AI Cards: Towards an Applied Framework for Machine-Readable AI and Risk Documentation Inspired by the EU AI Act

Delaram Golpayegani*, Isabelle Hupont^, Cecilia Panigutti^
Harshvardhan J. Pandit**, Sven Schade^, Declan O'Sullivan*, Dave Lewis*

* ADAPT Centre, Trinity College Dublin, Dublin, Ireland
^European Commission, Joint Research Centre (JRC), Ispra, Italy
**ADAPT Centre, Dublin City University, Dublin, Ireland

Annual Privacy Forum 2024
4 September 2024

A World Leading SFI Research Centre

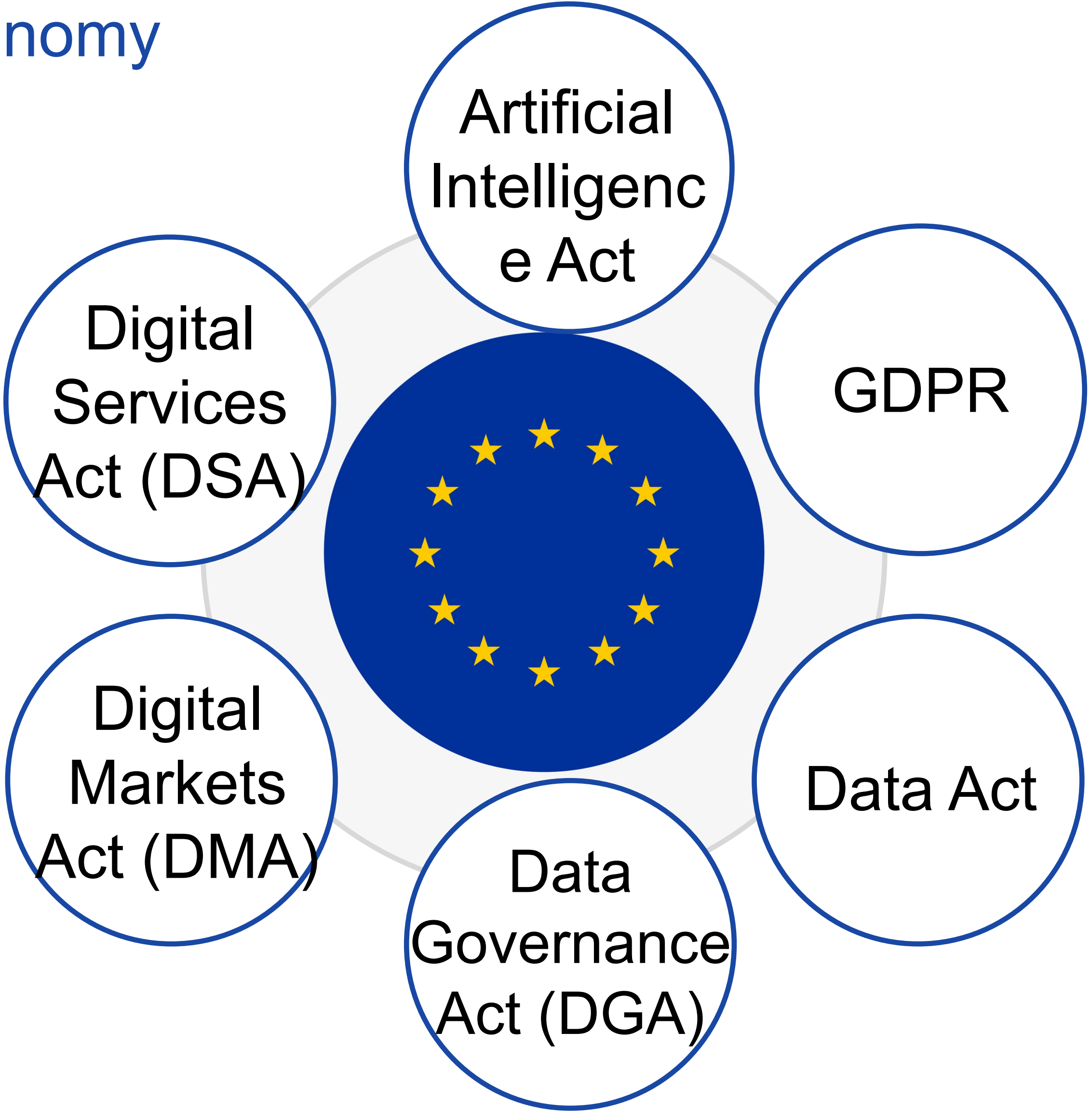
HOST INSTITUTION

HOST INSTITUTION

PARTNER INSTITUTIONS

Link to the paper: https://doi.org/10.1007/978-3-031-68024-3_3

The Big 5+1 EU Digital Regulations
for Data and AI Economy


Artificial Intelligenc e Act
GDPR
Data Act
Data Governance Act (DGA)
Digital Markets Act (DMA)
Digital Services Act (DSA)

The EU AI Act

Engaging Content
Engaging People

New Rules for

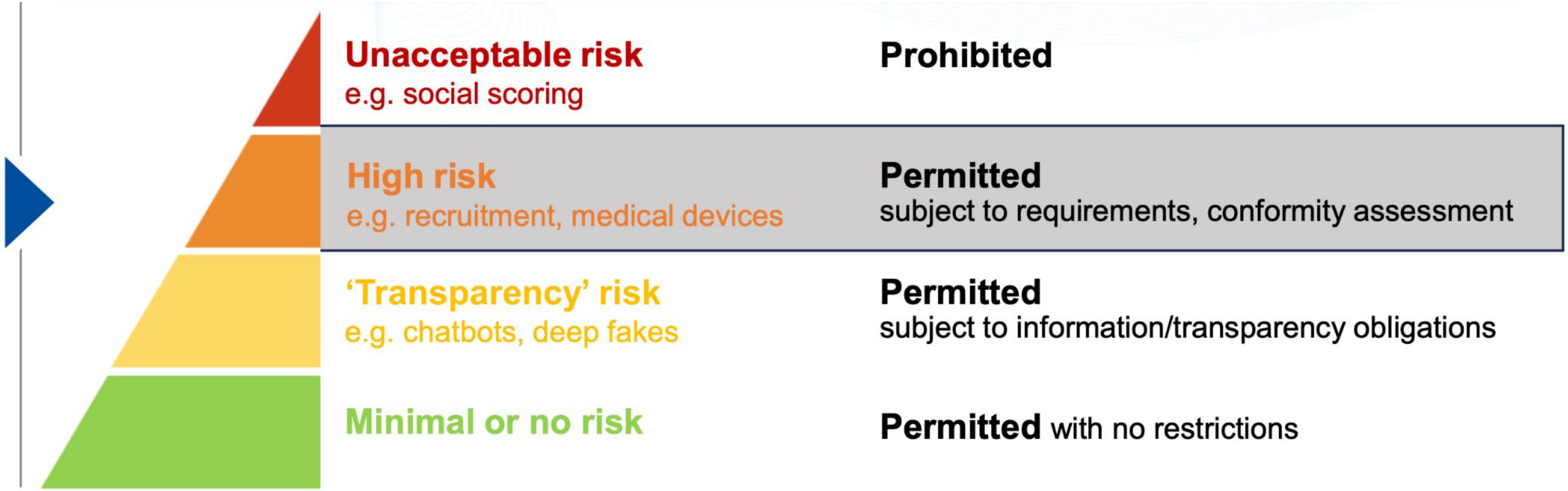
- **AI Systems**
- **GPAI Models** [General Purpose AI]

