Protect

AIRO

an Ontology for Representing AI Risks based on the Proposed EU AI Act and ISO Risk Management Standards

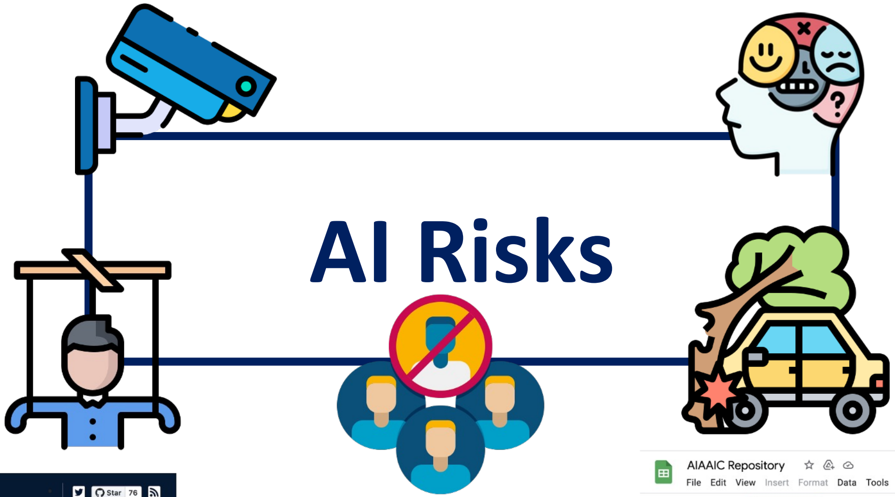
Delaram Golpayegani, Harshvardhan J. Pandit, Dave Lewis

ADAPT Centre, Trinity College Dublin, Ireland

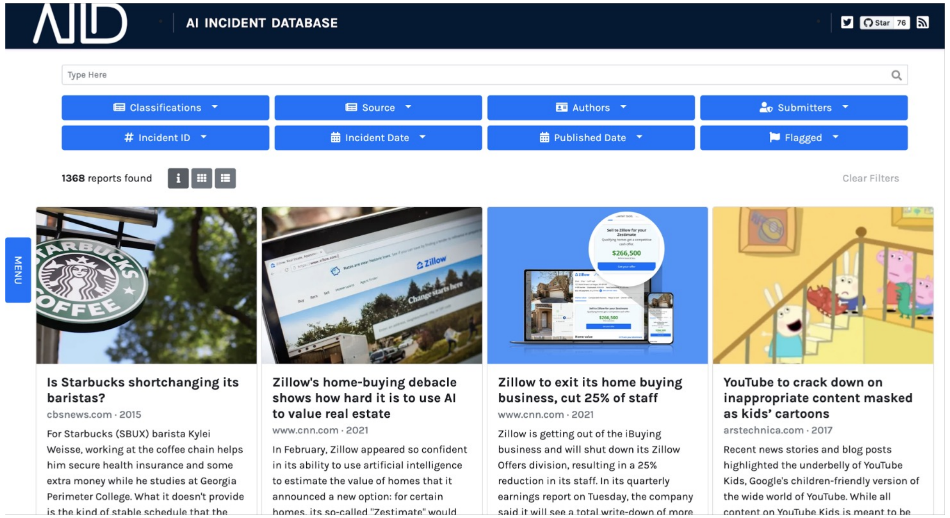
sgolpays@tcd.ie

PROTECT ITN
This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 813497.

AI Risks



AIID (AI Incident Database)
https://incidentdatabase.ai



1368 reports found

Is Starbucks shortchanging its baristas?
cbsnews.com · 2015
For Starbucks (SBUX) barista Kylei Weisse, working at the coffee chain helps him secure health insurance and some extra money while he studies at Georgia Perimeter College. What it doesn't provide is the kind of stable schedule that the

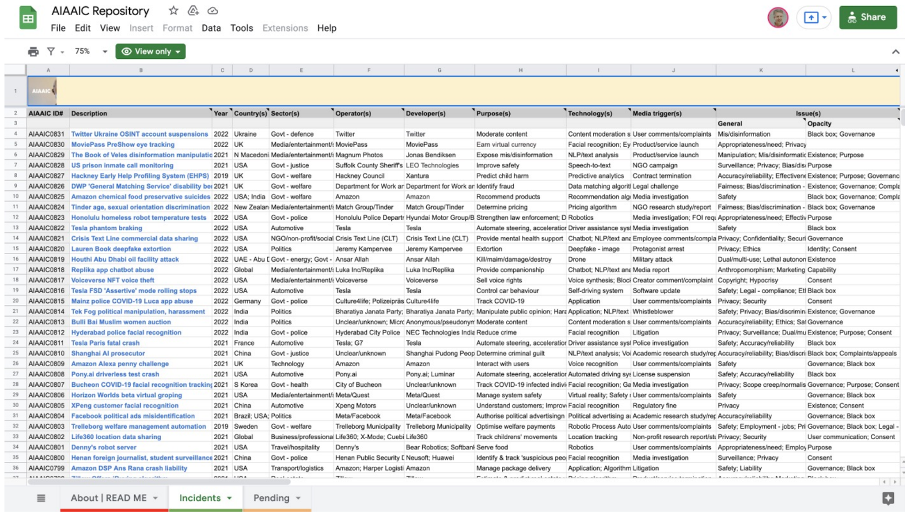
Zillow's home-buying debacle shows how hard it is to use AI to value real estate
www.cnn.com · 2021
In February, Zillow appeared so confident in its ability to use artificial intelligence to estimate the value of homes that it announced a new option: for certain homes, its so-called "Zestimate" would

Zillow to exit its home buying business, cut 25% of staff
www.cnn.com · 2021
Zillow is getting out of the iBuying business and will shut down its Zillow Offers division, resulting in a 25% reduction in its staff. In its quarterly earnings report on Tuesday, the company said it will see a total write-down of more

YouTube to crack down on inappropriate content masked as kids' cartoons
arstechnica.com · 2017
Recent news stories and blog posts highlighted the underbelly of YouTube Kids, Google's children-friendly version of the wide world of YouTube. While all content on YouTube Kids is meant to be

