 Governance failure
- 6.6 Environmental harm

7. AI System Safety, Failures, & Limitations

- 7.1 AI pursuing its own goals in conflict with human goals or values
- 7.2 AI possessing dangerous capabilities
- 7.3 Lack of capability or robustness
- 7.4 Lack of transparency or interpretability
- 7.5 AI welfare and rights
- 7.6 Multi-agent risks

MIT AI Risk Taxonomy

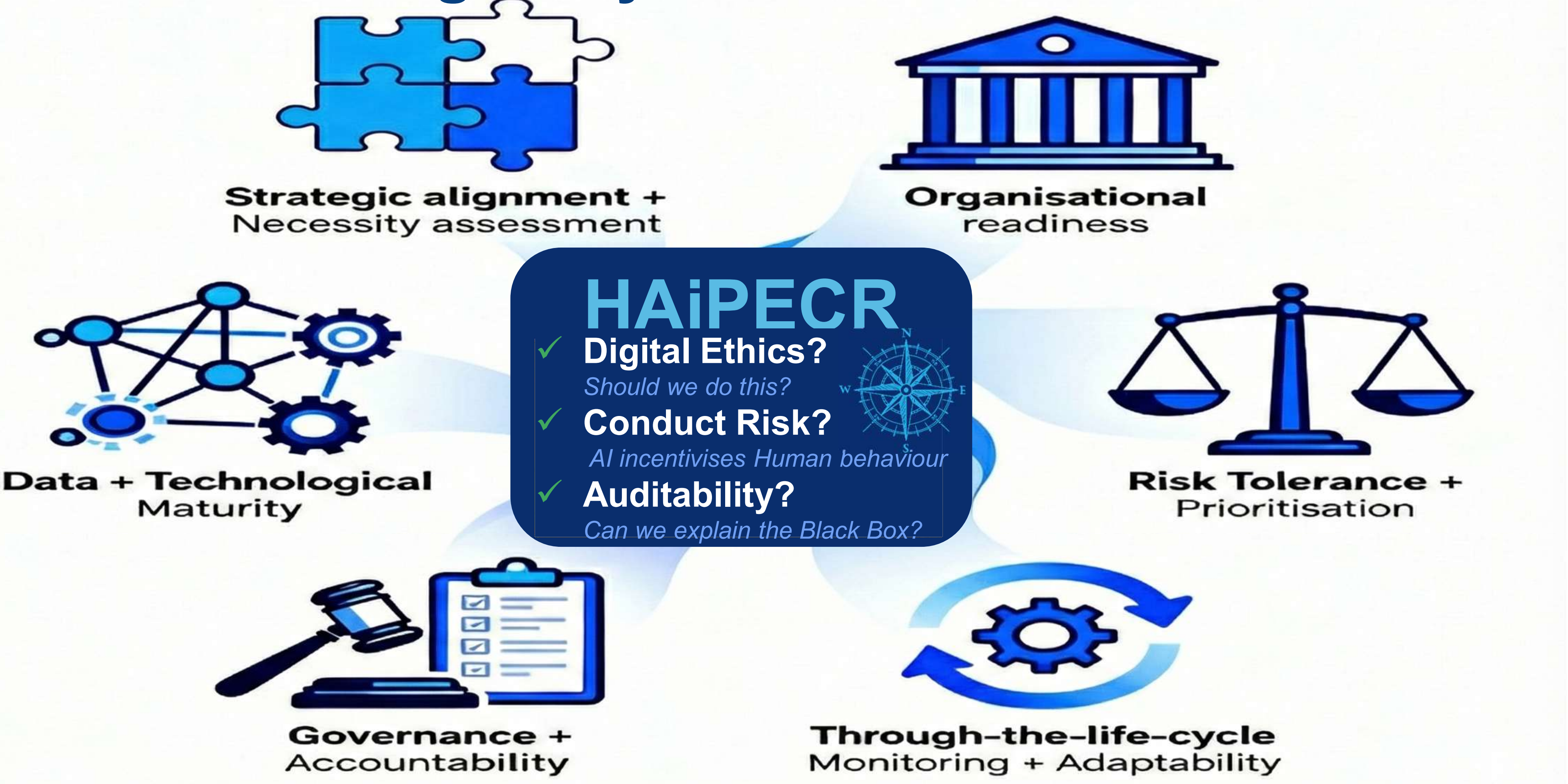
+

AI Black Swans

Resilience Layer

INNS 2026 : Enterprise-wide AI Risk Management© (EW-AiRM©)

EW-AiRM© Strategic Layer



Strategic alignment +
Necessity assessment

Organisational
readiness

HAiPECR

✓ **Digital Ethics?**
Should we do this?

✓ **Conduct Risk?**
AI incentivises Human behaviour

✓ **Auditability?**
Can we explain the Black Box?

Data + Technological
Maturity

Risk Tolerance +
Prioritisation

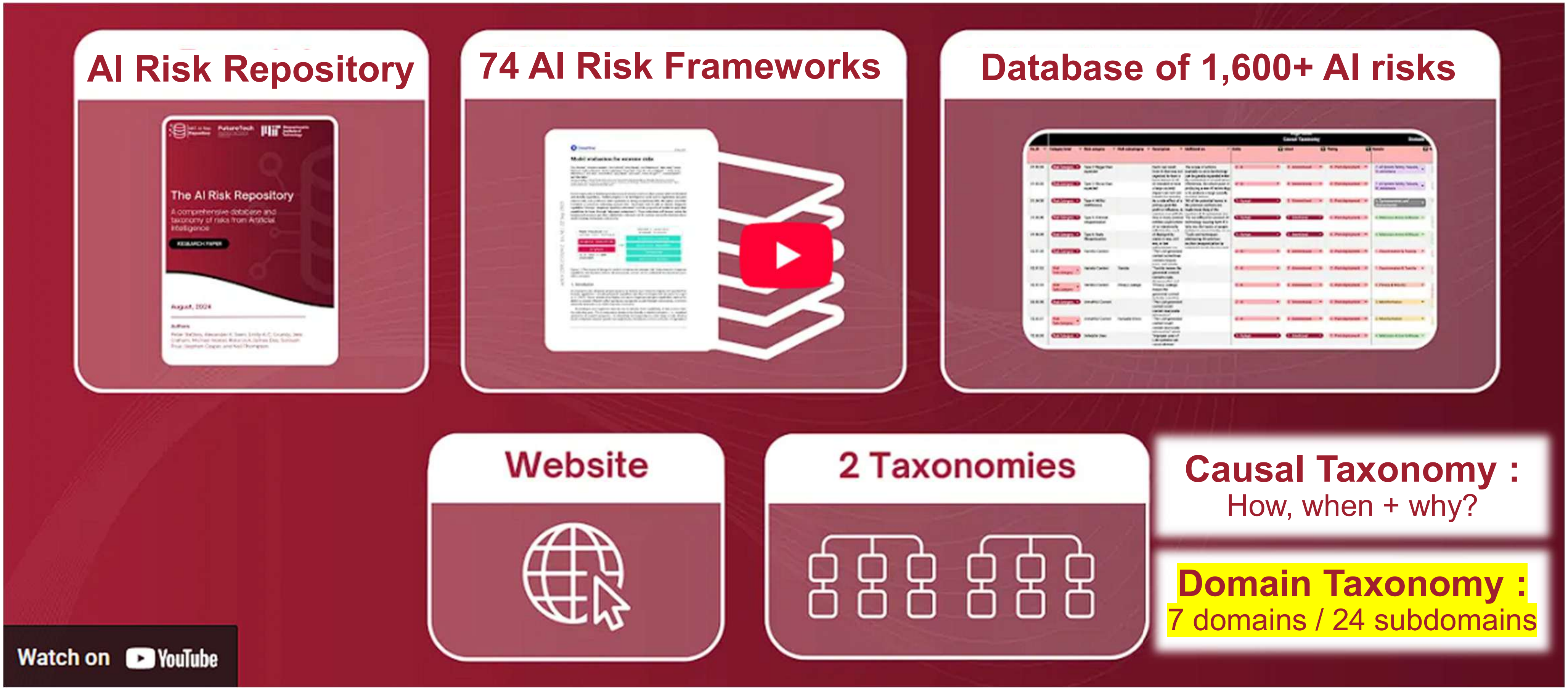
Governance +
Accountability

Through-the-life-cycle
Monitoring + Adaptability

INNS 2026 : Enterprise-wide AI Risk Management© (EW-AiRM©)