Promotes human-centric & trustworthy AI

Protects against harmful effects of AI on

- **Health**
- **Safety**
- **Fundamental Rights**

AI Systems Risk-Based Classification



Unacceptable risk
e.g. social scoring

Prohibited

High risk
e.g. recruitment, medical devices

Permitted
subject to requirements, conformity assessment

'Transparency' risk
e.g. chatbots, deep fakes

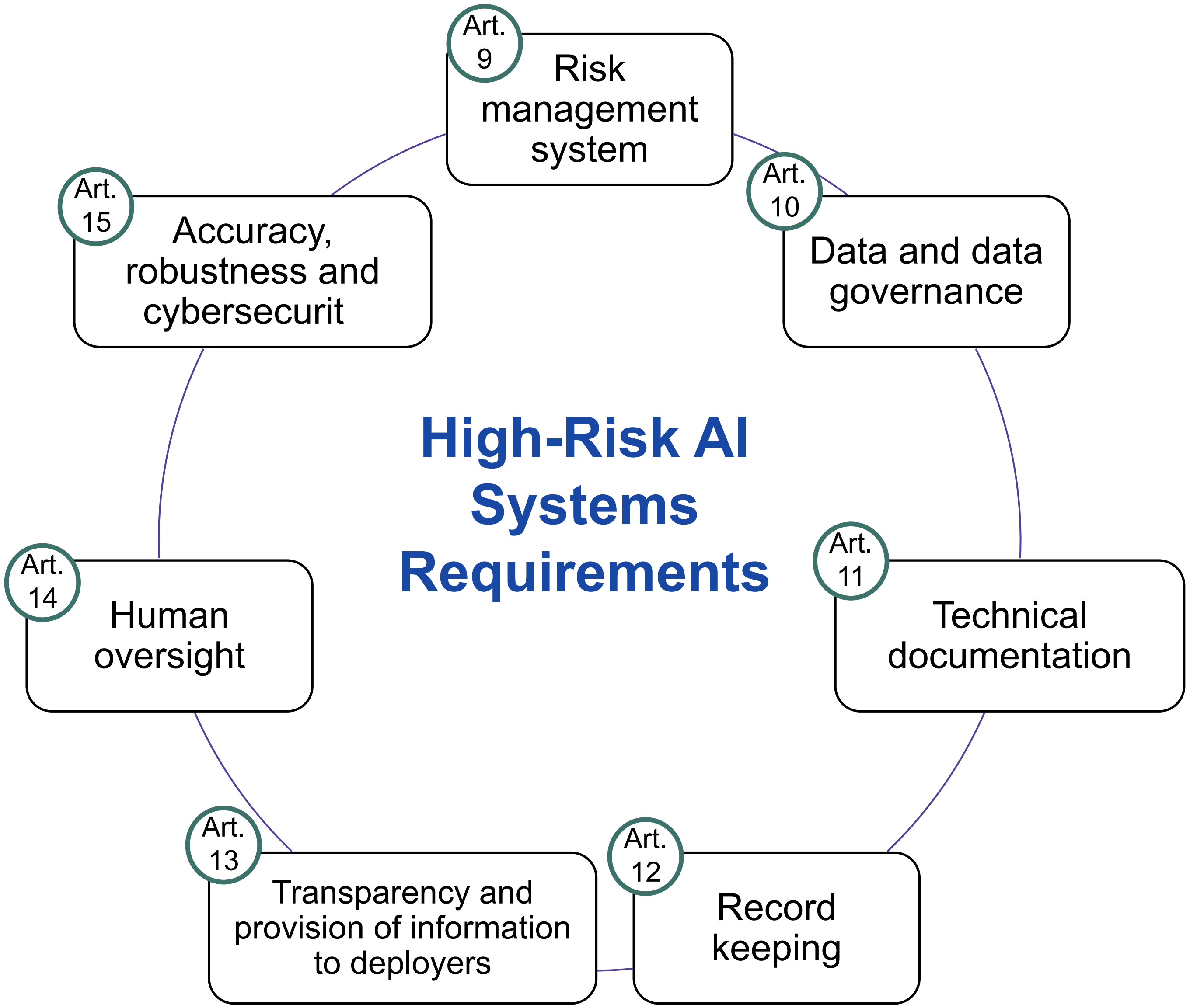
Permitted
subject to information/transparency obligations

Minimal or no risk

Permitted with no restrictions

From the EU AI Office webinar on risk management in the AI Act and related standards, 30 May 2024