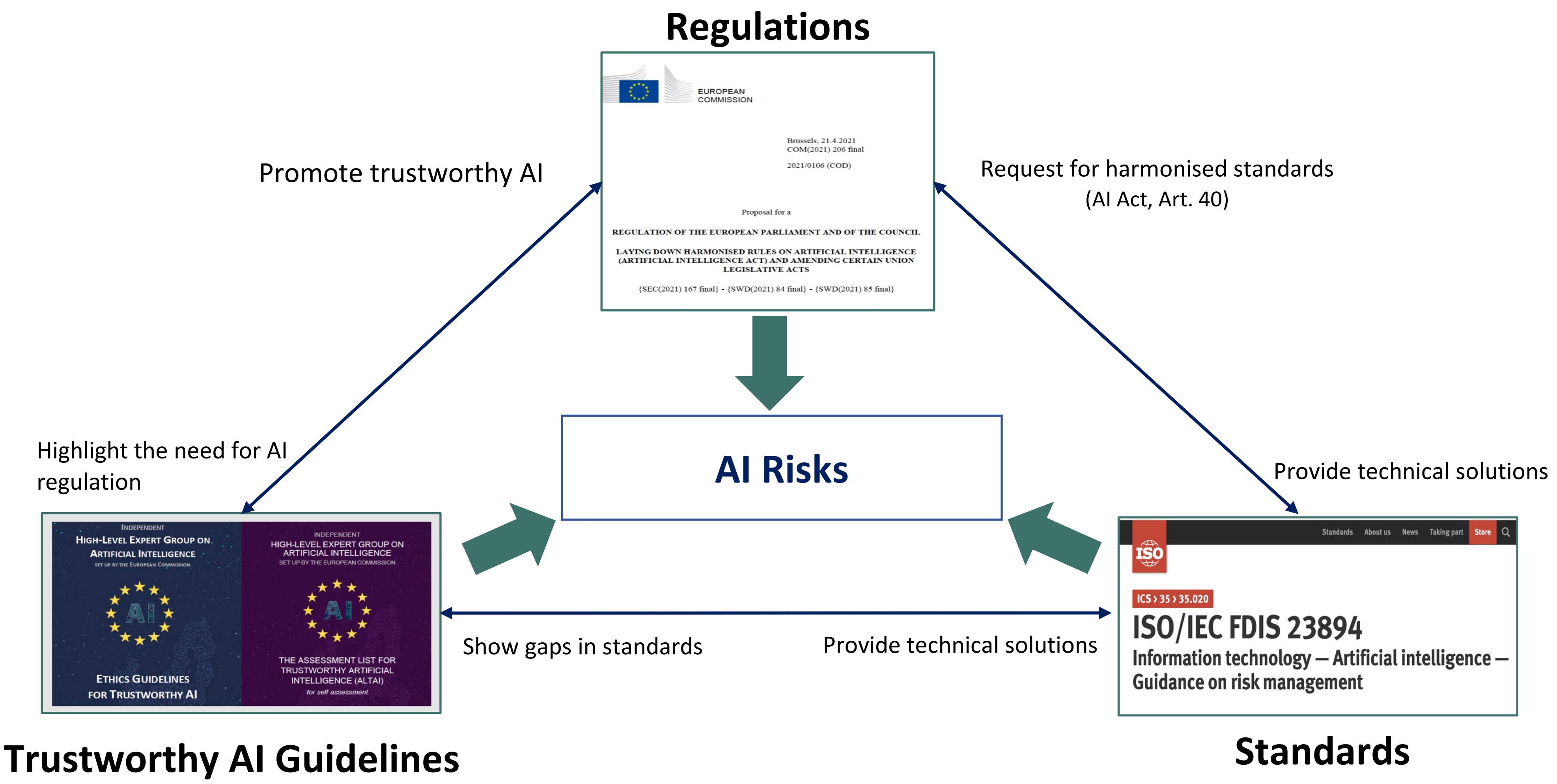
Icons from https://www.flaticon.com/

AIAAIC Repository
https://www.aiaaic.org/aiaaic-repository



AIAAIC ID#	Description	Year	Country(s)	Sector(s)	Operator(s)	Developer(s)	Purpose(s)	Technology(s)	Media trigger(s)	Issue(s)	
										General	Opacity
AIAAIC0831	Twitter Ukraine OSINT account suspensions	2022	Ukraine	Govt - defence	Twitter	Twitter	Moderate content	Content moderation s	User comments/complaints	Mis/disinformation	Black box; Governance
AIAAIC0830	MoviePass PreShow eye tracking	2022	UK	Media/entertainment/s	MoviePass	MoviePass	Earn virtual currency	Facial recognition; Ey	Product/service launch	Appropriateness/need; Privacy	
AIAAIC0829	The Book of Veles disinformation manipulatic	2021	N Macedoni	Media/entertainment/s	Magnum Photos	Jonas Bendiksen	Expose mis/disinformation	NLP/text analysis	Product/service launch	Manipulation; Mis/disinformatic	Existence; Purpose
AIAAIC0828	US prison inmate call monitoring	2021	USA	Govt - justice	Suffolk County Sheriff's	LEO Technologies	Improve safety	Speech-to-text	NGO campaign	Surveillance; Privacy; Bias/dis	Purpose
AIAAIC0827	Hackney Early Help Profiling System (EHPS)	2019	UK	Govt - welfare	Hackney Council	Xantura	Predict child harm	Predictive analytics	Contract termination	Accuracy/reliability; Effectivene	Existence; Purpose; Governanc
AIAAIC0826	DWP 'General Matching Service' disability be	2021	UK	Govt - welfare	Department for Work ar	Department for Work ar	Identify fraud	Data matching algorit	Legal challenge	Fairness; Bias/discrimination -	Existence; Governance; Compla
AIAAIC0825	Amazon chemical food preservative suicides	2022	USA; India	Govt - welfare	Amazon	Amazon	Recommend products	Recommendation alg	Media investigation	Safety	Black box; Governance; Compla
AIAAIC0824	Tinder age, sexual orientation discrimination	2022	New Zealan	Media/entertainment/s	Match Group/Tinder	Match Group/Tinder	Determine pricing	Pricing algorithm	NGO research study/report	Fairness; Bias/discrimination -	Black box; Governance
AIAAIC0823	Honolulu homeless robot temperature tests	2022	USA	Govt - police	Honolulu Police Departr	Hyundai Motor Group/B	Strengthen law enforcement; D	Robotics	Media investigation; FOI requ	Appropriateness/need; Effectiv	Purpose
AIAAIC0822	Tesla phantom braking	2022	USA	Automotive	Tesla	Tesla	Automate steering, acceleratior	Driver assistance syst	Media investigation	Safety	Black box
AIAAIC0821	Crisis Text Line commercial data sharing	2022	USA	NGO/non-profit/social	Crisis Text Line (CLT)	Crisis Text Line (CLT)	Provide mental health support	Chatbot; NLP/text ana	Employee comments/compla	Privacy; Confidentiality; Securi	Governance
AIAAIC0820	Lauren Book deepfake extortion	2022	USA	Politics	Jeremy Kampervee	Jeremy Kampervee	Extortion	Deepfake - image	Protagonist arrest	Privacy; Ethics	Identity; Consent
AIAAIC0819	Houthi Abu Dhabi oil facility attack	2022	UAE - Abu D	Govt - energy; Govt -	Ansar Allah	Ansar Allah	Kill/maim/damage/destroy	Drone	Military attack	Dual/multi-use; Lethal autonon	Existence
AIAAIC0818	Replika app chatbot abuse	2022	Global	Media/entertainment/s	Luka Inc/Replika	Luka Inc/Replika	Provide companionship	Chatbot; NLP/text ana	Media report	Anthropomorphism; Marketing	Capability
AIAAIC0817	Voiceverse NFT voice theft	2022	USA	Media/entertainment/s	Voiceverse	Voiceverse	Sell voice rights	Voice synthesis; Bloc	Creator comment/complaint	Copyright; Hypocrisy	Consent
AIAAIC0816	Tesla FSD 'Assertive' mode rolling stops	2022	USA	Automotive	Tesla	Tesla	Control car behaviour	Self-driving system	Software update	Safety; Legal - compliance; Et	Black box
AIAAIC0815	Mainz police COVID-19 Luca app abuse	2022	Germany	Govt - police	Culture4life; Polizeipräs	Culture4life	Track COVID-19	Application	User comments/complaints	Privacy; Security	Consent
AIAAIC0814	Tek Fog political manipulation, harassment	2022	India	Politics	Bharatiya Janata Party;	Bharatiya Janata Party;	Manipulate public opinion; Hara	Application; NLP/text	Whistleblower	Safety; Privacy; Bias/discrimin	Existence; Governance
AIAAIC0813	Bulli Bai Muslim women auction	2022	India	Politics	Unclear/unknown; Micro	Anonymous/pseudonym	Moderate content	Content moderation s	User comments/complaints	Accuracy/reliability; Ethics; Sa	Governance
AIAAIC0812	Hyderabad police facial recognition	2022	India	Govt - police	Hyderabad City Police	NEC Technologies India	Reduce crime	Facial recognition	Litigation	Privacy; Surveillance; Dual/mu	Existence; Purpose; Consent
AIAAIC0811	Tesla Paris fatal crash	2021	France	Automotive	Tesla; G7	Tesla	Automate steering, acceleratior	Driver assistance syst	Police investigation	Safety; Accuracy/reliability	Black box
AIAAIC0810	Shanghai AI prosecutor	2021	China	Govt - justice	Unclear/unknown	Shanghai Pudong Peop	Determine criminal guilt	NLP/text analysis; Voi	Academic research study/rep	Accuracy/reliability; Bias/discri	Black box; Complaints/appeals
AIAAIC0809	Amazon Alexa penny challenge	2021	UK	Technology	Amazon	Amazon	Interact with users	Voice recognition	User comments/complaints	Safety	Governance; Black box
AIAAIC0808	Pony.ai driverless test crash	2021	USA	Automotive	Pony.ai	Pony.ai; Luminar	Automate steering, acceleratior	Automated driving sys	License suspension	Safety; Accuracy/reliability	Black box
AIAAIC0807	Bucheon COVID-19 facial recognition tracking	2021	S Korea	Govt - health	City of Bucheon	Unclear/unknown	Track COVID-19 infected indivi	Facial recognition; Ga	Media investigation	Privacy; Scope creep/normalis	Governance; Purpose; Consent
AIAAIC0806	Horizon Worlds beta virtual groping	2021	USA	Media/entertainment/s	Meta/Quest	Meta/Quest	Manage system safety	Virtual reality; Safety	User comments/complaints	Safety	Governance; Black box
AIAAIC0805	XPeng customer facial recognition	2021	China	Automotive	Xpeng Motors	Unclear/unknown	Understand customers; Improv	Facial recognition	Regulatory fine	Privacy	Existence; Consent
AIAAIC0804	Facebook political ads misidentification	2021	Brazil; USA;	Politics	Meta/Facebook	Meta/Facebook	Authorise political advertisingn	Political advertising a	Academic research study/rep	Accuracy/reliability	Governance; Black box
AIAAIC0803	Trelleborg welfare management automation	2019	Sweden	Govt - welfare	Trelleborg Municipality	Trelleborg Municipality	Optimise welfare payments	Robotic Process Auto	User comments/complaints	Safety; Employment - jobs; Pri	Governance; Black box; Legal -
AIAAIC0802	Life360 location data sharing	2021	Global	Business/professiona	Life360; X-Mode; Cuebi	Life360	Track childrens' movements	Location tracking	Non-profit research report/stu	Privacy; Security	User communication; Consent
AIAAIC0801	Denny's robot server	2021	USA	Travel/hospitality	Denny's	Bear Robotics; Softbanl	Serve food	Robotics	User comments/complaints	Appropriateness/need; Employ	Purpose
AIAAIC0800	Henan foreign journalist, student surveillance	2021	China	Govt - police	Henan Public Security D	Neusoft; Huawei	Identify & track 'suspicious peo	Facial recognition	Media investigation	Surveillance; Privacy	Consent
AIAAIC0799	Amazon DSP Ans Rana crash liability	2021	USA	Transport/logistics	Amazon; Harper Logisti	Amazon	Manage package delivery	Application; Algorithm	Litigation	Safety; Liability	Governance; Black box

Efforts Addressing AI Risks



Regulations

EUROPEAN
COMMISSION

Brussels, 21.4.2021
COM(2021) 206 final

2021/0106 (COD)