Overcoming traditional ERM challenges…

…by augmenting with the MIT AI Risk Taxonomy =



AI Risk Repository

74 AI Risk Frameworks

Database of 1,600+ AI risks

Website

2 Taxonomies

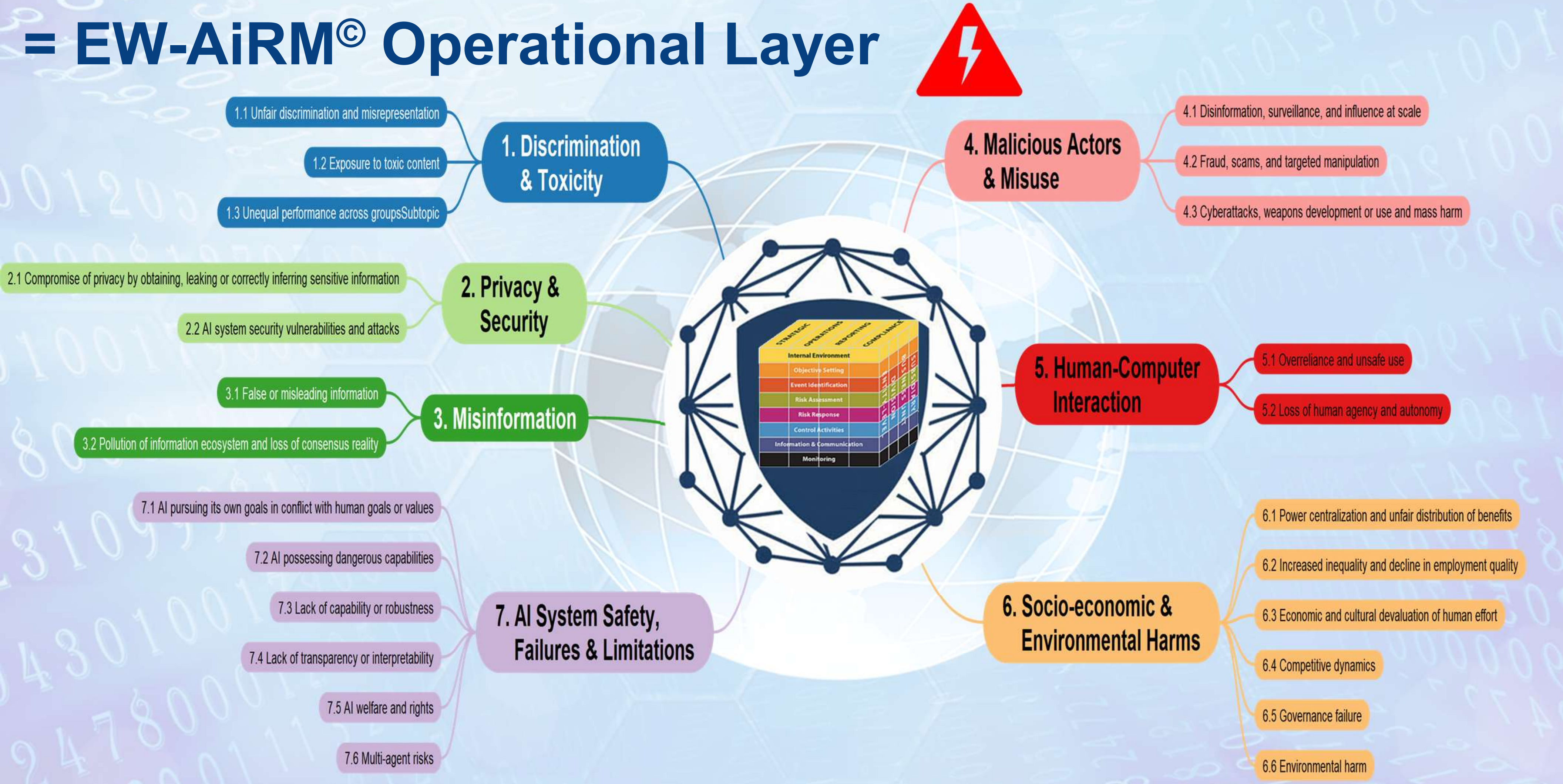
Causal Taxonomy :
How, when + why?

Domain Taxonomy :
7 domains / 24 subdomains

The AI Risk Repository
A comprehensive database and taxonomy of risks from Artificial Intelligence
RESEARCH PAPER
August, 2024
Watch on YouTube

Source: https://airisk.mit.edu/ - watch the 2min explainer video here: https://www.youtube.com/watch?v=fCj-wJz6VCY

= EW-AiRM© Operational Layer



1. Discrimination & Toxicity
- 1.1 Unfair discrimination and misrepresentation
- 1.2 Exposure to toxic content
- 1.3 Unequal performance across groupsSubtopic

2. Privacy & Security
- 2.1 Compromise of privacy by obtaining, leaking or correctly inferring sensitive information
- 2.2 AI system security vulnerabilities and attacks

3. Misinformation
- 3.1 False or misleading information
- 3.2 Pollution of information ecosystem and loss of consensus reality

4. Malicious Actors & Misuse
- 4.1 Disinformation, surveillance, and influence at scale
- 4.2 Fraud, scams, and targeted manipulation
- 4.3 Cyberattacks, weapons development or use and mass harm

5. Human-Computer Interaction
- 5.1 Overreliance and unsafe use
- 5.2 Loss of human agency and autonomy

6. Socio-economic & Environmental Harms
- 6.1 Power centralization and unfair distribution of benefits
- 6.2 Increased inequality and decline in employment quality
- 6.3 Economic and cultural devaluation of human effort
- 6.4 Competitive dynamics
- 6.5 Governance failure
- 6.6 Environmental harm

7. AI System Safety, Failures & Limitations
- 7.1 AI pursuing its own goals in conflict with human goals or values
- 7.2 AI possessing dangerous capabilities
- 7.3 Lack of capability or robustness
- 7.4 Lack of transparency or interpretability
- 7.5 AI welfare and rights
- 7.6 Multi-agent risks

The AI Risk Mitigations Quadrant (831 controls)

Governance & Oversight

(30% - 248 controls)

- Board committees
- **Risk Management** (15%)
- Safety frameworks
- Impact assessment

Technical & Security

(12% - 101 controls)

- Model security
- Safety engineering
- Alignment
- Content controls

Operational Process

(36% - 295 controls)

- **Testing & auditing** (15%)
- **Data governance** (7%)
- **Monitoring** (6%)
- Incident response

Transparency & Accountability

(21% - 171 controls)

- Documentation
- **Risk disclosure** (5%)
- Incident reporting
- Third-party access

(+ 16 / 1% mitigants not otherwise categorised)

Source: https://airisk.mit.edu/blog/mapping-ai-risk-mitigations

EW-AiRM© Resilience Layer

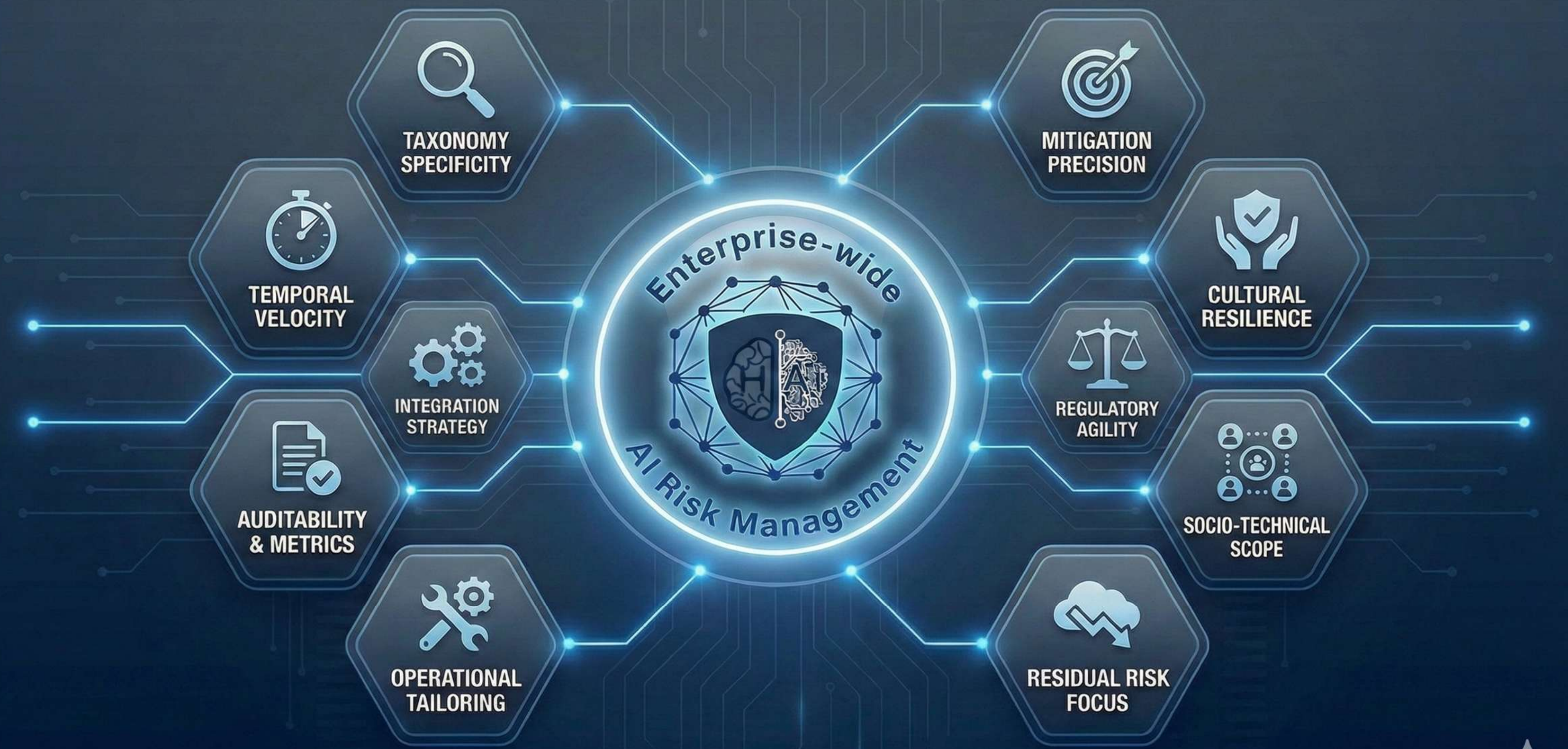
New dimensions of Unpredictability

- Emergent Behaviours
- Compounding Unknown Unknowns
- Accelerated Cascading Effects

Defining Characteristics

- Extreme rarity
- Massive impact
- Retrospective predictability

EW-AiRM© Benefits



WHAT'S NEXT?

