AI Safety Regulations and Requirements

FIMA - FCAI WORKSHOP: AI for Mobile Work Machines

Eetu Heikkilä

24.11.2021 **VTT – beyond the obvious**

Introduction

- AI is a key technology in enabling increasingly intelligent and autonomous mobile machinery.
- This presentation provides an overview of some emerging AI safety related regulations and requirements:
  - Upcoming EU regulation
  - Industry standards

AI Safety research

- Safety of AI is studied (publicly) in a surprisingly small scale. Few disciplines seem to produce most of the research.

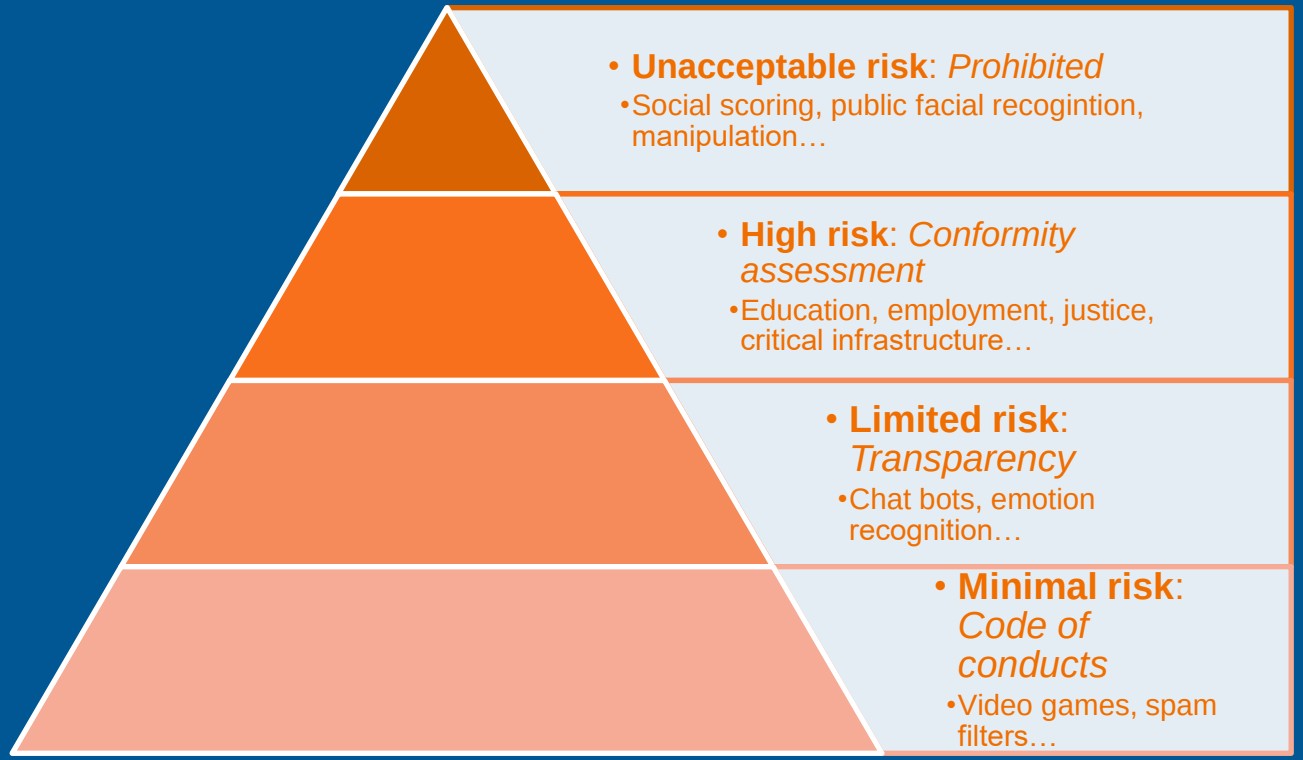
General AI safety, incl. AGI, superintelligence, technological singularity	AI in medical sciences and healthcare	AI in autonomous vehicles
• General safety issues, studied mostly at very high level of abstraction • Considerations of superintelligence • Close link to ethics and cybersecurity	• Effect on patient safety • Safety in supportive tasks (diagnosis, drug safety)	• Object detection, decision-making • Testing & validation

European Regulation

EU Regulation: Artificial Intelligence Act

- Regulation proposal: Laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts
- Definition of AI is wide!
  - *Machine learning approaches, including supervised, unsupervised and reinforcement learning, using a wide variety of methods including deep learning;*
  - *Logic- and knowledge-based approaches, including knowledge representation, inductive (logic) programming, knowledge bases, inference/deductive engines, (symbolic) reasoning and expert systems;*
  - *Statistical approaches, Bayesian estimation, search and optimization methods.*