High-Risk AI Systems Requirements



Art. 9 Risk management system

Art. 10 Data and data governance

Art. 11 Technical documentation

Art. 12 Record keeping

Art. 13 Transparency and provision of information to deployers

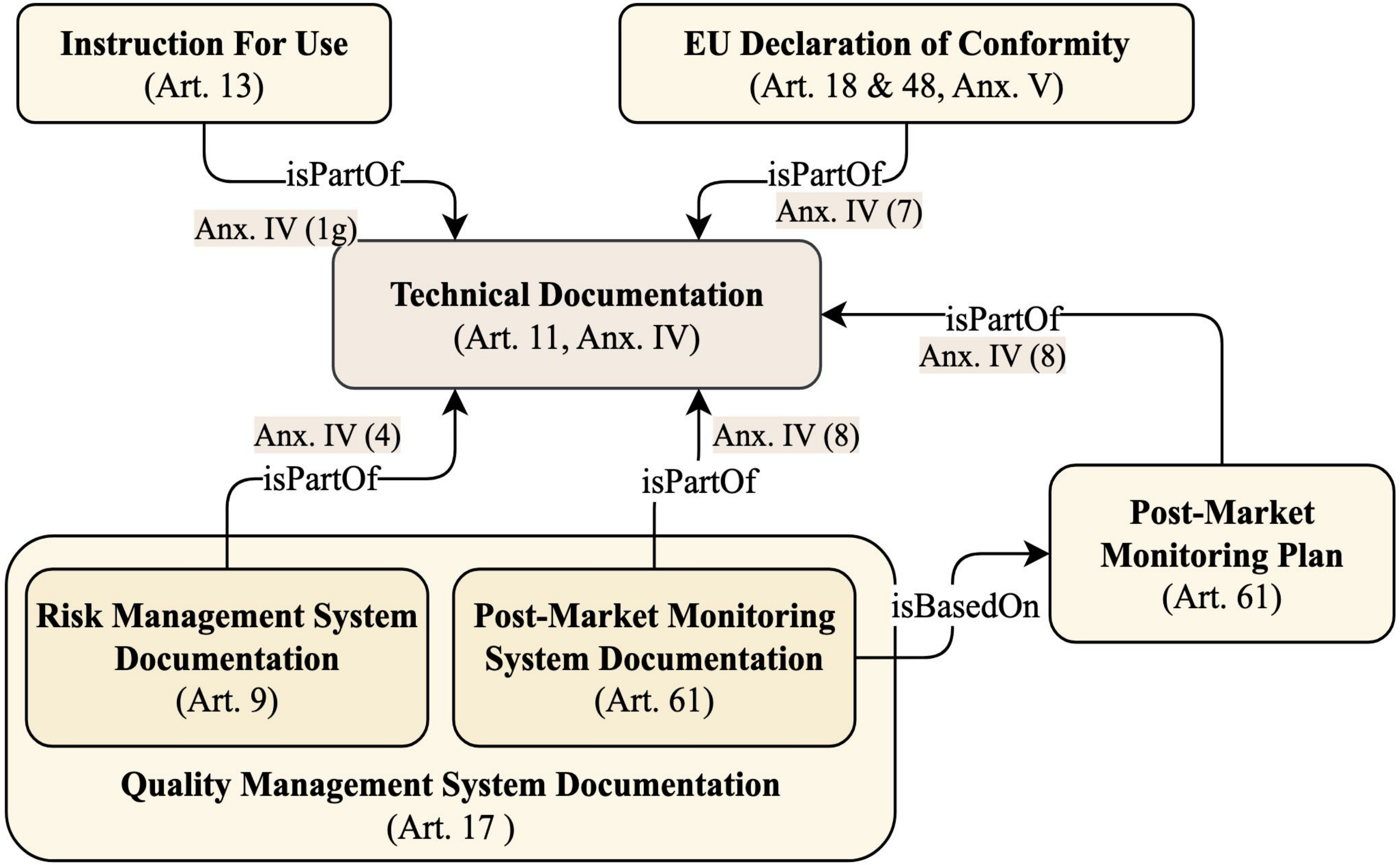
Art. 14 Human oversight

Art. 15 Accuracy, robustness and cybersecurit

Documentation Requirements

- Documentation → Transparency→ Trustworthy AI

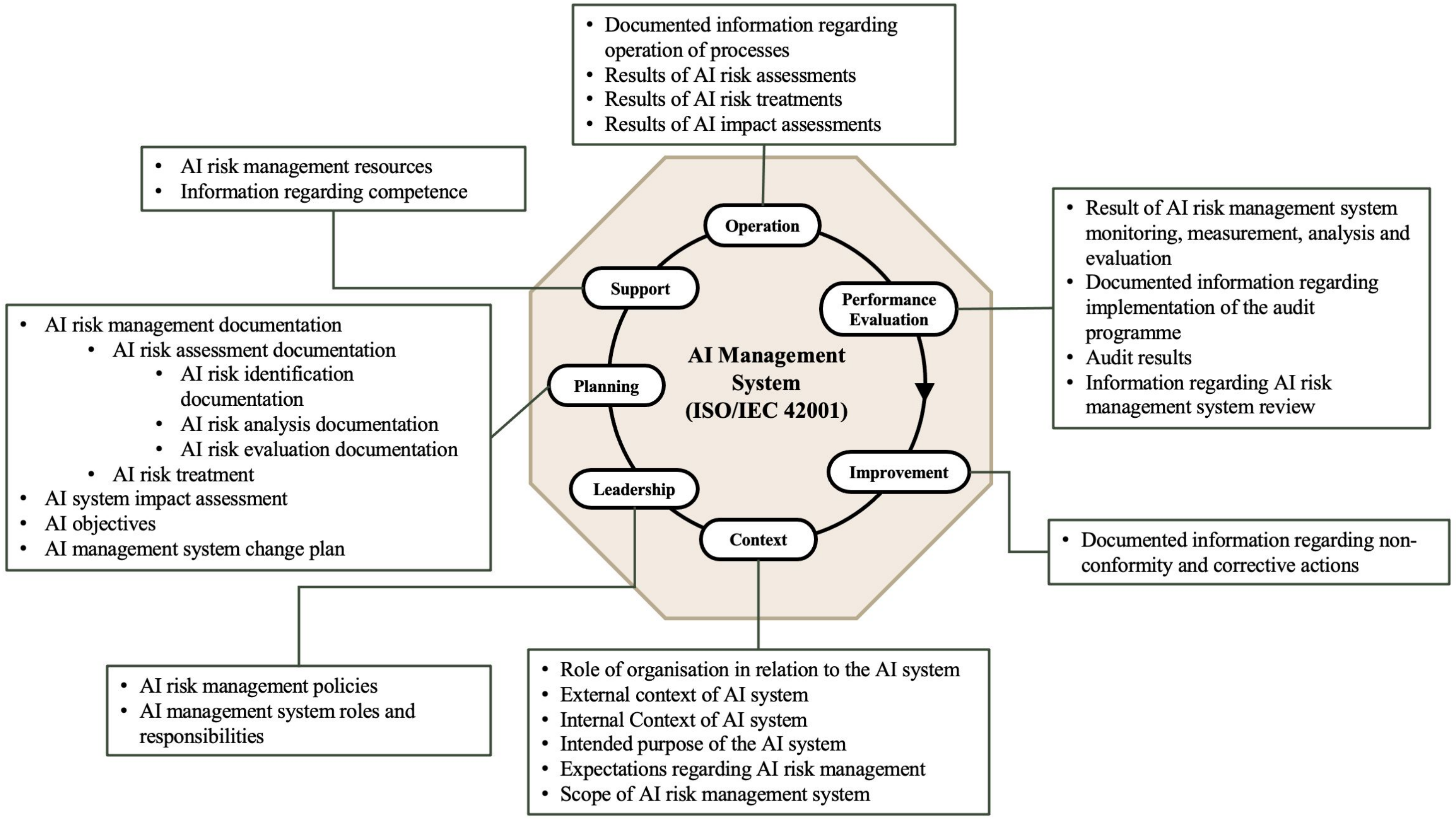
- “The technical documentation shall be drawn up in such a way as to demonstrate that the high-risk AI system complies with the requirements” (Art. 11)



Risk Management System Documentation - available today

ISO/IEC 42001 on AI management system
ISO/IEC 23894 on AI risk management

Note that these standards are not sufficient for compliance with the AI Act



- Documented information regarding operation of processes
- Results of AI risk assessments
- Results of AI risk treatments
- Results of AI impact assessments

- AI risk management resources
- Information regarding competence

- Result of AI risk management system monitoring, measurement, analysis and evaluation
- Documented information regarding implementation of the audit programme
- Audit results
- Information regarding AI risk management system review

- AI risk management documentation
    - AI risk assessment documentation
        - AI risk identification documentation
        - AI risk analysis documentation
        - AI risk evaluation documentation
    - AI risk treatment
- AI system impact assessment
- AI objectives
- AI management system change plan

- Documented information regarding non-conformity and corrective actions

- AI risk management policies
- AI management system roles and responsibilities

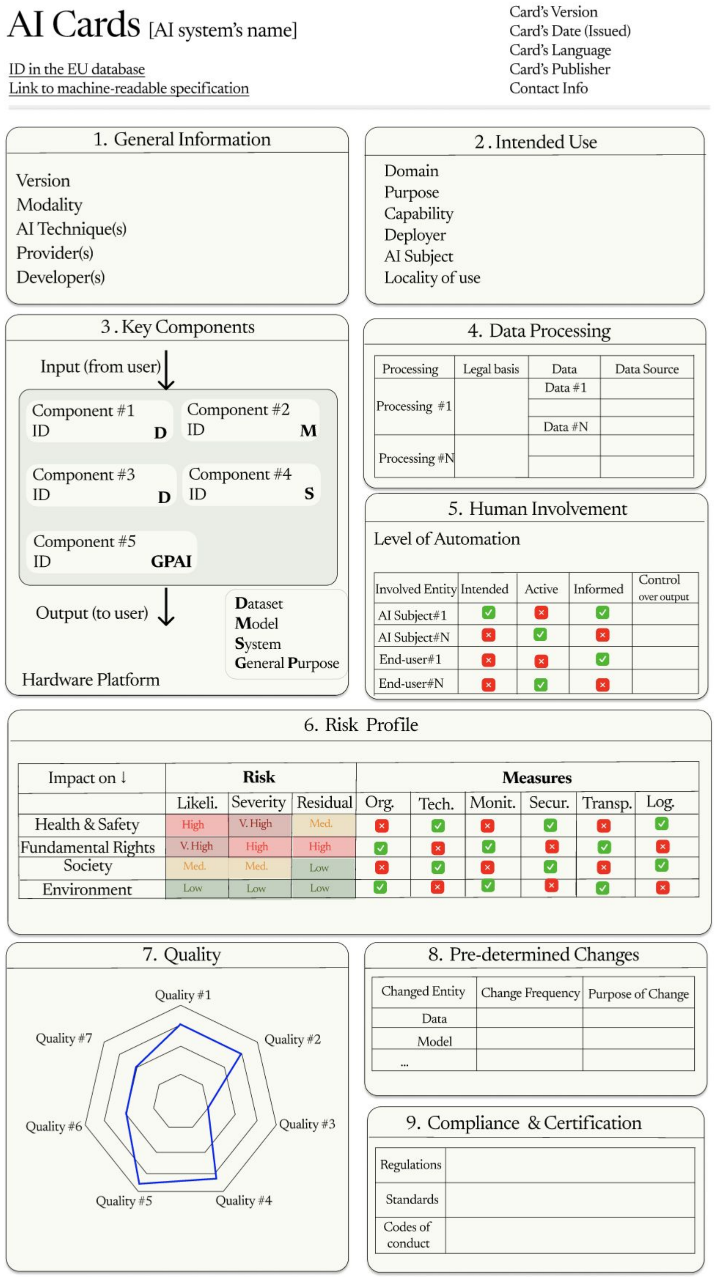
- Role of organisation in relation to the AI system
- External context of AI system
- Internal Context of AI system
- Intended purpose of the AI system
- Expectations regarding AI risk management
- Scope of AI risk management system