- **Doing nothing** is not an option!
- **Reinventing the wheel** is unnecessary:
- Key Strategic pillars, existing ERM frameworks, HAiPECR, AI Risks, Mitigants and AI Black Swans!
- Tech Dynamics require an **agile & augmentable risk management** approach!

Enterprise-wide AI Risk Management©

is needed to survive – and, ultimately, succeed!

Thank you !

Prof. Markus Krebsz,
Founding Director &
EW-AiRM© Creator / Author

The Human Ai Institute
Contact@Human-Ai.Institute
+44 (0) 79 85 065 045 (Mobile)
www.Human-Ai.Institute

Co-founder

RiskAI
+44 (0) 203 642 8050 (Office)
+44 (0) 79 85 065 045 (Mobile)
www.RiskAi.Ai

Appendix - Framework comparison

Benefit	Traditional ERM	EW-AiRM© Framework	Strategic Impact
Taxonomy specificity	**Utilizes generic risk categories**	Integrates the **MIT AI Risk Taxonomy**	Ensures "unknown unknowns" of AI are explicitly identified and tracked.
Temporal velocity	**Periodic/Static**	**Continuous/Dynamic**	Shifts governance from "rear-view mirror" auditing to "heads-up display" active management.
Integration strategy	**Siloed**	**Enterprise-Wide**	Breaks down silos, creating a unified, holistic risk language that connects code to corporate strategy.
Mitigation precision	**Generic Controls**	**Targeted Mitigants**	Moves from passive policy-setting to active technical defense and algorithmic correction.
Auditability & metrics	**Qualitative/Subjective**	**Auditable/Quantifiable**	Generates the evidence artifacts required by regulators (e.g., EU AI Act) and auditors.
Cultural resilience	**Compliance-Driven**	**Resilience-Driven**	Transforms risk from a "Department of No" into an enabler of speed and innovation.
Regulatory agility	**Reactive**	**Proactive/Adaptive**	Future-proofs the organization against a volatile and fragmenting global regulatory landscape.
Socio-Technical scope	**Techno-Centric**	**Human-Centric**	Protects against "Responsible AI" failures that cause massive reputational damage.
Residual Risk focus	**Inherent Risk Bias**	**Residual Risk Target**	Provides the Board with the "true" exposure number after defenses are applied.
Operational tailoring	**Rigid/Standardized**	**Profile-Based/Tailored**	Allows the framework to scale from a small fintech startup to a multinational conglomerate.

Appendix - Overarching common regulatory arrangement on products with embedded AI (CRA-AI)

I. Scope

Products and/or services with embedded digital technologies or an AI system

II. Product and/or service requirements

- Regulatory objectives and the level of risk
- Societal impact
- Digital considerations

III. Reference to International Standards

- **ISO/IEC 42001** series of standards on AI management systems
- **ISO/IEC 23894:2023** series of standards on AI – Guidance on Risk Management
- **ISO/IEC TR 22100-5** series of standards on the implications of AI machine learning
- **IEC 62443** series of standards on industrial automation and control systems
- **IEEE 7001-2021** series of standards for transparency of autonomous systems
- Organisation for Economic Co-operation and Development (OECD) Recommendation of the Council on Artificial Intelligence (**OECD/LEGAL/0449**)
- United Nations Educational, Scientific and Cultural Organization (**UNESCO**) Recommendation on the Ethics of Artificial Intelligence **(SHS/BIO/PI/2021/1)**
- World Health Organization (**WHO**) Ethics and governance of artificial intelligence for health: Guidance on large multi-modal models

IV. Conformity Assessment

- New regulatory approach
- Two different types of compliance (product/service + AI)
- Risk assessment framework

V. Market Surveillance

Methods for continuous compliance throughout the product life cycle. Mandatory independent audits. Product recalls and removal from markets, if necessary.

Appendix - Project Deliverables

UNECE
The regulatory compliance of products with embedded artificial intelligence or other digital technologies
WP.6 Working Party on Regulatory Cooperation and Standardization Policies
Economic Commission for Europe
Steering Committee on Trade Capacity and Standards
Working Party on Regulatory Cooperation and Standardization Policies
Summary
Please recycle

https://unece.org/sites/default/files/2023-12/2023-9-Compliance-AI_Eng.pdf

UNECE WP.6
AI White Paper
(Nov. 2023)

→ **UNECE WP.6**
AI Paper
(Dec 2023)

Newly designated
UNECE WP.6
Key area of work:

‘Digital Regulation of Goods and AI’
(Dec 2023)

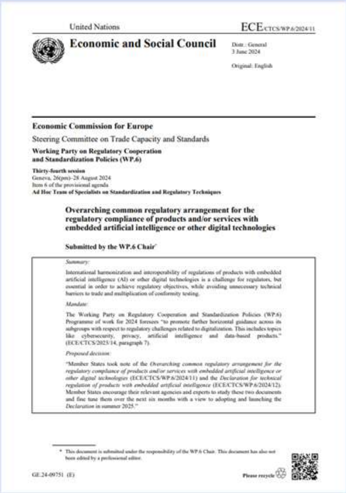
Working Party on Regulatory Cooperation and Standardization Policies (WP.6)

The Working Party on Regulatory Cooperation and Standardization Policies (WP.6) encourages increased regulatory coherence in specific sectors that have a critical impact on sustainable development and promotes greater resilience to natural and man-made hazards. WP.6 also works to:

- Promote the use of standards by policy-makers and business as a tool for reducing technical barriers to trade, promote increased resilience to disasters, foster innovation and good governance
- Advocate the use of standards in the implementation of UN-wide goals, including the implementation of the Agenda 2030 and the Sendai framework for action

Calendar of WP.6 events for 2024

https://unece.org/trade/wp6/digital-regulation-goods-artificial-intelligence



United Nations ECE/CTCS/WP.6/2024/11

Economic and Social Council

Distr.: General
3 June 2024

Original: English

Economic Commission for Europe
Steering Committee on Trade Capacity and Standards
Working Party on Regulatory Cooperation and Standardization Policies (WP.6)
Thirty-fourth session
Geneva, 26(pm)–28 August 2024
Item 6 of the provisional agenda
Ad Hoc Team of Specialists on Standardization and Regulatory Techniques

Overarching common regulatory arrangement for the regulatory compliance of products and/or services with embedded artificial intelligence or other digital technologies

Submitted by the WP.6 Chair*

Summary:

International harmonization and interoperability of regulations of products with embedded artificial intelligence (AI) or other digital technologies is a challenge for regulators, but essential in order to achieve regulatory objectives, while avoiding unnecessary technical barriers to trade and multiplication of conformity testing.

Mandate:

The Working Party on Regulatory Cooperation and Standardization Policies (WP.6) Programme of work for 2024 foresees "to promote further horizontal guidance across its subgroups with respect to regulatory challenges related to digitalization. This includes topics like cybersecurity, privacy, artificial intelligence and data-based products." (ECE/CTCS/2023/14, paragraph 7).

Proposed decision:

"Member States took note of the *Overarching common regulatory arrangement for the regulatory compliance of products and/or services with embedded artificial intelligence or other digital technologies* (ECE/CTCS/WP.6/2024/11) and the *Declaration for technical regulation of products with embedded artificial intelligence* (ECE/CTCS/WP.6/2024/12). Member States encourage their relevant agencies and experts to study these two documents and fine tune them over the next six months with a view to adopting and launching the *Declaration* in summer 2025."

* This document is submitted under the responsibility of the WP.6 Chair. This document has also not been edited by a professional editor.

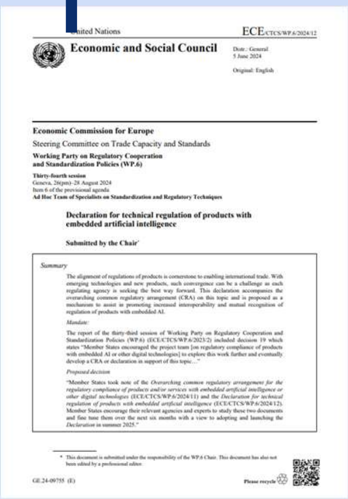
GE.24-09751 (E)

Please recycle

https://unece.org/sites/default/files/2024-07/ECE_CTCS_WP.6_2024_11_E.pdf

Overarching Common Regulatory Arrangement (CRA)
(June 2024)

Declaration for technical regulation of products with embedded artificial intelligence
(June 2024)



United Nations ECE/CTCS/WP.6/2024/12

Economic and Social Council

Distr.: General
5 June 2024

Original: English

Economic Commission for Europe
Steering Committee on Trade Capacity and Standards
Working Party on Regulatory Cooperation and Standardization Policies (WP.6)
Thirty-fourth session
Geneva, 26(pm)–28 August 2024
Item 6 of the provisional agenda
Ad Hoc Team of Specialists on Standardization and Regulatory Techniques

Declaration for technical regulation of products with embedded artificial intelligence

Submitted by the Chair*

Summary

The alignment of regulations of products is cornerstone to enabling international trade. With emerging technologies and new products, such convergence can be a challenge as each regulating agency is seeking the best way forward. This declaration accompanies the overarching common regulatory arrangement (CRA) on this topic and is proposed as a mechanism to assist in promoting increased interoperability and mutual recognition of regulation of products with embedded AI.