AI Act's hierarchy of risk levels

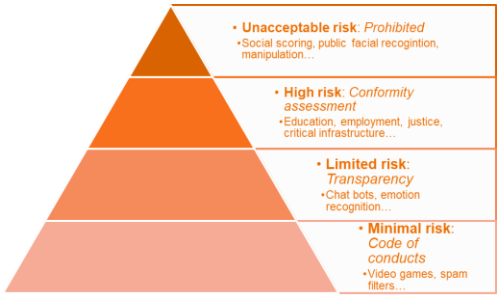


Based on Gaumond, E. 2021. Artificial Intelligence Act: What Is the European Approach for AI?

Unacceptable risk systems are prohibited

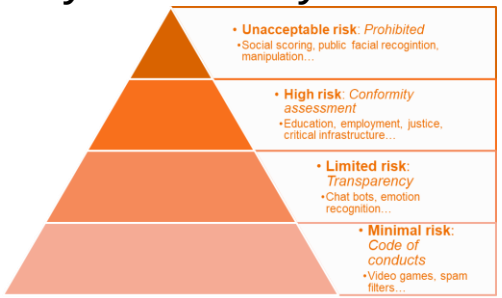
The following list of artificial intelligence practices are **prohibited** as contravening the Union values or violating fundamental rights protected under Union law:

- AI system that distort a person's behaviour, cause **physical or psychological harm**;
- AI system that **exploits vulnerabilities** of a specific group **of persons**; cause physical or psychological harm;
- AI systems for the evaluation or classification of the **trustworthiness** of natural **persons**
- AI for use of 'real-time' remote **biometric identification** systems in publicly accessible spaces (exceptions)



High risk systems

- In the AI Act proposal, **machinery** are considered as **high-risk systems**
- Other systems considered high risk include e.g.: Biometric identification and categorisation of natural persons, employment, workers management
- AI systems considered high-risk if the product in question undergoes the conformity assessment procedure with a third-party conformity assessment body.
- High-risk AI systems should bear the CE marking

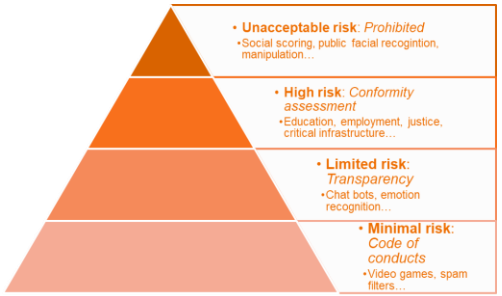


High risk systems

Some requirements:

- High-risk AI systems shall be designed and developed to ensure that their operation is sufficiently transparent to enable users to **understand and control how the high-risk AI** system produces its output.
- Quality and risk management systems

However, many statements are quite vague / unclear…



EU regulation on machinery products (replaces Machinery Directive)

(45) For example, software ensuring safety functions of machinery based on **artificial intelligence**, embedded or not in the machinery product, should be classified as a **high-risk machinery product** due to the characteristics of artificial intelligence such as data dependency, opacity, autonomy and connectivity, which might increase very much the probability and severity of harm and seriously affect the safety of the machinery product. Furthermore, the market for software ensuring safety functions of machinery products based on artificial intelligence is so far very small, which results in a lack of experience and data. Therefore, the **conformity assessment** of software ensuring safety functions **based on artificial intelligence should be carried out by a third party.**

Industry standards

ISO documents available (JTC 1/SC 42)

STANDARD AND/OR PROJECT UNDER THE DIRECT RESPONSIBILITY OF ISO/IEC JTC 1/SC 42 SECRETARIAT (9)	STAGE
ISO/IEC 20546:2019 Information technology — Big data — Overview and vocabulary	60.60
ISO/IEC TR 20547-1:2020 Information technology — Big data reference architecture — Part 1: Framework and application process	60.60
ISO/IEC TR 20547-2:2018 Information technology — Big data reference architecture — Part 2: Use cases and derived requirements	60.60
ISO/IEC 20547-3:2020 Information technology — Big data reference architecture — Part 3: Reference architecture	60.60
ISO/IEC TR 20547-5:2018 Information technology — Big data reference architecture — Part 5: Standards roadmap	60.60
ISO/IEC TR 24027:2021 Information technology — Artificial intelligence (AI) — Bias in AI systems and AI aided decision making	60.60
ISO/IEC TR 24028:2020 Information technology — Artificial intelligence — Overview of trustworthiness in artificial intelligence	60.60
ISO/IEC TR 24029-1:2021 Artificial Intelligence (AI) — Assessment of the robustness of neural networks — Part 1: Overview	60.60
ISO/IEC TR 24030:2021 Information technology — Artificial intelligence (AI) — Use cases	90.92