Types of information

1. Information about the context of the AI system and the organisation
2. Details of the risk management system in place
3. Risk management processes
4. Results of AI risk management

AI Cards (I)

Engaging Content
Engaging People



AI Cards [AI system's name]

ID in the EU database
Link to machine-readable specification

Card's Version
Card's Date (Issued)
Card's Language
Card's Publisher
Contact Info

1. General Information

Version
Modality
AI Technique(s)
Provider(s)
Developer(s)

2. Intended Use

Domain
Purpose
Capability
Deployer
AI Subject
Locality of use

3. Key Components

Input (from user)

Component #1
ID D

Component #2
ID M

Component #3
ID D

Component #4
ID S

Component #5
ID GPAI

Output (to user)

Dataset
Model
System
General Purpose

Hardware Platform

4. Data Processing

Processing	Legal basis	Data	Data Source
Processing #1		Data #1	
		Data #N	
Processing #N			

5. Human Involvement

Level of Automation

Involved Entity	Intended	Active	Informed	Control over output
AI Subject#1	✓	✗	✓	
AI Subject#N	✗	✓	✗	
End-user#1	✗	✗	✓	
End-user#N	✗	✓	✗	

6. Risk Profile

Impact on ↓	**Risk**			**Measures**					
	Likeli.	Severity	Residual	Org.	Tech.	Monit.	Secur.	Transp.	Log.
Health & Safety	High	V. High	Med.	✗	✓	✗	✓	✗	✓
Fundamental Rights	V. High	High	High	✓	✗	✓	✗	✓	✗
Society	Med.	Med.	Low	✗	✓	✗	✓	✗	✓
Environment	Low	Low	Low	✓	✗	✓	✗	✓	✗

7. Quality

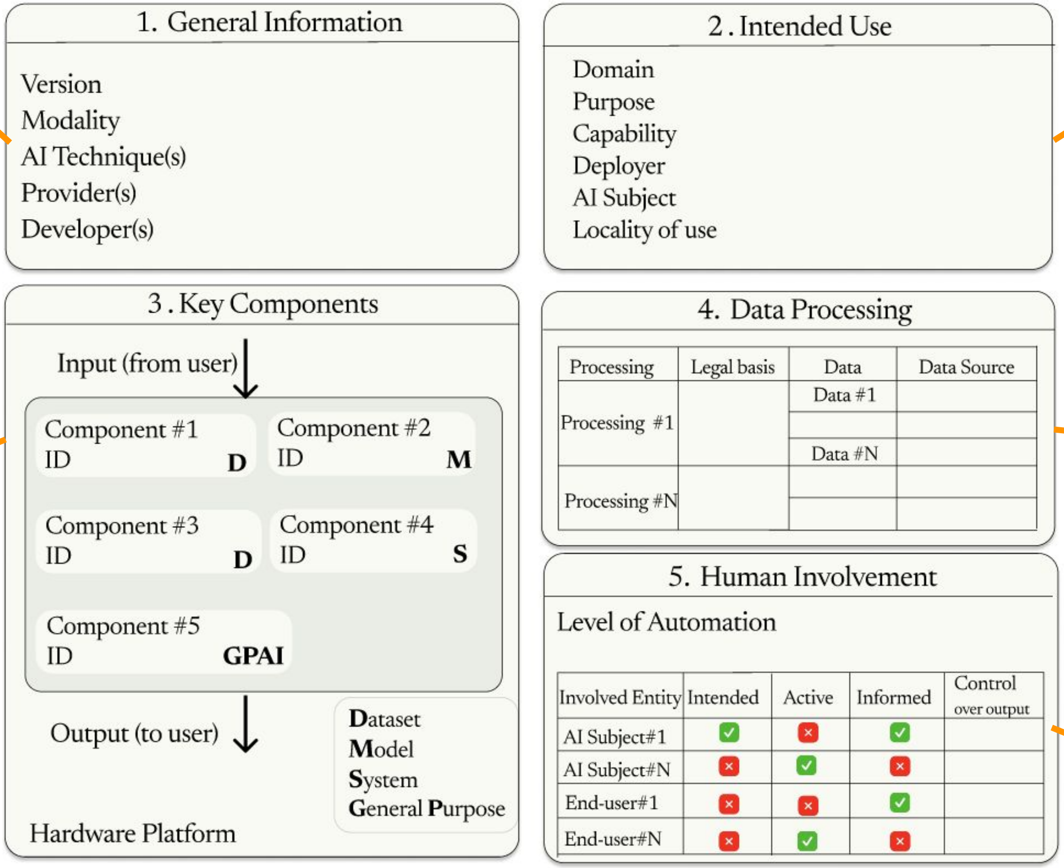
8. Pre-determined Changes

Changed Entity	Change Frequency	Purpose of Change
Data		
Model		
...		

9. Compliance & Certification

Regulations	
Standards	
Codes of conduct	

AI Cards (II)



1. General Information

Version
Modality
AI Technique(s)
Provider(s)
Developer(s)

2 . Intended Use

Domain
Purpose
Capability
Deployer
AI Subject
Locality of use

3 . Key Components

Input (from user)

Component #1 ID D	Component #2 ID M
Component #3 ID D	Component #4 ID S
Component #5 ID GPAI	

Output (to user)

Dataset
Model
System
General **P**urpose

Hardware Platform

4. Data Processing

Processing	Legal basis	Data	Data Source
Processing #1		Data #1	
		Data #N	
Processing #N			

5. Human Involvement

Level of Automation

Involved Entity	Intended	Active	Informed	Control over output
AI Subject#1	✓	✗	✓	
AI Subject#N	✗	✓	✗	
End-user#1	✗	✗	✓	
End-user#N	✗	✓	✗	

1. General Information about the system

2. Intended use of the AI system using 6 concepts

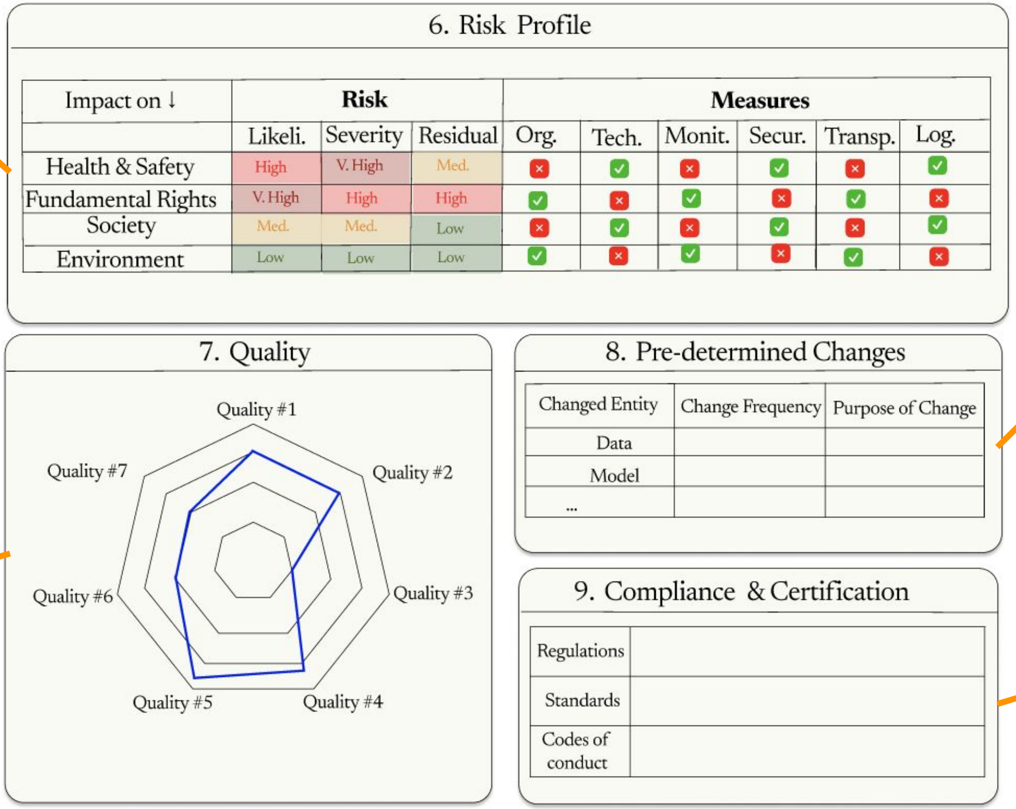
3. Information about the incorporating components