Mandate:

The report of the thirty-third session of Working Party on Regulatory Cooperation and Standardization Policies (WP.6) (ECE/CTCS/WP.6/2023/2) included decision 19 which states "Member States encouraged the project team [on regulatory compliance of products with embedded AI or other digital technologies] to explore this work further and eventually develop a CRA or declaration in support of this topic…"

Proposed decision

"Member States took note of the *Overarching common regulatory arrangement for the regulatory compliance of products and/or services with embedded artificial intelligence or other digital technologies* (ECE/CTCS/WP.6/2024/11) and the *Declaration for technical regulation of products with embedded artificial intelligence* (ECE/CTCS/WP.6/2024/12). Member States encourage their relevant agencies and experts to study these two documents and fine tune them over the next six months with a view to adopting and launching the *Declaration* in summer 2025."

* This document is submitted under the responsibility of the WP.6 Chair. This document has also not been edited by a professional editor.

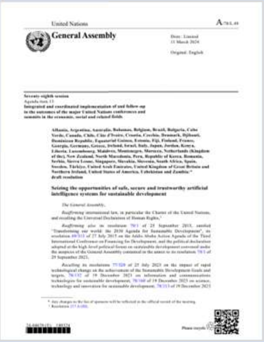
GE.24-09755 (E)

Please recycle

https://unece.org/sites/default/files/2024-07/ECE_CTCS_WP.6_2024_12_E.pdf

complemented by:

UNGA Resolution A/78/L.49
(Mar 2024)



United Nations A/78/L.49

General Assembly

Distr.: Limited
11 March 2024

Original: English

Seventy-eighth session
Agenda item 13
Integrated and coordinated implementation of and follow-up to the outcomes of the major United Nations conferences and summits in the economic, social and related fields

Albania, Argentina, Australia, Bahamas, Belgium, Brazil, Bulgaria, Cabo Verde, Canada, Chile, Côte d'Ivoire, Croatia, Czechia, Denmark, Djibouti, Dominican Republic, Equatorial Guinea, Estonia, Fiji, Finland, France, Georgia, Germany, Greece, Ireland, Israel, Italy, Japan, Jordan, Kenya, Liberia, Luxembourg, Maldives, Montenegro, Morocco, Netherlands (Kingdom of the), New Zealand, North Macedonia, Peru, Republic of Korea, Romania, Serbia, Sierra Leone, Singapore, Slovakia, Slovenia, South Africa, Spain, Sweden, Türkiye, United Arab Emirates, United Kingdom of Great Britain and Northern Ireland, United States of America, Uzbekistan and Zambia:* draft resolution

Seizing the opportunities of safe, secure and trustworthy artificial intelligence systems for sustainable development

The General Assembly,

Reaffirming international law, in particular the Charter of the United Nations, and recalling the Universal Declaration of Human Rights,[1]

[illegible]

Please recycle

https://daccess-ods.un.org/access.nsf/Get?OpenAgent&DS=A/78/L.49&Lang=E

“Seizing the opportunities of safe, secure and trustworthy AI systems for sustainable development”

UNGA Resolution A/78/L.86
(Jul 2024)

“Enhancing international cooperation on capacity-building of artificial intelligence”

United Nations A/78/L.86

General Assembly

Distr.: Limited
25 June 2024

Original: English

Seventy-eighth session
Agenda item 13
Integrated and coordinated implementation of and follow-up to the outcomes of the major United Nations conferences and summits in the economic, social and related fields

Algeria, Burundi, Cambodia, China, Congo, Cuba, Dominica, Egypt, Equatorial Guinea, Eritrea, Gabon, Indonesia, Iran (Islamic Republic of), Iraq, Kazakhstan, Lao People's Democratic Republic, Liberia, Nicaragua, Nigeria, Pakistan, Qatar, Sierra Leone, Syrian Arab Republic, Uganda, United Republic of Tanzania, Venezuela (Bolivarian Republic of), Zambia and Zimbabwe: draft resolution

Enhancing international cooperation on capacity-building of artificial intelligence

The General Assembly,

[illegible]

Please recycle

https://daccess-ods.un.org/access.nsf/Get?OpenAgent&DS=A/78/L.86&Lang=E

Appendix - Proposed next steps

Overarching Common Regulatory Arrangement

- Seek your **approval** as per the Programme of work 2024. (ECE/CTCS/WP.6/2023/14, paragraph 10e)
- Request your input in the development of **product-specific CRAs**

Declaration

- Encourage **relevant national agencies and experts** to study the Overarching CRA and Declaration
- Request **consider adopting the declaration** and **becoming a signatory**
- To formally launch in Summer 2025

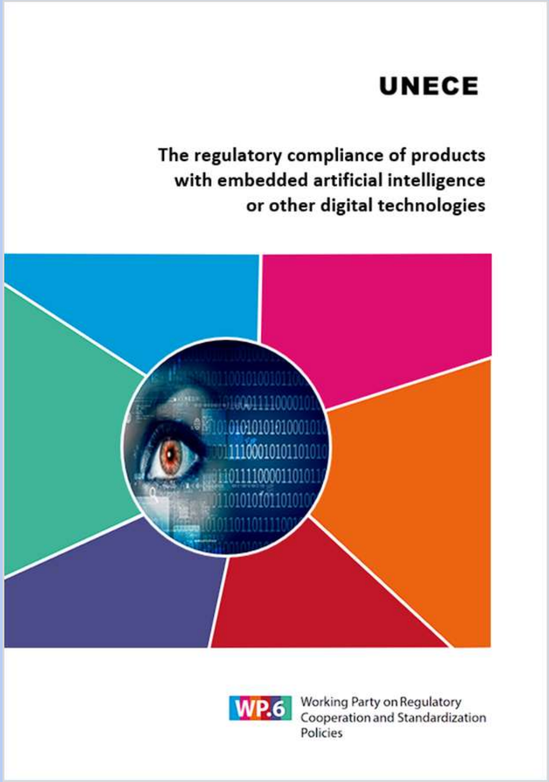
Future possibilities

- **Consider developing a UN convention / treaty** for certain products/services with embedded AI and/or other digital technologies

Appendix 1 – Autonomy & Agency in International Standards

Autonomy		**Agency**	
System autonomy / Supervised autonomy	IEEE Std 7001-2021	Human-in-the-loop Control points: Agency and Autonomy	ISO/IEC TR 24028:2020
Autonomy Levels for Unmanned Systems (ALFUS)	NIST SP 1011-II-1.0, 2007	Human agency and oversight	OECD/LEGAL/0449
Systems acting as “automata in process”	IEEE Std 1872-2015	Autonomy, agency, worth and dignity	SHS/BIO/PI/2021/1
Six levels of autonomy (0-5, for driverless cars)	SAE J3016_201806		
Levels of automation	ISO/IEC 42001 & ISO/IEC 23894		
Relationship between autonomy, heteronomy and automation (incl. a tabular representation)	ISO/IEC TR 5469 & ISO/IEC 22989		
(Degree / Level of) Autonomy	ISO/IEC TR 24028:2020 & OECD/LEGAL/0449		
Various references to autonomy	WHO – Guidance on Large Multi-Modal Models		

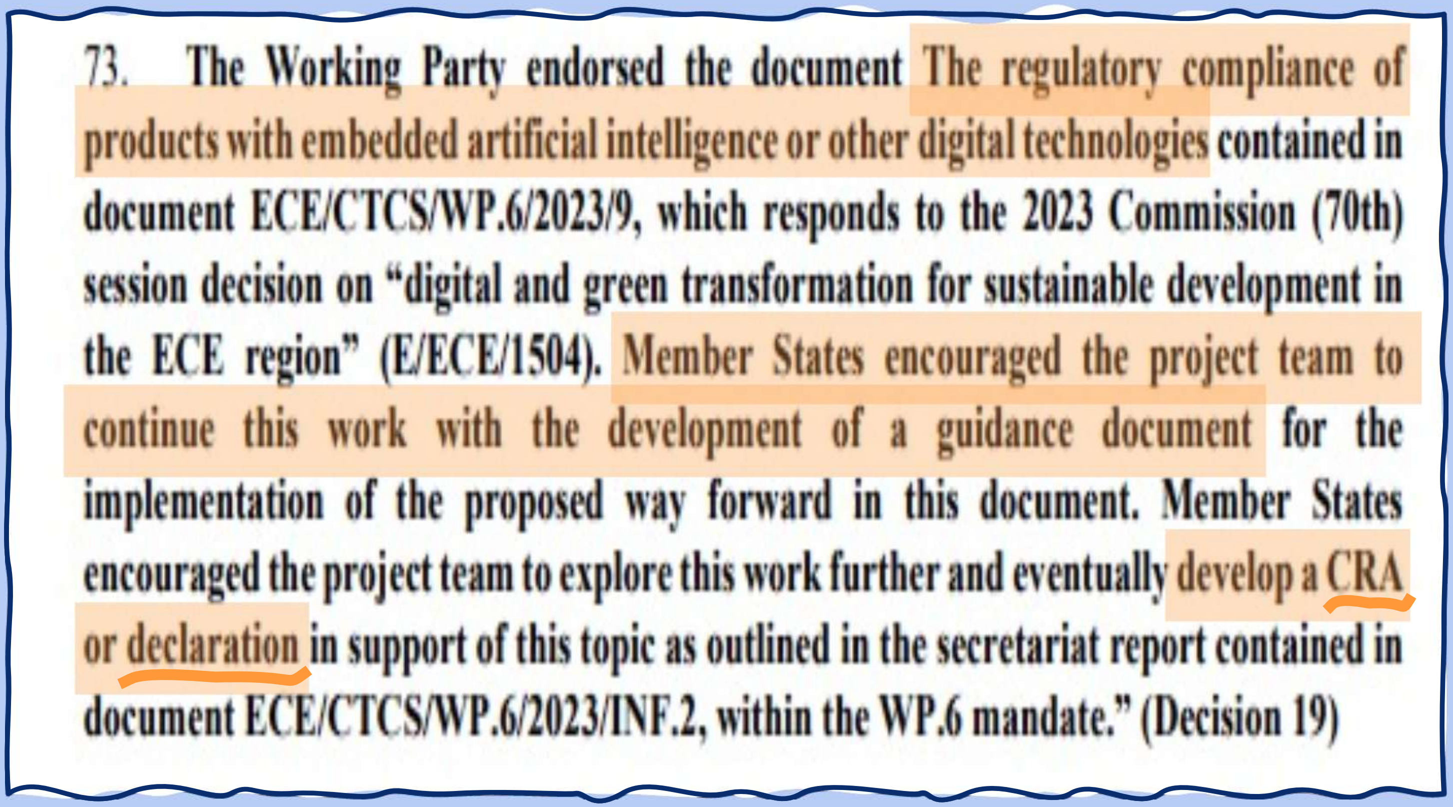
Appendix – Last year's project brief (33rd Annual session)