ISO documents being prepared (JTC 1/SC 42)

STANDARD AND/OR PROJECT UNDER THE DIRECT RESPONSIBILITY OF ISO/IEC JTC 1/SC 42 SECRETARIAT (23)	STAGE	ICS
ISO/IEC DTS 4213 Information technology — Artificial Intelligence — Assessment of machine learning classification performance	30.60	35.020
ISO/IEC AWI 5259-1 Artificial intelligence — Data quality for analytics and machine learning (ML) — Part 1: Overview, terminology, and examples	20.00	
ISO/IEC AWI 5259-2 Artificial intelligence — Data quality for analytics and machine learning (ML) — Part 2: Data quality measures	20.00	
ISO/IEC AWI 5259-3 Artificial intelligence — Data quality for analytics and machine learning (ML) — Part 3: Data quality management requirements and guidelines	20.00	
ISO/IEC AWI 5259-4 Artificial intelligence — Data quality for analytics and machine learning (ML) — Part 4: Data quality process framework	20.00	
ISO/IEC AWI 5338 Information technology — Artificial intelligence — AI system life cycle processes	20.00	
ISO/IEC AWI 5339 Information Technology — Artificial Intelligence — Guidelines for AI applications	20.00	
ISO/IEC AWI 5392 Information technology — Artificial intelligence — Reference architecture of knowledge engineering	20.00	
ISO/IEC AWI TR 5469 Artificial intelligence — Functional safety and AI systems	10.99	
ISO/IEC AWI TS 5471 Artificial intelligence — Quality evaluation guidelines for AI systems	20.00	
ISO/IEC AWI TS 6254 Information technology — Artificial intelligence — Objectives and approaches for explainability of ML models and AI systems	20.00	
ISO/IEC AWI TS 8200 Information technology — Artificial intelligence — Controllability of automated artificial intelligence systems	20.00	
ISO/IEC DIS 22989 Information technology — Artificial intelligence — Artificial intelligence concepts and terminology	40.60	35.020 01.040.35
ISO/IEC DIS 23053 Framework for Artificial Intelligence (AI) Systems Using Machine Learning (ML)	40.60	35.020
ISO/IEC DIS 23894 Information technology — Artificial intelligence — Risk management	40.00	35.020
ISO/IEC CD 24029-2 Artificial intelligence (AI) — Assessment of the robustness of neural networks — Part 2: Methodology for the use of formal methods	30.20	35.020
ISO/IEC AWI TR 24030 Information technology — Artificial intelligence (AI) — Use cases	20.00	
ISO/IEC DTR 24368 Information technology — Artificial intelligence — Overview of ethical and societal concerns	30.60	35.020
ISO/IEC TR 24372 Information technology — Artificial intelligence (AI) — Overview of computational approaches for AI systems	60.00	35.020
ISO/IEC DIS 24668 Information technology — Artificial intelligence — Process management framework for big data analytics	40.20	35.020
ISO/IEC CD 25059 Software engineering — Systems and software Quality Requirements and Evaluation (SQuaRE) — Quality model for AI systems	30.20	35.080
ISO/IEC DIS 38507 Information technology — Governance of IT — Governance implications of the use of artificial intelligence by organizations	40.60	35.020
ISO/IEC CD 42001 Information Technology — Artificial intelligence — Management system	30.60	35.020 03.100.70

ISO/TR 22100-5:2021. Safety of machinery — Relationship with ISO 12100 — Part 5: Implications of artificial intelligence machine learning

- This document suggests that risk(s) introduced by AI in machinery applications can be addressed by the methodology for risk assessment and risk reduction as prescribed in ISO 12100 where risks of the AI are addressed according to the intended use and use limits (predetermined boundaries) specified by the machine manufacturer.

Conclusions

- Safety of AI systems is a key issue in development of increasingly intelligent and autonomous mobile machinery.
- AI regulations and standards are emerging. Their development needs to be closely followed to ensure that systems can enter market.

VTT

beyond

the obvious

Eetu Heikkilä
eetu.heikkila@vtt.fi
+358 40 849 5790

@VTTFinland

www.vtt.fi