4. Information about processing of data (including info about legal basis and source of data)

5. Involvement of humans and level of automation

AI Cards (III)

6. High-level summary of risk management



6. Risk Profile

Impact on ↓	Risk			Measures					
	Likeli.	Severity	Residual	Org.	Tech.	Monit.	Secur.	Transp.	Log.
Health & Safety	High	V. High	Med.	✗	✓	✗	✓	✗	✓
Fundamental Rights	V. High	High	High	✓	✗	✓	✗	✓	✗
Society	Med.	Med.	Low	✗	✓	✗	✓	✗	✓
Environment	Low	Low	Low	✓	✗	✓	✗	✓	✗

7. Quality

8. Pre-determined Changes

Changed Entity	Change Frequency	Purpose of Change
Data		
Model		
...		

9. Compliance & Certification

Regulations	
Standards	
Codes of conduct	

7. Illustration of key qualities of the AI system

8. List of pre-determined changes

8.Regulation & Certification information

Example:
An AI-Based Student Proctoring System

Engaging Content
Engaging People

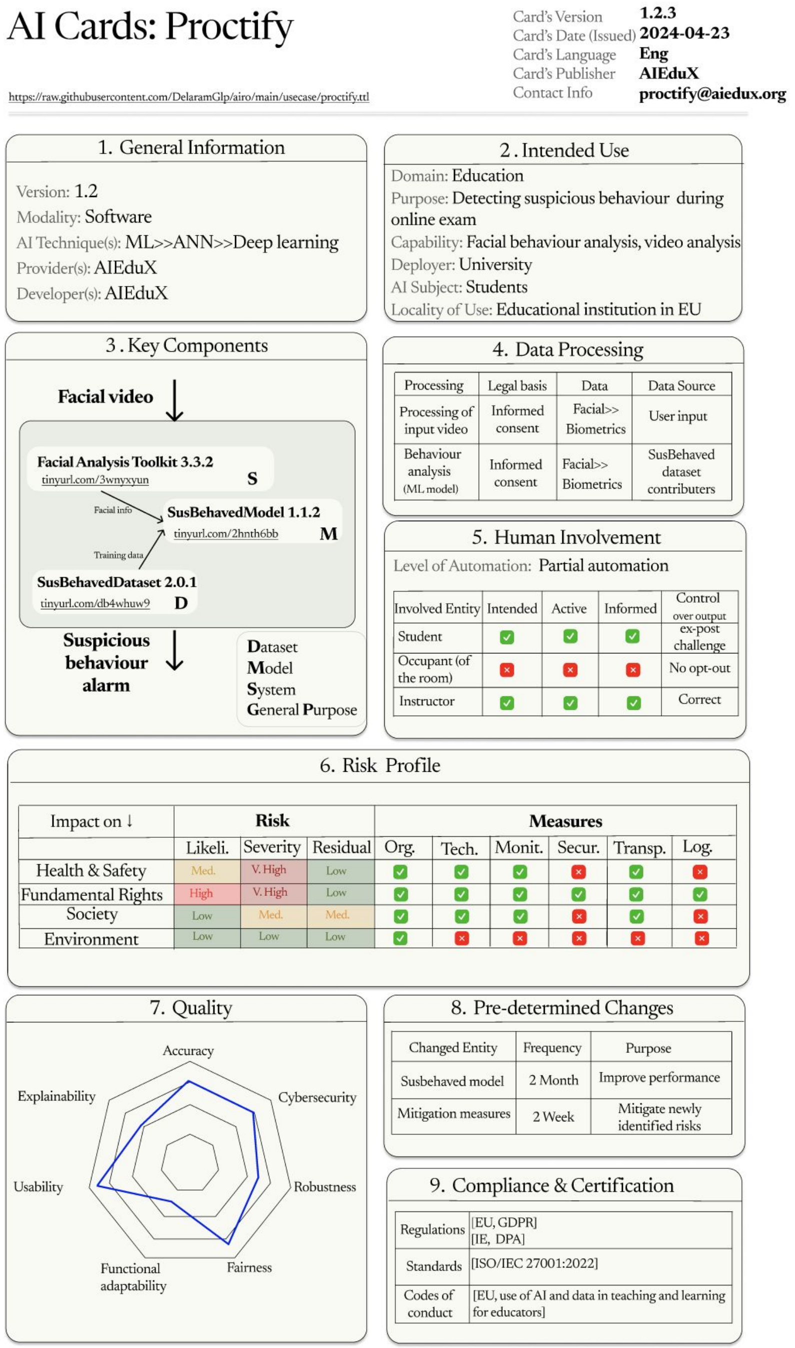
Human-readable description

Proctify is intended to be used in the education domain, for detecting suspicious behaviour of students during online exams in universities. Facial behaviour analysis and video analysis are used for detecting suspicious behaviour

Machine-readable specification

```
ex:proctify
    airo:isAppliedWithinDomain ex:education ;
    airo:hasPurpose ex:detecting_suspicious_bahviour_during_online_exam
    airo:hasCapability ex:facial_behaviour_analysis ;
    airo:hasCapability ex:video_analysis ;
    airo:isUsedBy ex:university ;
    airo:hasAISubject ex:student ;
```

https://delaramglp.github.io/aicards/example/



AI Cards: Proctify

https://raw.githubusercontent.com/DelaramGlp/airo/main/usecase/proctify.ttl

Card's Version	1.2.3
Card's Date (Issued)	2024-04-23
Card's Language	Eng
Card's Publisher	AIEduX
Contact Info	proctify@aiedux.org

1. General Information

Version: 1.2
Modality: Software
AI Technique(s): ML>>ANN>>Deep learning
Provider(s): AIEduX
Developer(s): AIEduX

2 . Intended Use

Domain: Education
Purpose: Detecting suspicious behaviour during online exam
Capability: Facial behaviour analysis, video analysis
Deployer: University
AI Subject: Students
Locality of Use: Educational institution in EU

3 . Key Components

Facial video

Facial Analysis Toolkit 3.3.2 tinyurl.com/3wnyxyun S

SusBehavedModel 1.1.2 tinyurl.com/2hnth6bb M

SusBehavedDataset 2.0.1 tinyurl.com/db4whuw9 D

Suspicious behaviour alarm

Dataset
Model
System
General **P**urpose

4. Data Processing

Processing	Legal basis	Data	Data Source
Processing of input video	Informed consent	Facial>> Biometrics	User input
Behaviour analysis (ML model)	Informed consent	Facial>> Biometrics	SusBehaved dataset contributers

5. Human Involvement

Level of Automation: Partial automation

Involved Entity	Intended	Active	Informed	Control over output
Student	✓	✓	✓	ex-post challenge
Occupant (of the room)	✗	✗	✗	No opt-out
Instructor	✓	✓	✓	Correct