UNECE

The regulatory compliance of products with embedded artificial intelligence or other digital technologies

WP.6 Working Party on Regulatory Cooperation and Standardization Policies

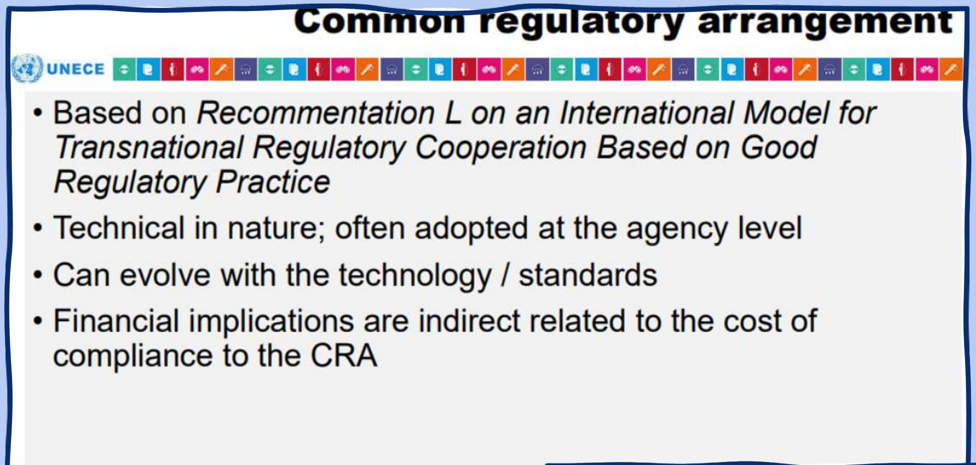


73. The Working Party endorsed the document The regulatory compliance of products with embedded artificial intelligence or other digital technologies contained in document ECE/CTCS/WP.6/2023/9, which responds to the 2023 Commission (70th) session decision on "digital and green transformation for sustainable development in the ECE region" (E/ECE/1504). Member States encouraged the project team to continue this work with the development of a guidance document for the implementation of the proposed way forward in this document. Member States encouraged the project team to explore this work further and eventually develop a CRA or declaration in support of this topic as outlined in the secretariat report contained in document ECE/CTCS/WP.6/2023/INF.2, within the WP.6 mandate." (Decision 19)

Source:

Report on the thirty-third session of the Working Party on Regulatory Cooperation and Standardization Policies ECE/CTCS/WP.6/2023/2
https://unece.org/sites/default/files/2023-12/ECE_CTCS_WP.6_2023_02_E.pdf

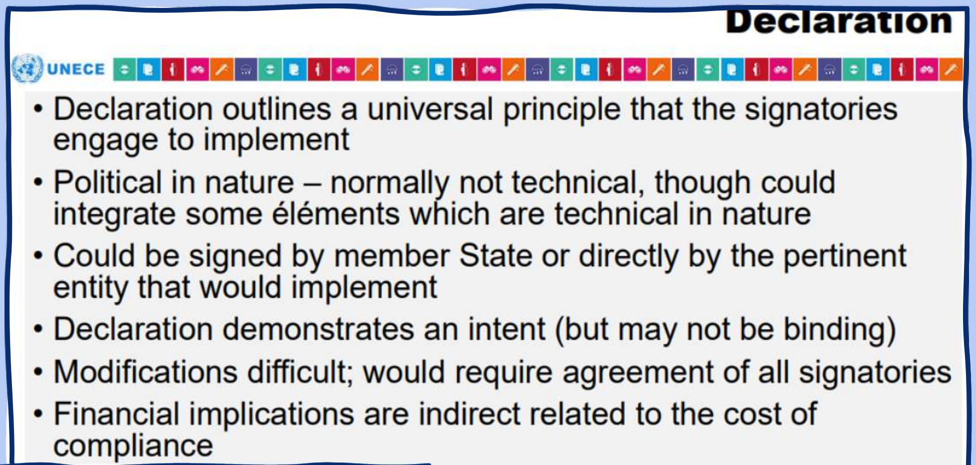
Appendix – Instrument features comparison



Common regulatory arrangement

UNECE

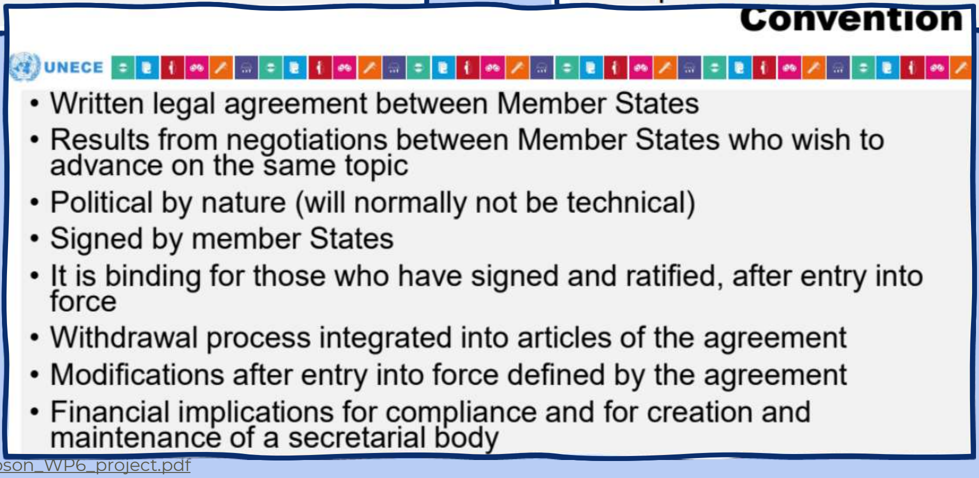
- Based on *Recommentation L on an International Model for Transnational Regulatory Cooperation Based on Good Regulatory Practice*
- Technical in nature; often adopted at the agency level
- Can evolve with the technology / standards
- Financial implications are indirect related to the cost of compliance to the CRA



Declaration

UNECE

- Declaration outlines a universal principle that the signatories engage to implement
- Political in nature – normally not technical, though could integrate some éléments which are technical in nature
- Could be signed by member State or directly by the pertinent entity that would implement
- Declaration demonstrates an intent (but may not be binding)
- Modifications difficult; would require agreement of all signatories
- Financial implications are indirect related to the cost of compliance



Convention

UNECE

- Written legal agreement between Member States
- Results from negotiations between Member States who wish to advance on the same topic
- Political by nature (will normally not be technical)
- Signed by member States
- It is binding for those who have signed and ratified, after entry into force
- Withdrawal process integrated into articles of the agreement
- Modifications after entry into force defined by the agreement
- Financial implications for compliance and for creation and maintenance of a secretarial body

Source: https://unece.org/sites/default/files/2023-12/LThompson_WP6_project.pdf

Appendix – Existing UNECE CRAs to date

Title / Symbol	Date	ENG	FRA	RUS
ECE/CTCS/WP.6/2024/11 Overarching common regulatory arrangement for the **regulatory compliance of products and/or services with embedded artificial intelligence or other digital technologies**	2024	PDF	PDF	PDF
ECE/TRADE/391/Rev.1 Common Regulatory Framework for **Equipment Used In Environments with an Explosive Atmosphere**	2022	HTML		
Common Regulatory Arrangements on **Cybersecurity**	2019	PDF	PDF	PDF
Common Regulatory Objectives on **Earth-moving Machinery**	2009	PDF		
Common Regulatory Objectives (Telecom) for **ICT Equipment, for Bluetooth Equipment, for GSM Equipment, for IMT-2000 Equipment, for PC Equipment, for PSTN Equipment and for WLAN Equipment**	2004	PDF		