Proposal for a

REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL

LAYING DOWN HARMONISED RULES ON ARTIFICIAL INTELLIGENCE (ARTIFICIAL INTELLIGENCE ACT) AND AMENDING CERTAIN UNION LEGISLATIVE ACTS

{SEC(2021) 167 final} - {SWD(2021) 84 final} - {SWD(2021) 85 final}

Promote trustworthy AI

Request for harmonised standards
(AI Act, Art. 40)

AI Risks

Highlight the need for AI regulation

Provide technical solutions

INDEPENDENT
HIGH-LEVEL EXPERT GROUP ON ARTIFICIAL INTELLIGENCE
SET UP BY THE EUROPEAN COMMISSION

ETHICS GUIDELINES FOR TRUSTWORTHY AI

INDEPENDENT
HIGH-LEVEL EXPERT GROUP ON ARTIFICIAL INTELLIGENCE
SET UP BY THE EUROPEAN COMMISSION

THE ASSESSMENT LIST FOR TRUSTWORTHY ARTIFICIAL INTELLIGENCE (ALTAI)
for self assessment

Show gaps in standards

Provide technical solutions

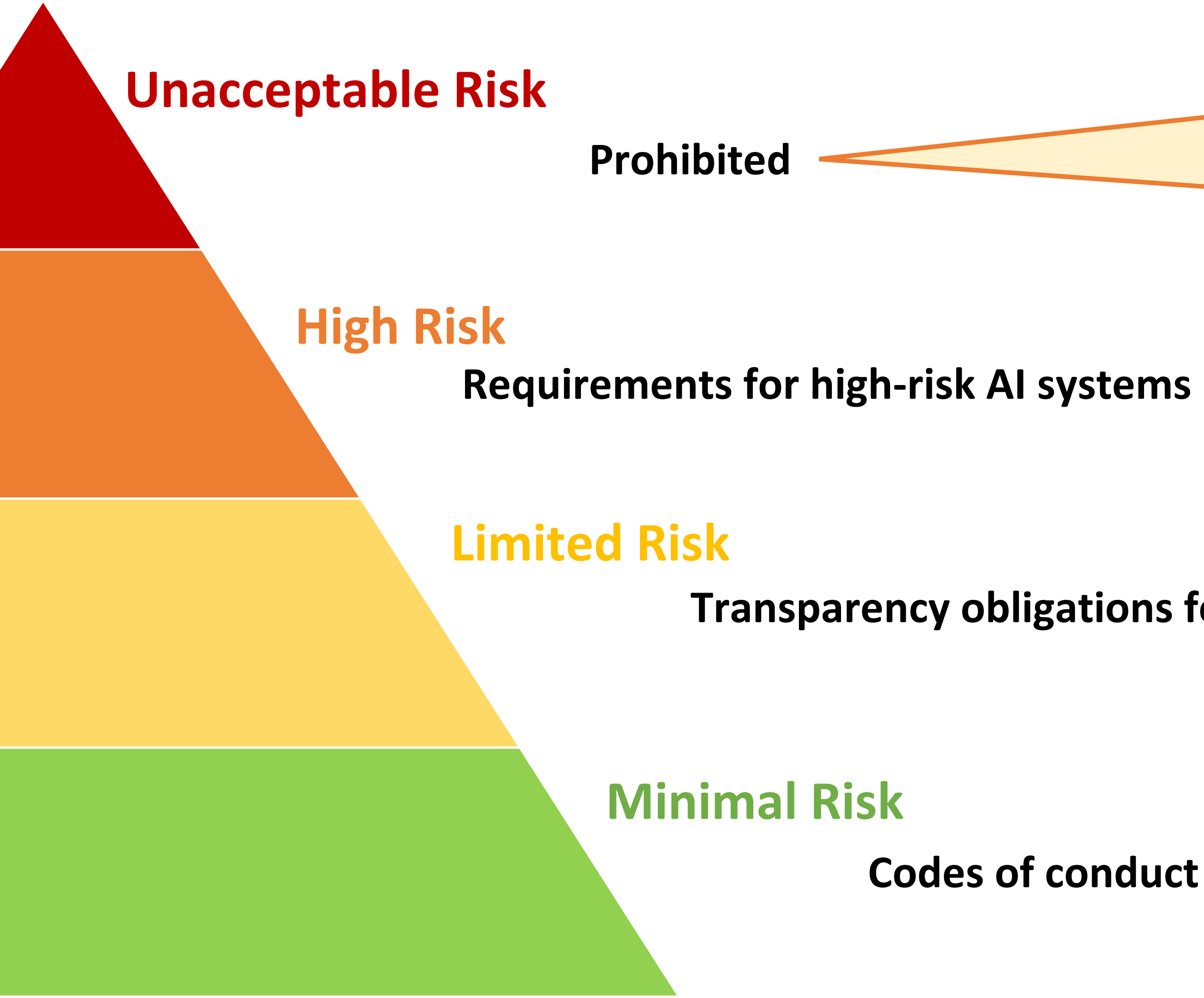
ICS › 35 › 35.020

ISO/IEC FDIS 23894
Information technology — Artificial intelligence — Guidance on risk management

Trustworthy AI Guidelines

Standards

AI Act Risk Pyramid



Unacceptable Risk

Prohibited

High Risk

Requirements for high-risk AI systems

Limited Risk

Transparency obligations for certain AI systems

Minimal Risk

Codes of conduct

High Risk to health, safety, and fundamental rights of people:
1) Product or safety component of a product covered by Annex II
2) AI system used in Annex III areas

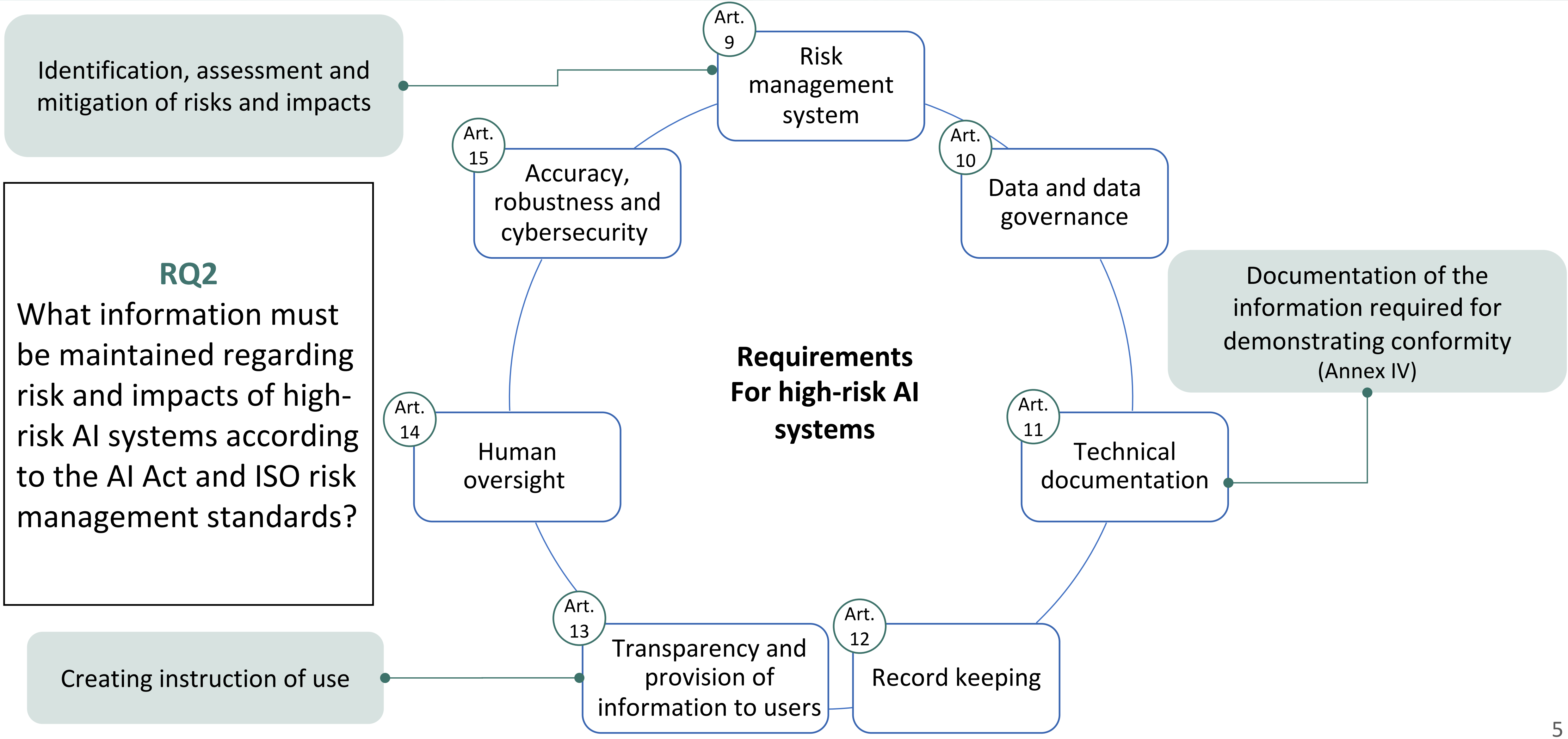
RQ1

What is the information required to determine whether an AI system is ‘high-risk’ as per the AI Act?