6. Risk Profile

Impact on ↓	**Risk**			**Measures**					
	Likeli.	Severity	Residual	Org.	Tech.	Monit.	Secur.	Transp.	Log.
Health & Safety	Med.	V. High	Low	✓	✓	✓	✗	✓	✗
Fundamental Rights	High	V. High	Low	✓	✓	✓	✓	✓	✓
Society	Low	Med.	Med.	✓	✓	✓	✗	✓	✗
Environment	Low	Low	Low	✓	✗	✗	✗	✗	✗

7. Quality

8. Pre-determined Changes

Changed Entity	Frequency	Purpose
Susbehaved model	2 Month	Improve performance
Mitigation measures	2 Week	Mitigate newly identified risks

9. Compliance & Certification

Regulations	[EU, GDPR] [IE, DPA]
Standards	[ISO/IEC 27001:2022]
Codes of conduct	[EU, use of AI and data in teaching and learning for educators]

Benefits of the Machine-Readable Representation

Engaging Content
Engaging People

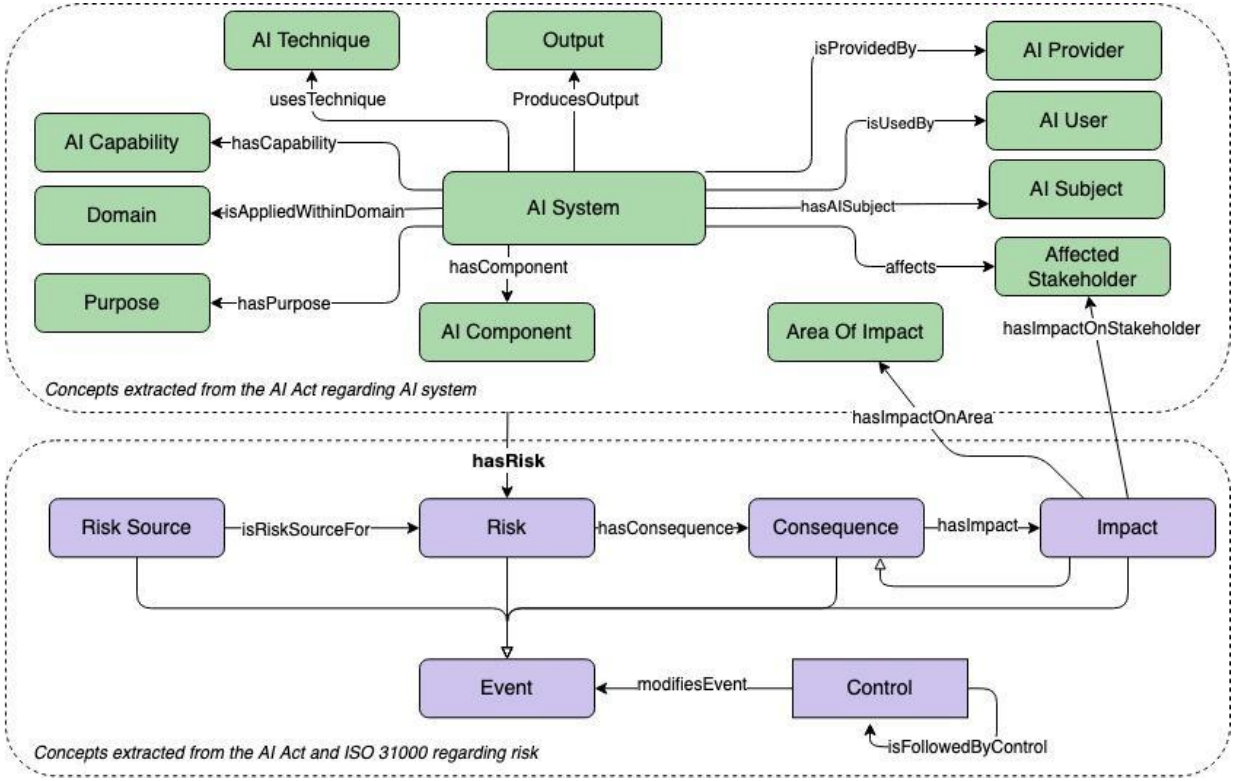
- Consistency
- Interoperability
- Integration with other documents, e.g. HuggingFace's Model Cards
- Automated generation of AI Cards (using SPARQL queries)
- Querying to support compliance checking

https://w3id.org/airo

AI Risk Ontology (AIRO)

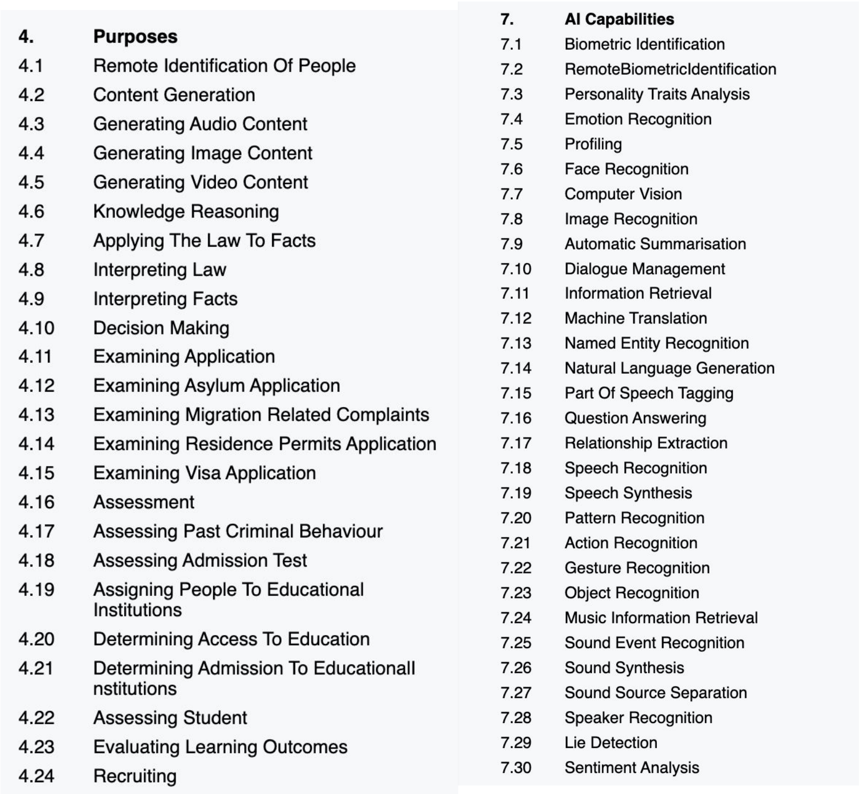
https://w3id.org/vair

Vocabulary of AI Risks (VAIR)