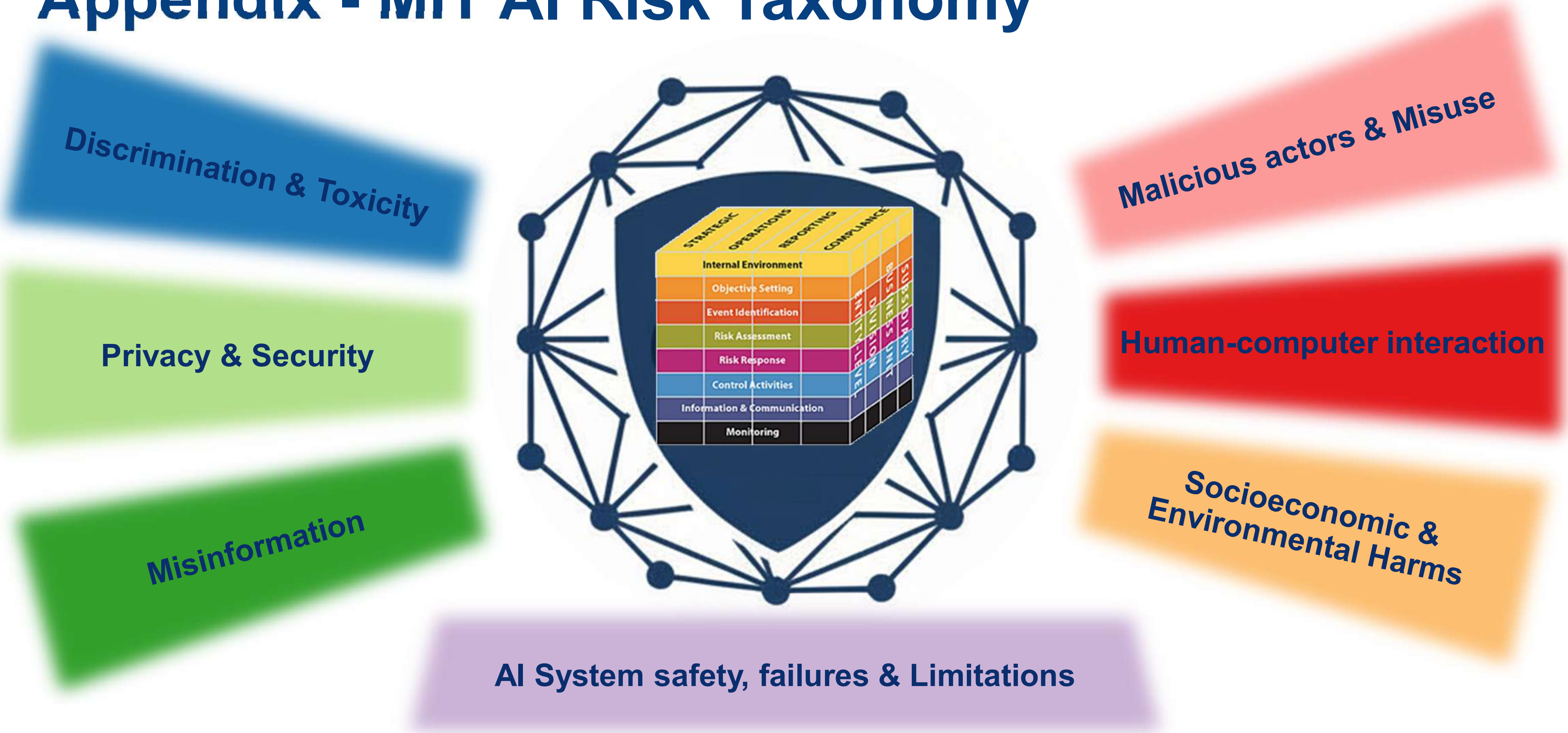
Source: https://unece.org/trade/wp6/regulatory-cooperation#accordion_1

Appendix – Existing WP.6 Declarations

Title / Symbol	Date	ENG	FRA	RUS
Declaration on Gender-Responsive Standards and Standards Development, notably: • There are **86 signatories** from **every region of the world** - **mostly standards development organizations (SDOs)** – national, regional, international as well as some VSS. • The **signatories are de-facto members** of the Team of Specialists on Gender-Responsive Standards • There is a **follow-up to this declaration through the ToS-GRS** long after the declaration has been finalized.	2019	PDF	PDF	PDF

Source: https://unece.org/trade/wp6/Gender-Resp%20-Stdards-declaration

Appendix - MIT AI Risk Taxonomy



Discrimination & Toxicity

Privacy & Security

Misinformation

Malicious actors & Misuse

Human-computer interaction

Socioeconomic & Environmental Harms

AI System safety, failures & Limitations

Appendix - Discrimination & Toxicity

Unfair discrimination and misrepresentation	Unequal treatment of individuals or groups by AI, often based on race, gender, or other sensitive characteristics, resulting in unfair outcomes and unfair representation of those groups.
Exposure to toxic Content	AI that exposes users to harmful, abusive, unsafe or inappropriate content. May involve providing advice or encouraging action. Examples of toxic content include hate speech, violence, extremism, illegal acts, or child sexual abuse material, as well as content that violates community norms such as profanity, inflammatory political speech, or pornography.
Unequal performance across groups	Accuracy and effectiveness of AI decisions and actions is dependent on group membership, where decisions in AI system design and biased training data lead to unequal outcomes, reduced benefits, increased effort, and alienation of users.

Appendix - Privacy & Security

Compromise of privacy by obtaining, leaking, or correctly inferring sensitive information	AI systems that memorize and leak sensitive personal data or infer private information about individuals without their consent. Unexpected or unauthorized sharing of data and information can compromise user expectation of privacy, assist identity theft, or cause loss of confidential intellectual property
AI system security vulnerabilities and attacks	AI that exposes users to harmful, abusive, unsafe or inappropriate content. May involve providing advice or encouraging action. Examples of toxic content include hate speech, violence, extremism, illegal acts, or child sexual abuse material, as well as content that violates community norms such as profanity, inflammatory political speech, or pornography.

Appendix - Misinformation

False or misleading information	AI systems that inadvertently generate or spread incorrect or deceptive information, which can lead to inaccurate beliefs in users and undermine their autonomy. Humans that make decisions based on false beliefs can experience physical, emotional, or material harms
Pollution of information ecosystem and loss of consensus reality	Highly personalized AI-generated misinformation that creates "filter bubbles" where individuals only see what matches their existing beliefs, undermining shared reality and weakening social cohesion and political processes

Appendix - Malicious actors & Misuse

Disinformation, surveillance, and influence at scale	Using AI systems to conduct large-scale disinformation campaigns, malicious surveillance, or targeted and sophisticated automated censorship and propaganda, with the aim of manipulating political processes, public opinion, and behaviours.
Fraud, scams, and targeted manipulation	Using AI systems to gain a personal advantage over others such as through cheating, fraud, scams, blackmail, or targeted manipulation of beliefs or behaviour. Examples include AI-facilitated plagiarism for research or education, impersonating a trusted or fake individual for illegitimate financial benefit, or creating humiliating or sexual imagery.
Cyberattacks, weapon development or use, and mass harm	Using AI systems to develop cyber weapons (e.g., by coding cheaper, more effective malware), develop new or enhance existing weapons (e.g., Lethal Autonomous Weapons or chemical, biological, radiological, nuclear, and high-yield explosives), or use weapons to cause mass harm

Appendix - Human-computer interaction

Overreliance and unsafe use	Anthropomorphizing, trusting, or relying on AI systems by users, leading to emotional or material dependence and to inappropriate relationships with or expectations of AI systems. Trust can be exploited by malicious actors (e.g., to harvest information or enable manipulation), or result in harm from inappropriate use of AI in critical situations (such as a medical emergency). Over reliance on AI systems can compromise autonomy and weaken social ties
Loss of human agency and autonomy	Delegating by humans of key decisions to AI systems, or AI systems that make decisions that diminish human control and autonomy, potentially leading to humans feeling disempowered, losing the ability to shape a fulfilling life trajectory, or becoming cognitively enfeebled

Appendix - Socioeconomic & Environmental Harms (1/2)

Power centralization and unfair distribution of benefits	AI-driven concentration of power and resources within certain entities or groups, especially those with access to or ownership of powerful AI systems, leading to inequitable distribution of benefits and increased societal inequality
Exposure to toxic Content	Social and economic inequalities caused by widespread use of AI, such as by automating jobs, reducing the quality of employment, or producing exploitative dependencies between workers and their employers.
Economic and cultural devaluation of human effort	AI systems capable of creating economic or cultural value, including through reproduction of human innovation or creativity (e.g., art, music, writing, coding, invention), destabilizing economic and social systems that rely on human effort. The ubiquity of AI-generated content may lead to reduced appreciation for human skills, disruption of creative and knowledge-based industries, and homogenization of cultural experiences

Appendix - Socioeconomic & Environmental Harms (2/2)

Competitive dynamics	Competition by AI developers or state-like actors in an AI "race" by rapidly developing, deploying, and applying AI systems to maximize strategic or economic advantage, increasing the risk they release unsafe and error-prone systems..
Governance failure	Inadequate regulatory frameworks and oversight mechanisms that fail to keep pace with AI development, leading to ineffective governance and the inability to manage AI risks appropriately.
Environmental harm	The development and operation of AI systems that cause environmental harm, such as through energy consumption of data centers or the materials and carbon footprints associated with AI hardware.