Requirements for High-Risk AI Systems

RQ2

What information must be maintained regarding risk and impacts of high-risk AI systems according to the AI Act and ISO risk management standards?



Identification, assessment and mitigation of risks and impacts

Art. 9 Risk management system

Art. 10 Data and data governance

Documentation of the information required for demonstrating conformity (Annex IV)

Art. 11 Technical documentation

Art. 12 Record keeping

Art. 13 Transparency and provision of information to users

Creating instruction of use

Art. 14 Human oversight

Art. 15 Accuracy, robustness and cybersecurity

Requirements For high-risk AI systems

Challenge

- Maintaining, querying, and sharing information associated with risks for compliance checking, demonstrating accountability, and building trust

- Challenges:
  - The pace of changes in AI systems
  - The amount of risk-related information
  - The complexities in the AI value chain

RQ3
To what extent can semantic web technologies assist with representing information and generating documentation for high-risk AI systems required by the AI Act?

Using semantic web technologies:
- enables automation
- Interoperability

Research Questions

1 What is the information required to determine whether an AI system is ‘high-risk’ as per the AI Act?

2 What information must be maintained regarding risk and impacts of high-risk AI systems according to the AI Act and ISO risk management standards?

Identify information requirements from:

- the AI Act
- ISO 31000 family

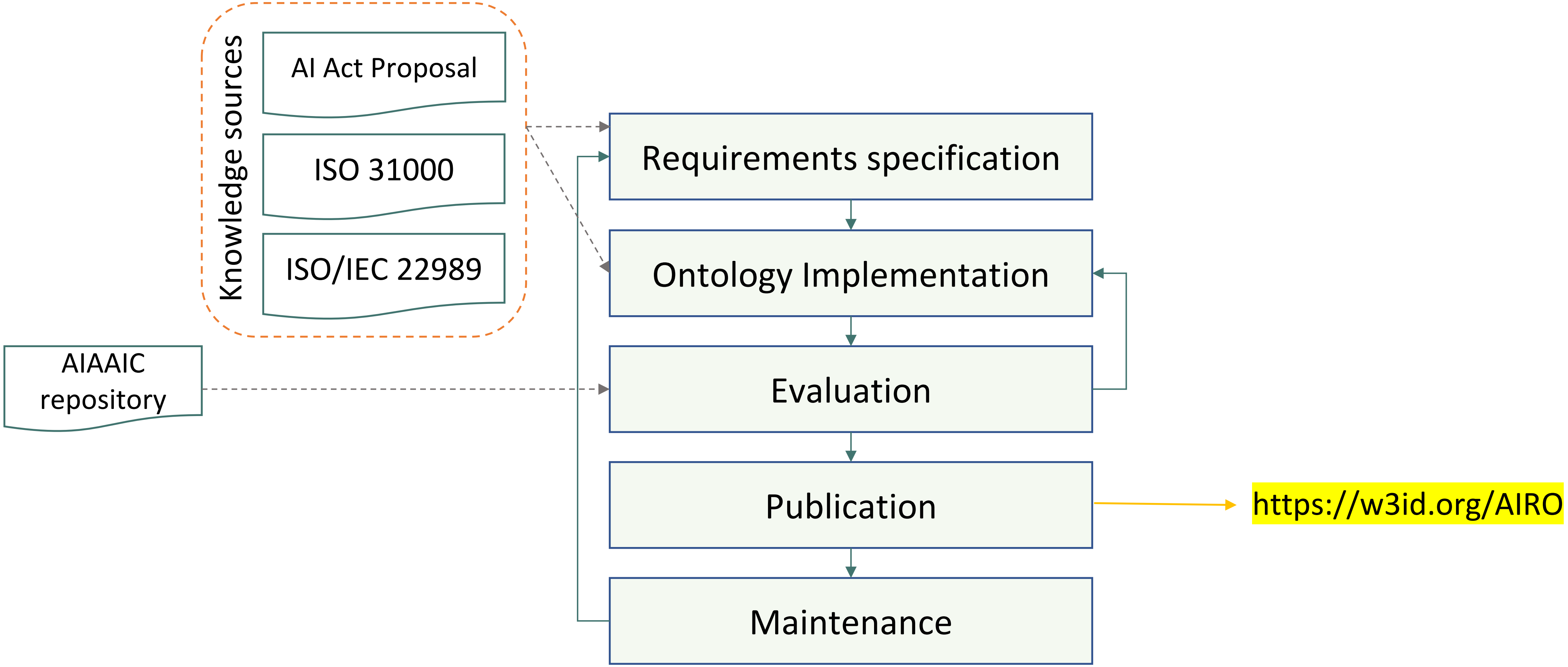
3 To what extent can semantic web technologies assist with representing information and generating documentation for high-risk AI systems required by the AI Act?

Create AIRO (AI Risk Ontology) demonstrate its applicability in real-world cases

State of the Art

Topic	Summary	Relation to this work
AI risk management standards	ISO 31000:2018 Risk management– Guidelines ISO 31073:2022 Risk management — Vocabulary	Used for identifying risk concepts
AI risk taxonomies	Existing taxonomies of AI risks, harms, risk sources, & mitigation measures	Reusing the taxonomies for populating AIRO
Risk models & ontologies	- Generic risk models - Domain-specific risk models	Reusing risk concepts

Ontology Development Methodology

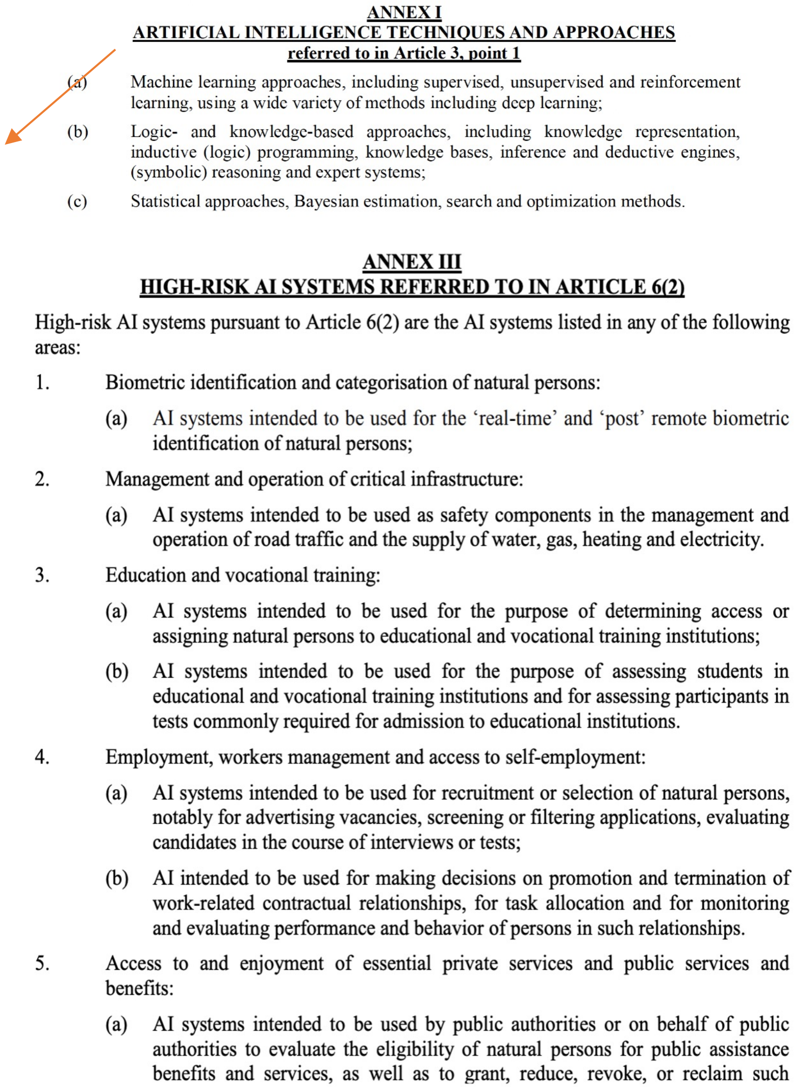


AIRO Requirements

Describing High-Risk AI Systems

Questions to identify whether an AI system is high-risk according to Annex III

Question	concept	Relation with AISystem
What techniques are utilised in the system?	**AI Technique**	usesAITechnique
What domain is the system intended to be used in?	**Domain**	isAppliedWithinDomain
What is the intended purpose of the system?	**Purpose**	hasPurpose
What is the application of the system?	**AI Application**	hasApplication
Who is the intended user of the system?	**AI User**	hasAIUser
Who is the subject of the system?	**AI Subject**	hasAISubject
In which environment is the system used?	**Environment Of Use**	isUsedInEnvironment



ANNEX I
ARTIFICIAL INTELLIGENCE TECHNIQUES AND APPROACHES
referred to in Article 3, point 1

(a) Machine learning approaches, including supervised, unsupervised and reinforcement learning, using a wide variety of methods including deep learning;

(b) Logic- and knowledge-based approaches, including knowledge representation, inductive (logic) programming, knowledge bases, inference and deductive engines, (symbolic) reasoning and expert systems;

(c) Statistical approaches, Bayesian estimation, search and optimization methods.