Concepts extracted from the AI Act regarding AI system

Concepts extracted from the AI Act and ISO 31000 regarding risk

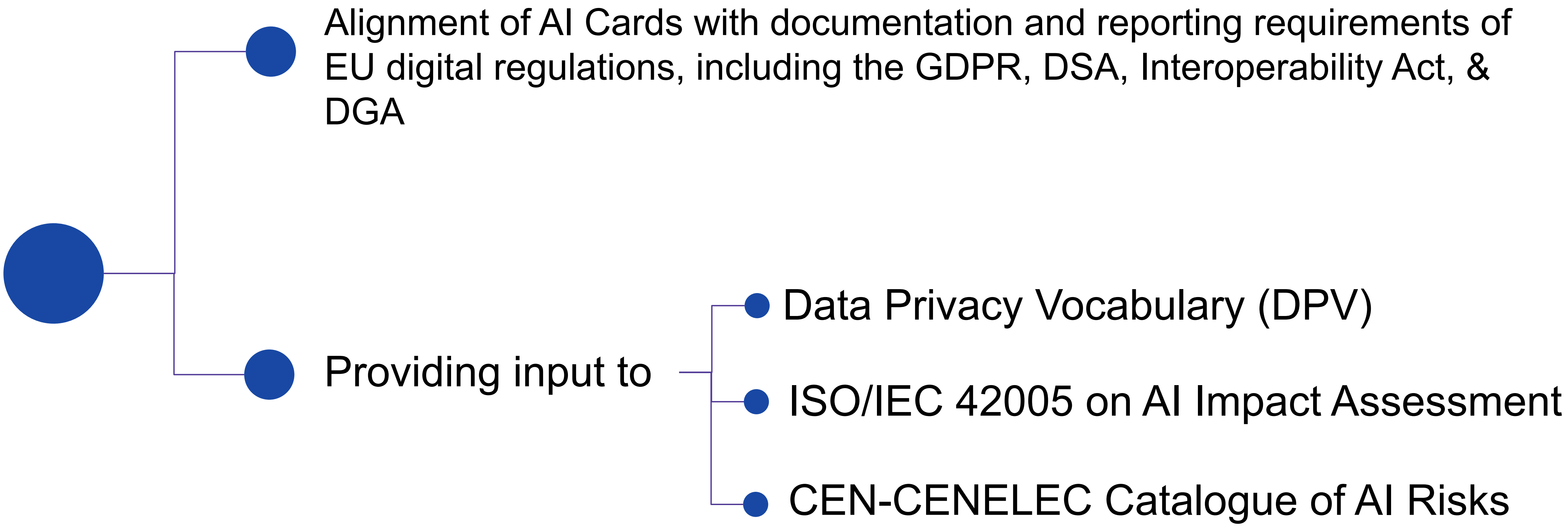


4.	Purposes
4.1	Remote Identification Of People
4.2	Content Generation
4.3	Generating Audio Content
4.4	Generating Image Content
4.5	Generating Video Content
4.6	Knowledge Reasoning
4.7	Applying The Law To Facts
4.8	Interpreting Law
4.9	Interpreting Facts
4.10	Decision Making
4.11	Examining Application
4.12	Examining Asylum Application
4.13	Examining Migration Related Complaints
4.14	Examining Residence Permits Application
4.15	Examining Visa Application
4.16	Assessment
4.17	Assessing Past Criminal Behaviour
4.18	Assessing Admission Test
4.19	Assigning People To Educational Institutions
4.20	Determining Access To Education
4.21	Determining Admission To Educationall nstitutions
4.22	Assessing Student
4.23	Evaluating Learning Outcomes
4.24	Recruiting

7.	AI Capabilities
7.1	Biometric Identification
7.2	RemoteBiometricIdentification
7.3	Personality Traits Analysis
7.4	Emotion Recognition
7.5	Profiling
7.6	Face Recognition
7.7	Computer Vision
7.8	Image Recognition
7.9	Automatic Summarisation
7.10	Dialogue Management
7.11	Information Retrieval
7.12	Machine Translation
7.13	Named Entity Recognition
7.14	Natural Language Generation
7.15	Part Of Speech Tagging
7.16	Question Answering
7.17	Relationship Extraction
7.18	Speech Recognition
7.19	Speech Synthesis
7.20	Pattern Recognition
7.21	Action Recognition
7.22	Gesture Recognition
7.23	Object Recognition
7.24	Music Information Retrieval
7.25	Sound Event Recognition
7.26	Sound Synthesis
7.27	Sound Source Separation
7.28	Speaker Recognition
7.29	Lie Detection
7.30	Sentiment Analysis

AIRO and VAIR are going to be integrated with DPV (https://w3id.org/dpv)

Future Work



Alignment of AI Cards with documentation and reporting requirements of EU digital regulations, including the GDPR, DSA, Interoperability Act, & DGA

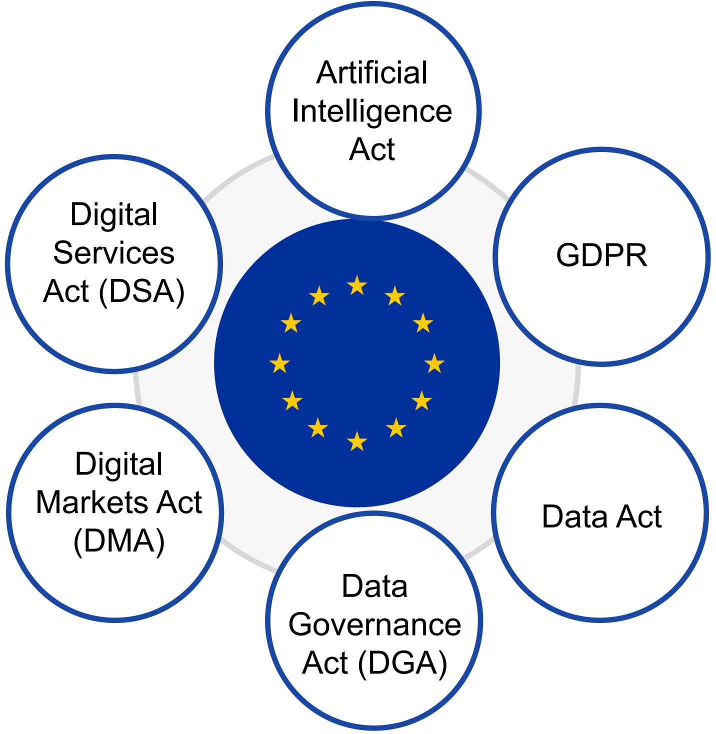
Providing input to

- Data Privacy Vocabulary (DPV)
- ISO/IEC 42005 on AI Impact Assessment
- CEN-CENELEC Catalogue of AI Risks

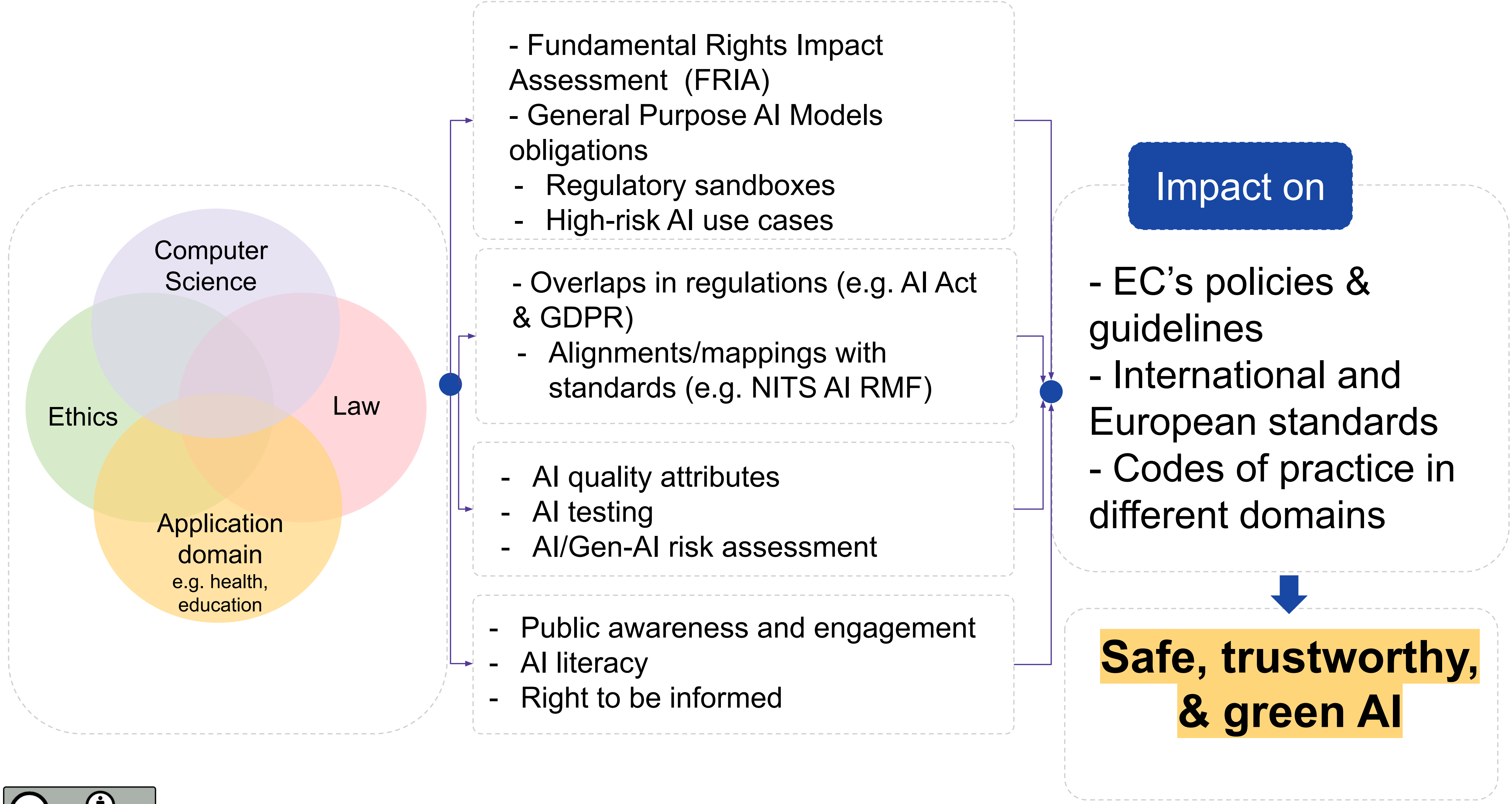
Navigating in the New European Regulatory Environment (5+1 regulations)