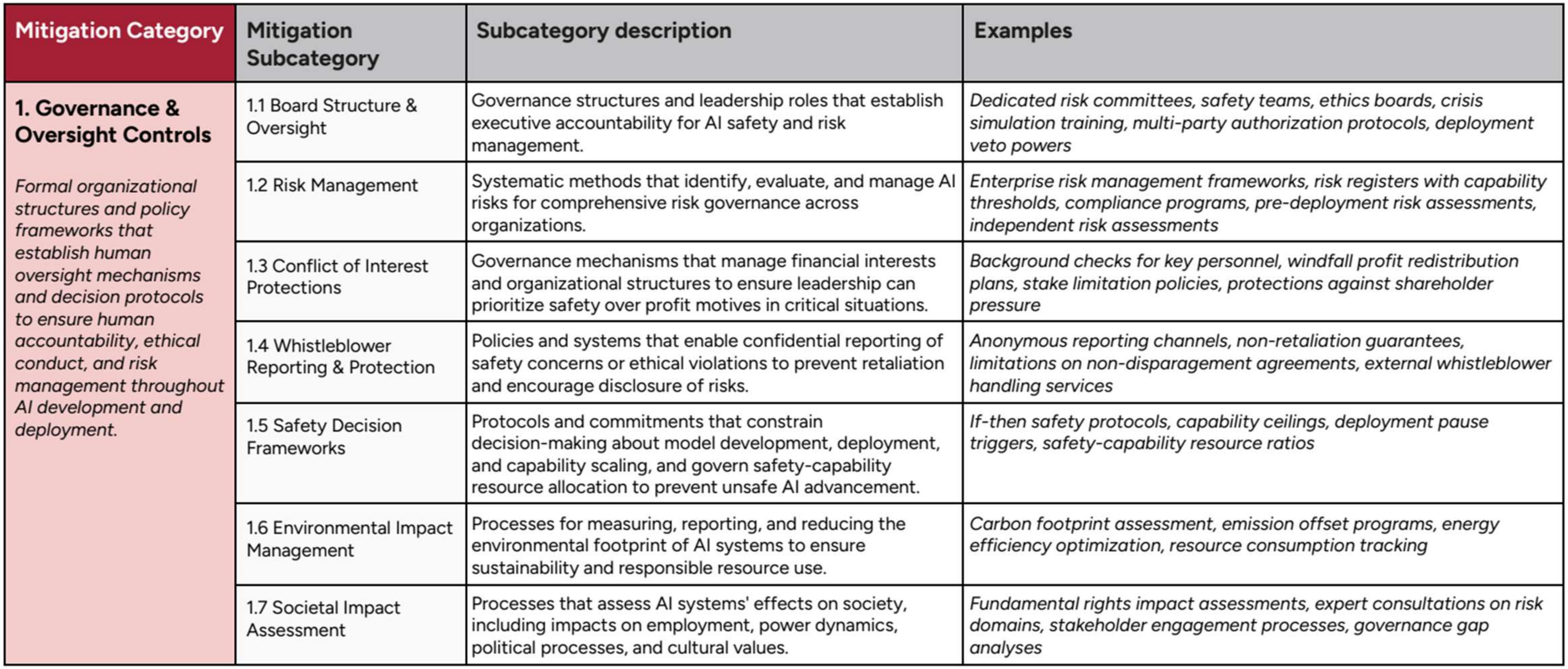
Appendix - AI System safety, failures & Limitations (1/2)

AI pursuing its own goals in conflict with human goals or values	AI systems that act in conflict with ethical standards or human goals or values, especially the goals of designers or users. These misaligned behaviours may be introduced by humans during design and development, such as through reward hacking and goal misgeneralisation, and may result in AI using dangerous capabilities such as manipulation, deception, or situational awareness to seek power, self-proliferate, or achieve other goals.
AI possessing dangerous capabilities	AI systems that develop, access, or are provided with capabilities that increase their potential to cause mass harm through deception, weapons development and acquisition, persuasion and manipulation, political strategy, cyber-offense, AI development, situational awareness, and self-proliferation. These capabilities may cause mass harm due to malicious human actors, misaligned AI systems, or failure in the AI system
Lack of capability or robustness	AI systems that fail to perform reliably or effectively under varying conditions, exposing them to errors and failures that can have significant consequences, especially in critical applications or areas that require moral reasoning

Appendix - AI System safety, failures & Limitations (2/2)

Lack of transparency or interpretability	Challenges in understanding or explaining the decision-making processes of AI systems, which can lead to mistrust, difficulty in enforcing compliance standards or holding relevant actors accountable for harms, and the inability in identify and correct errors
AI welfare and rights	Ethical considerations regarding the treatment of potentially sentient AI entities, including discussions around their potential rights and welfare, particularly as AI systems become more advanced and autonomous.
Multi-agent risks	Risks from multi-agent interactions due to incentives (which can lead to conflict or collusion) and/or the structure of multi-agent systems, which can create cascading failures, selection pressures, new security vulnerabilities, and a lack of shared information and trust.

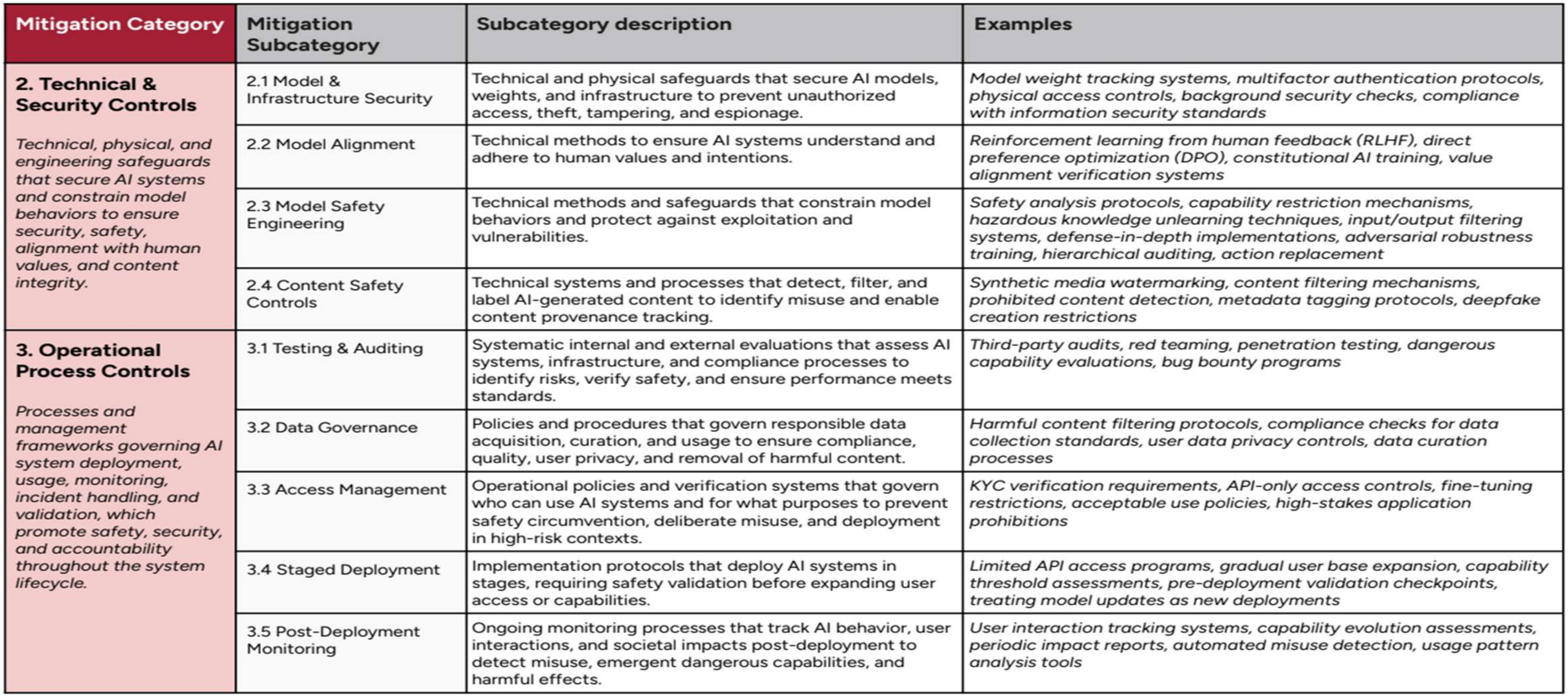
Appendix – Risk Mitigation Taxonomy (1/3)



Mitigation Category	Mitigation Subcategory	Subcategory description	Examples
1. Governance & Oversight Controls *Formal organizational structures and policy frameworks that establish human oversight mechanisms and decision protocols to ensure human accountability, ethical conduct, and risk management throughout AI development and deployment.*	1.1 Board Structure & Oversight	Governance structures and leadership roles that establish executive accountability for AI safety and risk management.	*Dedicated risk committees, safety teams, ethics boards, crisis simulation training, multi-party authorization protocols, deployment veto powers*
	1.2 Risk Management	Systematic methods that identify, evaluate, and manage AI risks for comprehensive risk governance across organizations.	*Enterprise risk management frameworks, risk registers with capability thresholds, compliance programs, pre-deployment risk assessments, independent risk assessments*
	1.3 Conflict of Interest Protections	Governance mechanisms that manage financial interests and organizational structures to ensure leadership can prioritize safety over profit motives in critical situations.	*Background checks for key personnel, windfall profit redistribution plans, stake limitation policies, protections against shareholder pressure*
	1.4 Whistleblower Reporting & Protection	Policies and systems that enable confidential reporting of safety concerns or ethical violations to prevent retaliation and encourage disclosure of risks.	*Anonymous reporting channels, non-retaliation guarantees, limitations on non-disparagement agreements, external whistleblower handling services*
	1.5 Safety Decision Frameworks	Protocols and commitments that constrain decision-making about model development, deployment, and capability scaling, and govern safety-capability resource allocation to prevent unsafe AI advancement.	*If-then safety protocols, capability ceilings, deployment pause triggers, safety-capability resource ratios*
	1.6 Environmental Impact Management	Processes for measuring, reporting, and reducing the environmental footprint of AI systems to ensure sustainability and responsible resource use.	*Carbon footprint assessment, emission offset programs, energy efficiency optimization, resource consumption tracking*
	1.7 Societal Impact Assessment	Processes that assess AI systems' effects on society, including impacts on employment, power dynamics, political processes, and cultural values.	*Fundamental rights impact assessments, expert consultations on risk domains, stakeholder engagement processes, governance gap analyses*