ANNEX III
HIGH-RISK AI SYSTEMS REFERRED TO IN ARTICLE 6(2)

High-risk AI systems pursuant to Article 6(2) are the AI systems listed in any of the following areas:

1. Biometric identification and categorisation of natural persons:

   (a) AI systems intended to be used for the 'real-time' and 'post' remote biometric identification of natural persons;

2. Management and operation of critical infrastructure:

   (a) AI systems intended to be used as safety components in the management and operation of road traffic and the supply of water, gas, heating and electricity.

3. Education and vocational training:

   (a) AI systems intended to be used for the purpose of determining access or assigning natural persons to educational and vocational training institutions;

   (b) AI systems intended to be used for the purpose of assessing students in educational and vocational training institutions and for assessing participants in tests commonly required for admission to educational institutions.

4. Employment, workers management and access to self-employment:

   (a) AI systems intended to be used for recruitment or selection of natural persons, notably for advertising vacancies, screening or filtering applications, evaluating candidates in the course of interviews or tests;

   (b) AI intended to be used for making decisions on promotion and termination of work-related contractual relationships, for task allocation and for monitoring and evaluating performance and behavior of persons in such relationships.

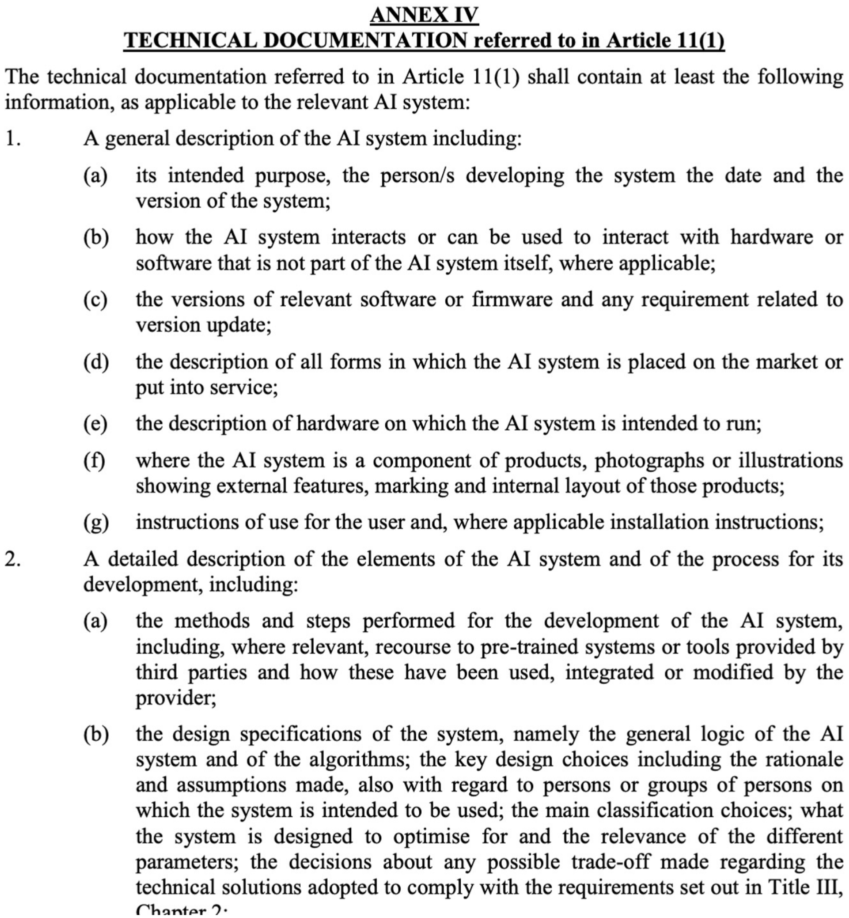
5. Access to and enjoyment of essential private services and public services and benefits:

   (a) AI systems intended to be used by public authorities or on behalf of public authorities to evaluate the eligibility of natural persons for public assistance benefits and services, as well as to grant, reduce, revoke, or reclaim such

AIRO Requirements

Technical Documentation

Annex IV	Required Info	Domain	Relation	Range
1(a)	System’s intended purpose	AISystem	hasPurpose	Purpose
	System’s developers	AISystem	isDevelopedBy	AIDeveloper
	System's date	AISystem	dcterms:date	
	System's version	AISystem	hasVersion	Version
…				
4	Risks of AI system	AISystem	hasRisk	Risk
	Sources of the risk	RiskSource	isRiskSourceFor	Risk
	Consequences of the risk	Risk	hasConsequence	Consequence
	Harmful impacts of risk	Consequence	hasImpact	Impact
	Probability of risk	Risk	hasLikelihood	Likelihood
	Severity of impact	Impact	hasSeverity	Severity
	…			



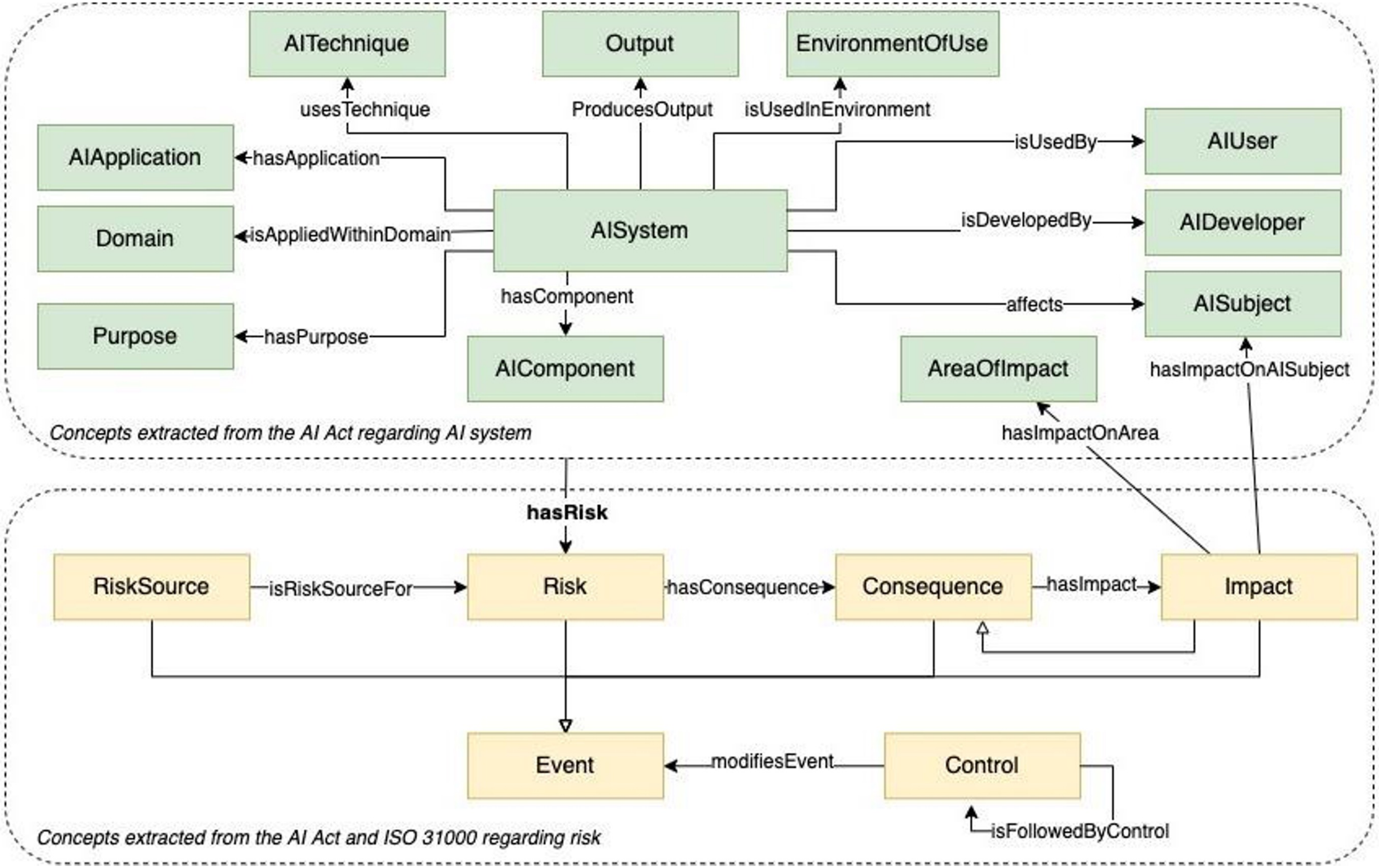
ANNEX IV

TECHNICAL DOCUMENTATION referred to in Article 11(1)

The technical documentation referred to in Article 11(1) shall contain at least the following information, as applicable to the relevant AI system:

1. A general description of the AI system including:
   (a) its intended purpose, the person/s developing the system the date and the version of the system;
   (b) how the AI system interacts or can be used to interact with hardware or software that is not part of the AI system itself, where applicable;
   (c) the versions of relevant software or firmware and any requirement related to version update;
   (d) the description of all forms in which the AI system is placed on the market or put into service;
   (e) the description of hardware on which the AI system is intended to run;
   (f) where the AI system is a component of products, photographs or illustrations showing external features, marking and internal layout of those products;
   (g) instructions of use for the user and, where applicable installation instructions;
2. A detailed description of the elements of the AI system and of the process for its development, including:
   (a) the methods and steps performed for the development of the AI system, including, where relevant, recourse to pre-trained systems or tools provided by third parties and how these have been used, integrated or modified by the provider;
   (b) the design specifications of the system, namely the general logic of the AI system and of the algorithms; the key design choices including the rationale and assumptions made, also with regard to persons or groups of persons on which the system is intended to be used; the main classification choices; what the system is designed to optimise for and the relevance of the different parameters; the decisions about any possible trade-off made regarding the technical solutions adopted to comply with the requirements set out in Title III, Chapter 2;