RegTech Solutions

- Risk and Operations Tech
- Compliance and Reporting Tech
  - To support compliance and conformity assessment
- Supervisory Tech
- Ethical Tech


Artificial Intelligence Act
GDPR
Data Act
Data Governance Act (DGA)
Digital Markets Act (DMA)
Digital Services Act (DSA)

Legal AI ≠ Ethical AI



- Fundamental Rights Impact Assessment (FRIA)
- General Purpose AI Models obligations
- Regulatory sandboxes
- High-risk AI use cases

- Overlaps in regulations (e.g. AI Act & GDPR)
- Alignments/mappings with standards (e.g. NITS AI RMF)

- AI quality attributes
- AI testing
- AI/Gen-AI risk assessment

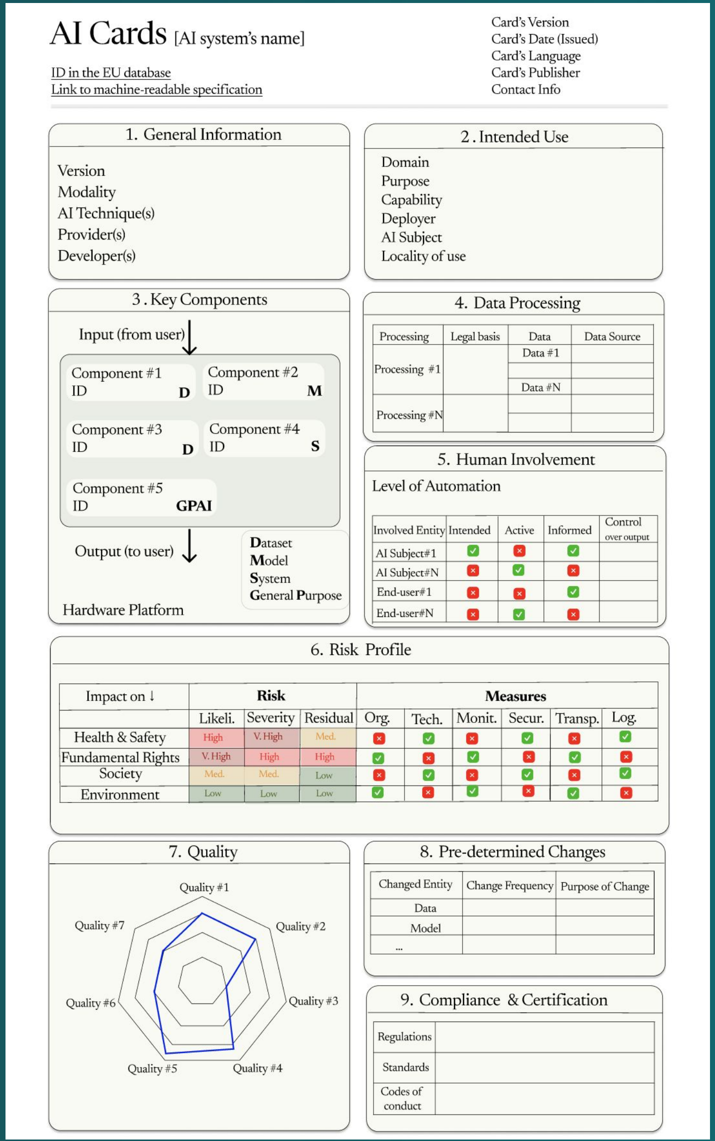
- Public awareness and engagement
- AI literacy
- Right to be informed

Impact on

- EC's policies & guidelines
- International and European standards
- Codes of practice in different domains

Safe, trustworthy, & green AI

AI Cards | Delaram Golpayegani et al. | ADAPT Scientific Conference | 14 May 2024 | delaram.golpayegani@adaptcentre.ie



AI Cards [AI system's name]

ID in the EU database
Link to machine-readable specification

Card's Version
Card's Date (Issued)
Card's Language
Card's Publisher
Contact Info

1. General Information

Version
Modality
AI Technique(s)
Provider(s)
Developer(s)

2 . Intended Use

Domain
Purpose
Capability
Deployer
AI Subject
Locality of use

3 . Key Components

Input (from user)

Component #1
ID D

Component #2
ID M

Component #3
ID D

Component #4
ID S

Component #5
ID GPAI

Output (to user)

Dataset
Model
System
General Purpose

Hardware Platform

4. Data Processing

Processing	Legal basis	Data	Data Source
Processing #1		Data #1	
		Data #N	
Processing #N			

5. Human Involvement

Level of Automation

Involved Entity	Intended	Active	Informed	Control over output
AI Subject#1	✓	✗	✓	
AI Subject#N	✗	✓	✗	
End-user#1	✗	✗	✓	
End-user#N	✗	✓	✗	

6. Risk Profile

Impact on ↓	Risk			Measures					
	Likeli.	Severity	Residual	Org.	Tech.	Monit.	Secur.	Transp.	Log.
Health & Safety	High	V. High	Med.	✗	✓	✗	✓	✗	✓
Fundamental Rights	V. High	High	High	✓	✗	✓	✗	✓	✗
Society	Med.	Med.	Low	✗	✓	✗	✓	✗	✓
Environment	Low	Low	Low	✓	✗	✓	✗	✓	✗

7. Quality

8. Pre-determined Changes

Changed Entity	Change Frequency	Purpose of Change
Data		
Model		
...		

9. Compliance & Certification

Regulations	
Standards	
Codes of conduct	

Engaging Content
Engaging People

AI Cards: Towards an Applied Framework for Machine-Readable AI and Risk Documentation Inspired by the EU AI Act

Delaram Golpayegani, Isabelle Hupont, Cecilia Panigutti
Harshvardhan J. Pandit, Sven Schade, Declan O'Sullivan, Dave Lewis

delaram.golpayegani@adaptcentre.ie

The views expressed in this article are purely those of the authors and should not, under any circumstances, be regarded as an official position of the European Commission.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 813497 (PROTECT ITN) and is funded by Science Foundation Ireland through the SFI Research Centres Programme through Grant#13/RC/2106_P2.