Source: https://airisk.mit.edu/blog/mapping-ai-risk-mitigations

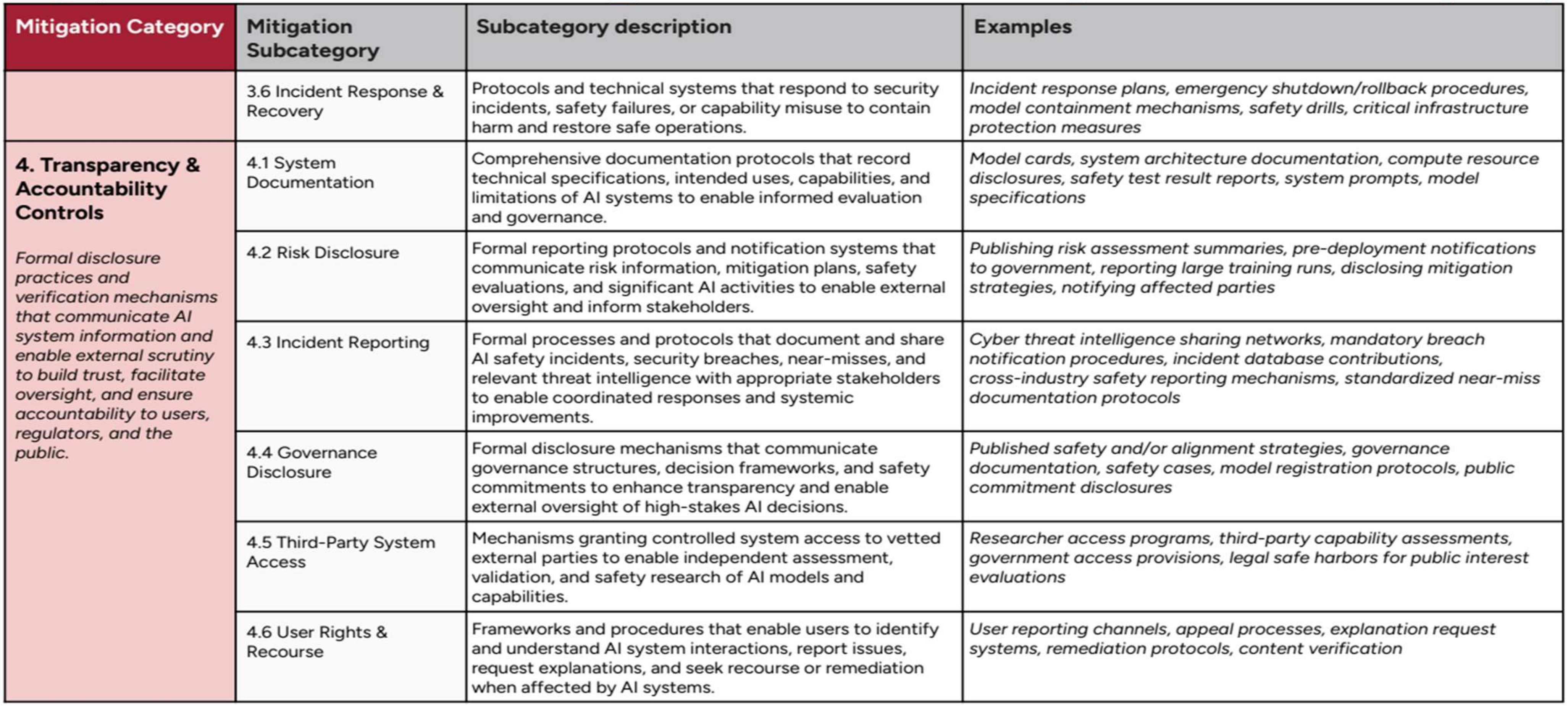
Appendix – Risk Mitigation Taxonomy (2/3)



Mitigation Category	Mitigation Subcategory	Subcategory description	Examples
2. Technical & Security Controls *Technical, physical, and engineering safeguards that secure AI systems and constrain model behaviors to ensure security, safety, alignment with human values, and content integrity.*	2.1 Model & Infrastructure Security	Technical and physical safeguards that secure AI models, weights, and infrastructure to prevent unauthorized access, theft, tampering, and espionage.	*Model weight tracking systems, multifactor authentication protocols, physical access controls, background security checks, compliance with information security standards*
	2.2 Model Alignment	Technical methods to ensure AI systems understand and adhere to human values and intentions.	*Reinforcement learning from human feedback (RLHF), direct preference optimization (DPO), constitutional AI training, value alignment verification systems*
	2.3 Model Safety Engineering	Technical methods and safeguards that constrain model behaviors and protect against exploitation and vulnerabilities.	*Safety analysis protocols, capability restriction mechanisms, hazardous knowledge unlearning techniques, input/output filtering systems, defense-in-depth implementations, adversarial robustness training, hierarchical auditing, action replacement*
	2.4 Content Safety Controls	Technical systems and processes that detect, filter, and label AI-generated content to identify misuse and enable content provenance tracking.	*Synthetic media watermarking, content filtering mechanisms, prohibited content detection, metadata tagging protocols, deepfake creation restrictions*
3. Operational Process Controls *Processes and management frameworks governing AI system deployment, usage, monitoring, incident handling, and validation, which promote safety, security, and accountability throughout the system lifecycle.*	3.1 Testing & Auditing	Systematic internal and external evaluations that assess AI systems, infrastructure, and compliance processes to identify risks, verify safety, and ensure performance meets standards.	*Third-party audits, red teaming, penetration testing, dangerous capability evaluations, bug bounty programs*
	3.2 Data Governance	Policies and procedures that govern responsible data acquisition, curation, and usage to ensure compliance, quality, user privacy, and removal of harmful content.	*Harmful content filtering protocols, compliance checks for data collection standards, user data privacy controls, data curation processes*
	3.3 Access Management	Operational policies and verification systems that govern who can use AI systems and for what purposes to prevent safety circumvention, deliberate misuse, and deployment in high-risk contexts.	*KYC verification requirements, API-only access controls, fine-tuning restrictions, acceptable use policies, high-stakes application prohibitions*
	3.4 Staged Deployment	Implementation protocols that deploy AI systems in stages, requiring safety validation before expanding user access or capabilities.	*Limited API access programs, gradual user base expansion, capability threshold assessments, pre-deployment validation checkpoints, treating model updates as new deployments*
	3.5 Post-Deployment Monitoring	Ongoing monitoring processes that track AI behavior, user interactions, and societal impacts post-deployment to detect misuse, emergent dangerous capabilities, and harmful effects.	*User interaction tracking systems, capability evolution assessments, periodic impact reports, automated misuse detection, usage pattern analysis tools*

Source: https://airisk.mit.edu/blog/mapping-ai-risk-mitigations

Appendix – Risk Mitigation Taxonomy (3/3)



Mitigation Category	Mitigation Subcategory	Subcategory description	Examples
	3.6 Incident Response & Recovery	Protocols and technical systems that respond to security incidents, safety failures, or capability misuse to contain harm and restore safe operations.	*Incident response plans, emergency shutdown/rollback procedures, model containment mechanisms, safety drills, critical infrastructure protection measures*
4. Transparency & Accountability Controls *Formal disclosure practices and verification mechanisms that communicate AI system information and enable external scrutiny to build trust, facilitate oversight, and ensure accountability to users, regulators, and the public.*	4.1 System Documentation	Comprehensive documentation protocols that record technical specifications, intended uses, capabilities, and limitations of AI systems to enable informed evaluation and governance.	*Model cards, system architecture documentation, compute resource disclosures, safety test result reports, system prompts, model specifications*
	4.2 Risk Disclosure	Formal reporting protocols and notification systems that communicate risk information, mitigation plans, safety evaluations, and significant AI activities to enable external oversight and inform stakeholders.	*Publishing risk assessment summaries, pre-deployment notifications to government, reporting large training runs, disclosing mitigation strategies, notifying affected parties*
	4.3 Incident Reporting	Formal processes and protocols that document and share AI safety incidents, security breaches, near-misses, and relevant threat intelligence with appropriate stakeholders to enable coordinated responses and systemic improvements.	*Cyber threat intelligence sharing networks, mandatory breach notification procedures, incident database contributions, cross-industry safety reporting mechanisms, standardized near-miss documentation protocols*
	4.4 Governance Disclosure	Formal disclosure mechanisms that communicate governance structures, decision frameworks, and safety commitments to enhance transparency and enable external oversight of high-stakes AI decisions.	*Published safety and/or alignment strategies, governance documentation, safety cases, model registration protocols, public commitment disclosures*
	4.5 Third-Party System Access	Mechanisms granting controlled system access to vetted external parties to enable independent assessment, validation, and safety research of AI models and capabilities.	*Researcher access programs, third-party capability assessments, government access provisions, legal safe harbors for public interest evaluations*
	4.6 User Rights & Recourse	Frameworks and procedures that enable users to identify and understand AI system interactions, report issues, request explanations, and seek recourse or remediation when affected by AI systems.	*User reporting channels, appeal processes, explanation request systems, remediation protocols, content verification*

Source: https://airisk.mit.edu/blog/mapping-ai-risk-mitigations