AIRO https://w3id.org/AIRO



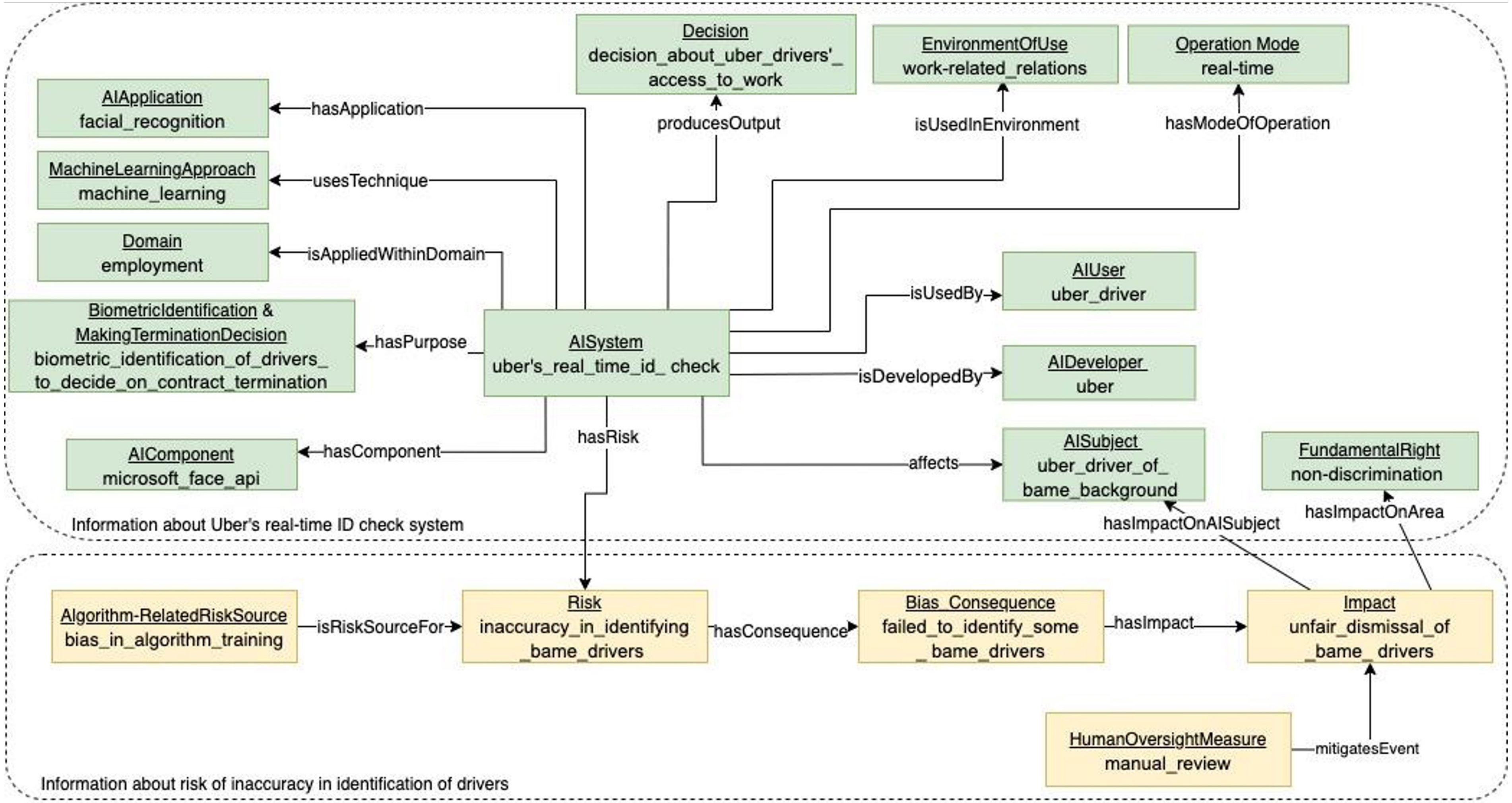
Use-cases

Use-case #1: Uber’s Real-time ID Check System
Purpose: Ensure the system is used by the registered driver
Main issue: Discrimination against drivers of BAME background

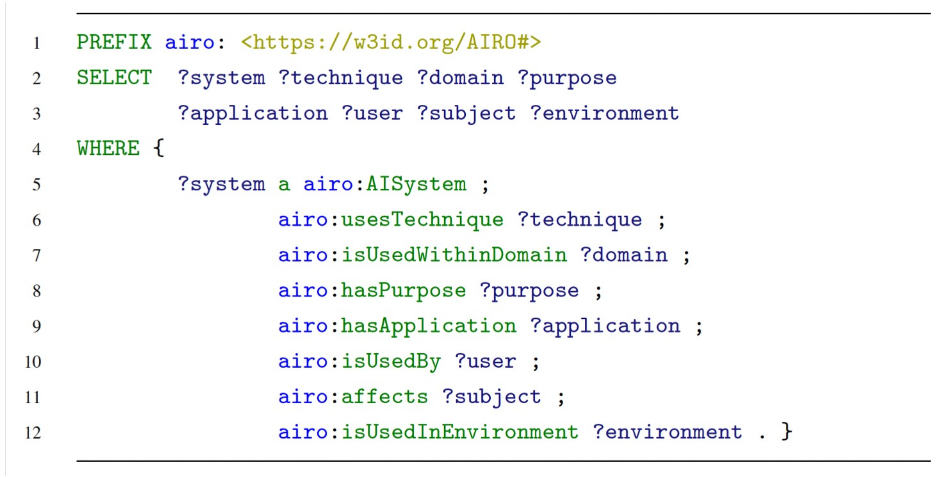
Use-case #2: VioGen Domestic Violence System
Purpose: Determine the eligibility to access police protection by predicting the likelihood of a victim of gender violence to be assaulted by the same perpetrator again
Main issue: Inaccuracy of predictions

Icons from https://www.flaticon.com/

Use-case #1: Uber's Real-time ID Check System

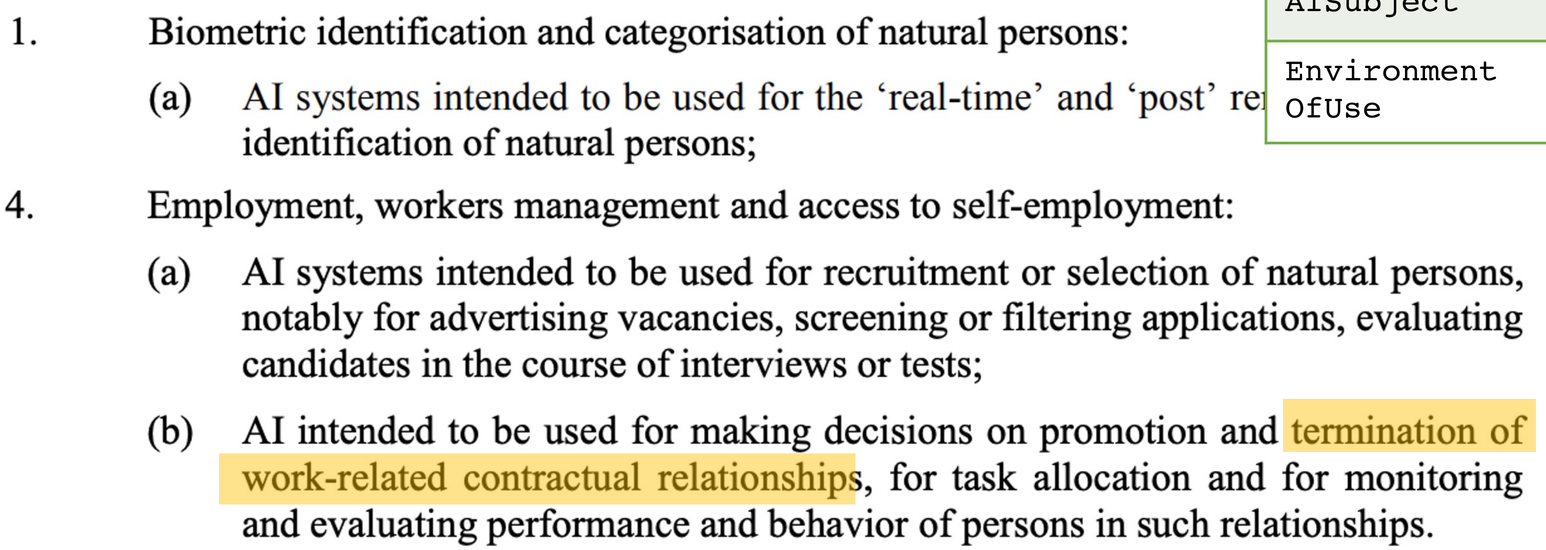


Identification of High-Risk AI Systems



```
PREFIX airo: <https://w3id.org/AIRO#>
SELECT  ?system ?technique ?domain ?purpose
        ?application ?user ?subject ?environment
WHERE {
        ?system a airo:AISystem ;
                airo:usesTechnique ?technique ;
                airo:isUsedWithinDomain ?domain ;
                airo:hasPurpose ?purpose ;
                airo:hasApplication ?application ;
                airo:isUsedBy ?user ;
                airo:affects ?subject ;
                airo:isUsedInEnvironment ?environment . }
```

AIRO concept	
AISystem	uber’s real time id check
AITechnique	machine learning techniques
Domain	employment
Purpose	biometric identification of drivers to decide on contract termination
AIApplication	facial recognition
AIUser	uber driver
AISubject	uber driver of bame background
Environment OfUse	work related relations



1. Biometric identification and categorisation of natural persons:

   (a) AI systems intended to be used for the ‘real-time’ and ‘post’ re identification of natural persons;

4. Employment, workers management and access to self-employment:

   (a) AI systems intended to be used for recruitment or selection of natural persons, notably for advertising vacancies, screening or filtering applications, evaluating candidates in the course of interviews or tests;

   (b) AI intended to be used for making decisions on promotion and termination of work-related contractual relationships, for task allocation and for monitoring and evaluating performance and behavior of persons in such relationships.

- Manual analysis

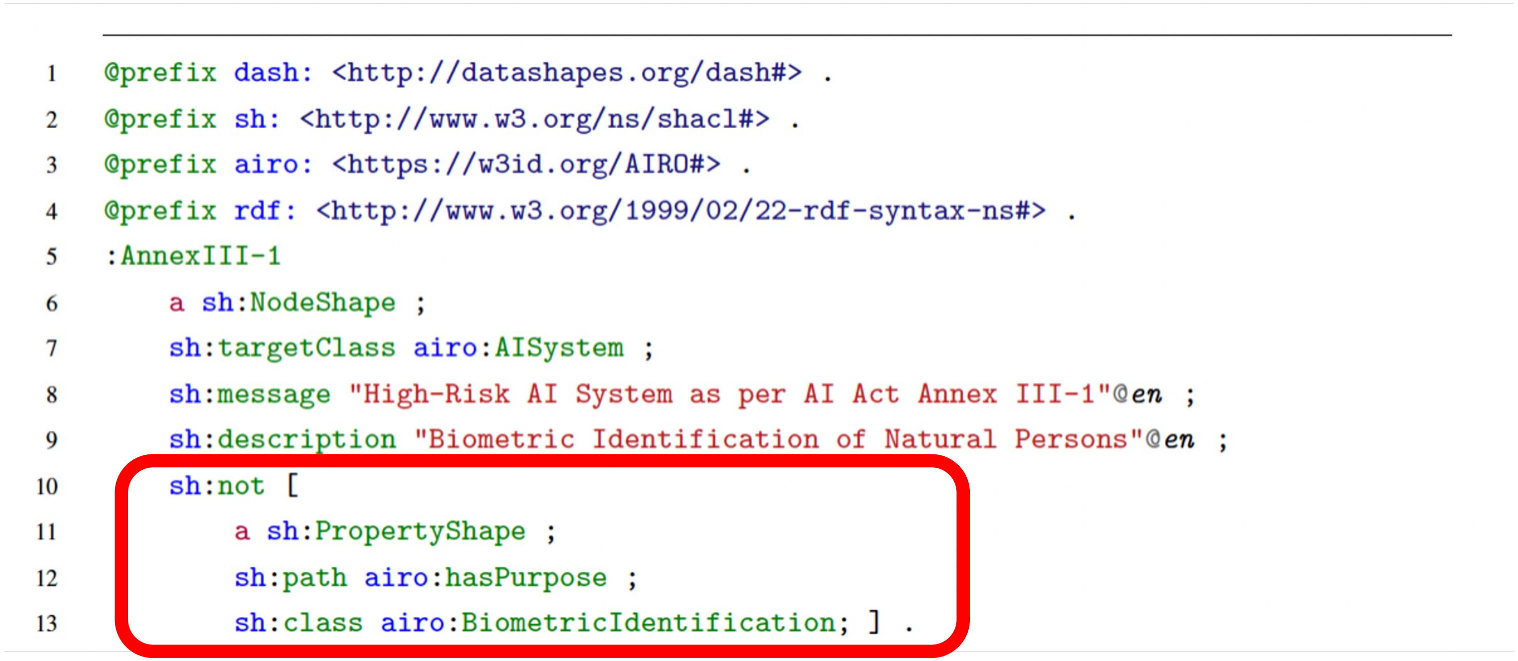
High Risk

SHACL Shapes for Automatic Identification of High-Risk AI

- ”Rules” to determine whether AI satisfies conditions for being “high-risk”
- Choose your favourite flavour of rule languages & mechanisms

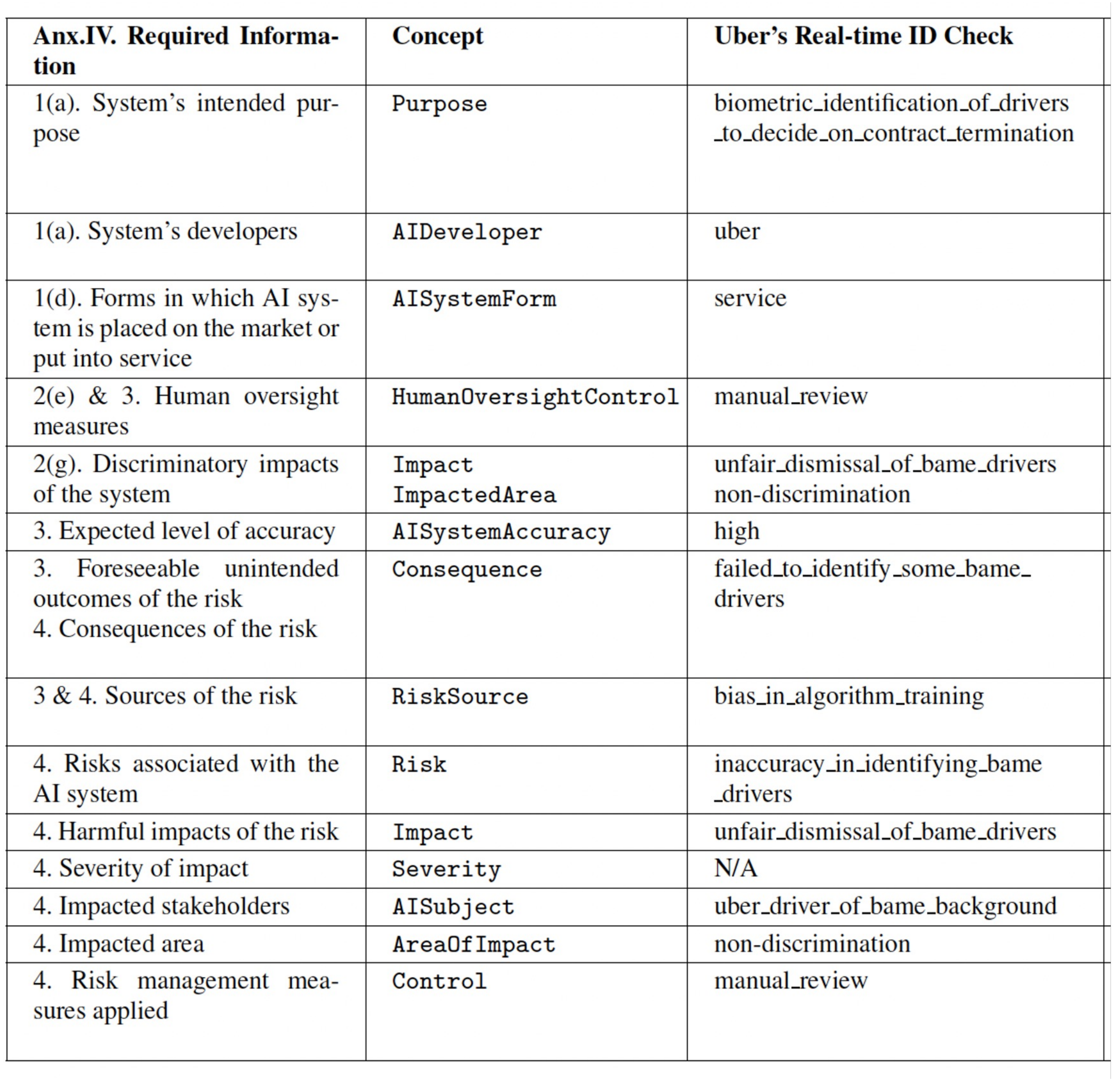
- We chose **SHACL**
- Why:
  - Flexible, Standardised
  - Extensible with plugins/features
  - Built-in documentation of outputs
  - Integrate to instead check outputs e.g. another rule engine

- We implement SHACL shapes for clauses defined in Annex III that determine high-risk
- Validation is to NOT satisfy the expressed criteria



```
@prefix dash: <http://datashapes.org/dash#> .
@prefix sh: <http://www.w3.org/ns/shacl#> .
@prefix airo: <https://w3id.org/AIRO#> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
:AnnexIII-1
    a sh:NodeShape ;
    sh:targetClass airo:AISystem ;
    sh:message "High-Risk AI System as per AI Act Annex III-1"@en ;
    sh:description "Biometric Identification of Natural Persons"@en ;
    sh:not [
        a sh:PropertyShape ;
        sh:path airo:hasPurpose ;
        sh:class airo:BiometricIdentification; ] .
```

Generating Technical Documentation



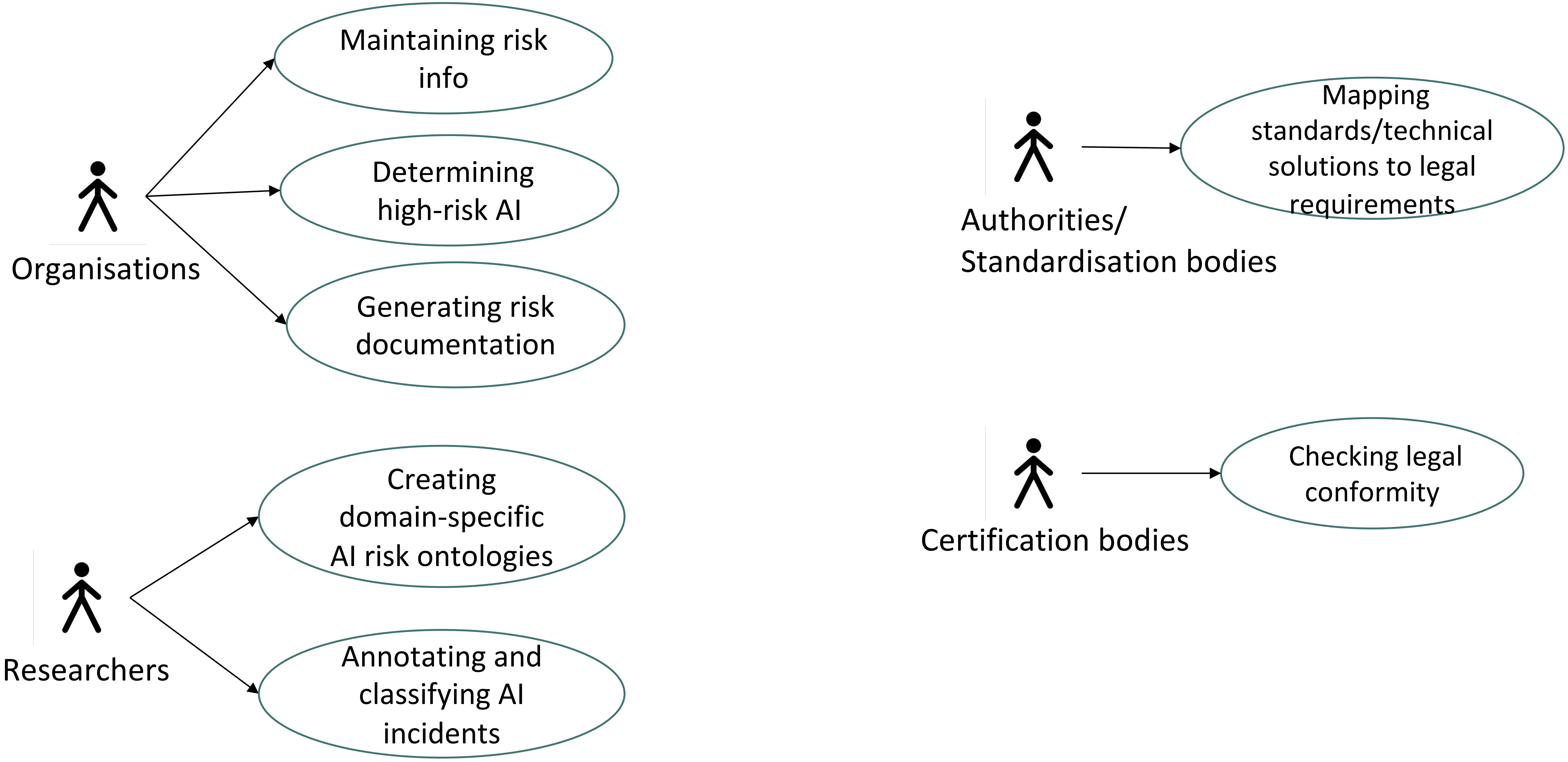
Anx.IV. Required Information	Concept	Uber's Real-time ID Check
1(a). System's intended purpose	Purpose	biometric_identification_of_drivers_to_decide_on_contract_termination
1(a). System's developers	AIDeveloper	uber
1(d). Forms in which AI system is placed on the market or put into service	AISystemForm	service
2(e) & 3. Human oversight measures	HumanOversightControl	manual_review
2(g). Discriminatory impacts of the system	Impact ImpactedArea	unfair_dismissal_of_bame_drivers non-discrimination
3. Expected level of accuracy	AISystemAccuracy	high
3. Foreseeable unintended outcomes of the risk 4. Consequences of the risk	Consequence	failed_to_identify_some_bame_drivers
3 & 4. Sources of the risk	RiskSource	bias_in_algorithm_training
4. Risks associated with the AI system	Risk	inaccuracy_in_identifying_bame_drivers
4. Harmful impacts of the risk	Impact	unfair_dismissal_of_bame_drivers
4. Severity of impact	Severity	N/A
4. Impacted stakeholders	AISubject	uber_driver_of_bame_background
4. Impacted area	AreaOfImpact	non-discrimination
4. Risk management measures applied	Control	manual_review

Domain Challenge

the incident reports do not provide *detailed information*

Why:
Implementation Details

Benefit to Stakeholders



Organisations

Maintaining risk info

Determining high-risk AI

Generating risk documentation

Researchers

Creating domain-specific AI risk ontologies

Annotating and classifying AI incidents

Authorities/ Standardisation bodies

Mapping standards/technical solutions to legal requirements

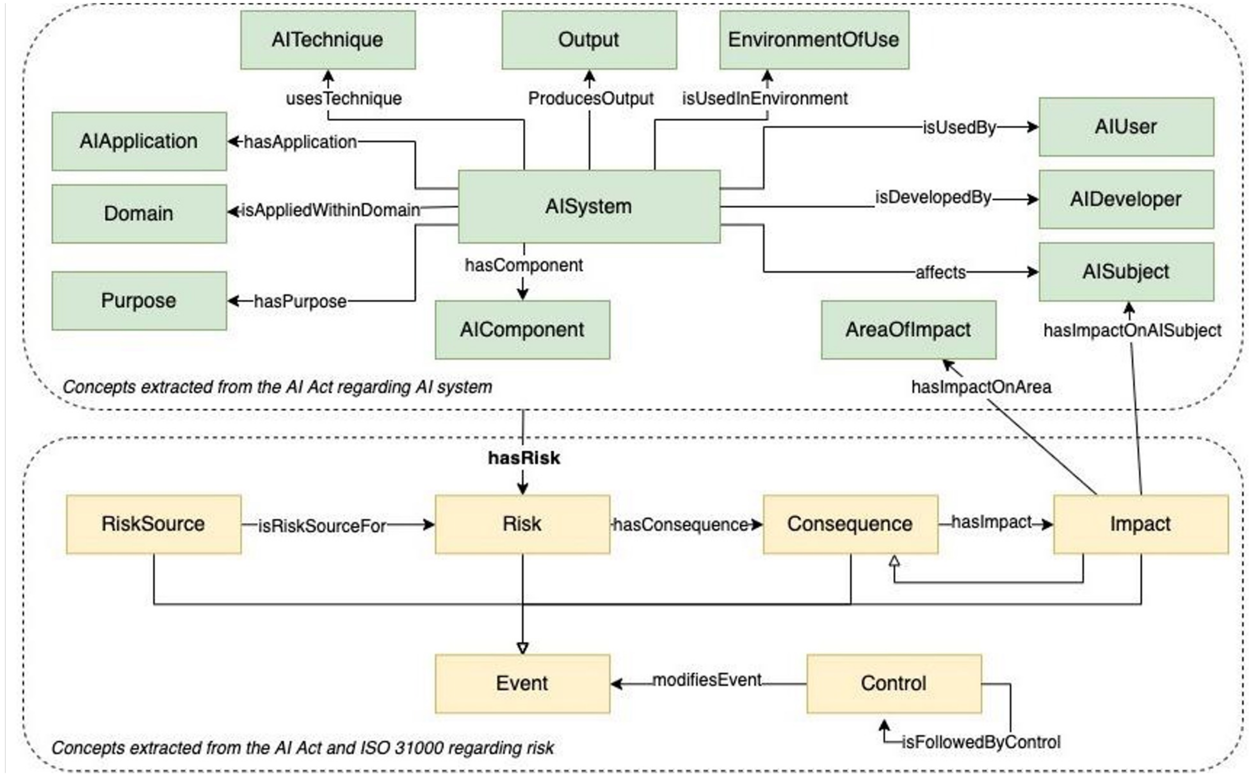
Certification bodies

Checking legal conformity

Future Work

- Enhance AIRO to:
  - represent known categories of AI risks identified from real-world incidents
  - express provenance of AI risk and impact assessments

- Incorporate changes from the AI Act update and recently developed ISO standards
- Create rules for determining High-Risk AI
- Develop tools for risk documentation generation and sharing
- Apply AIRO's AI impact assessment for the GDPR's DPIA



AIRO: an Ontology for Representing AI Risks based on the Proposed EU AI Act and ISO Risk Management Standards

Delaram Golpayegani,
Harshvardhan J. Pandit, Dave Lewis

Email: sgolpays@tcd.ie

Ontology: https://w3id.org/AIRO

GitHub: https://github.com/delaramglp/AIRO