Trustworthy AI – an international initiative towards mindful, holistic, and inclusive validation of Artificial Intelligence (AI)

https://trustworthyai-lab.icube.unistra.fr/

Université de Strasbourg

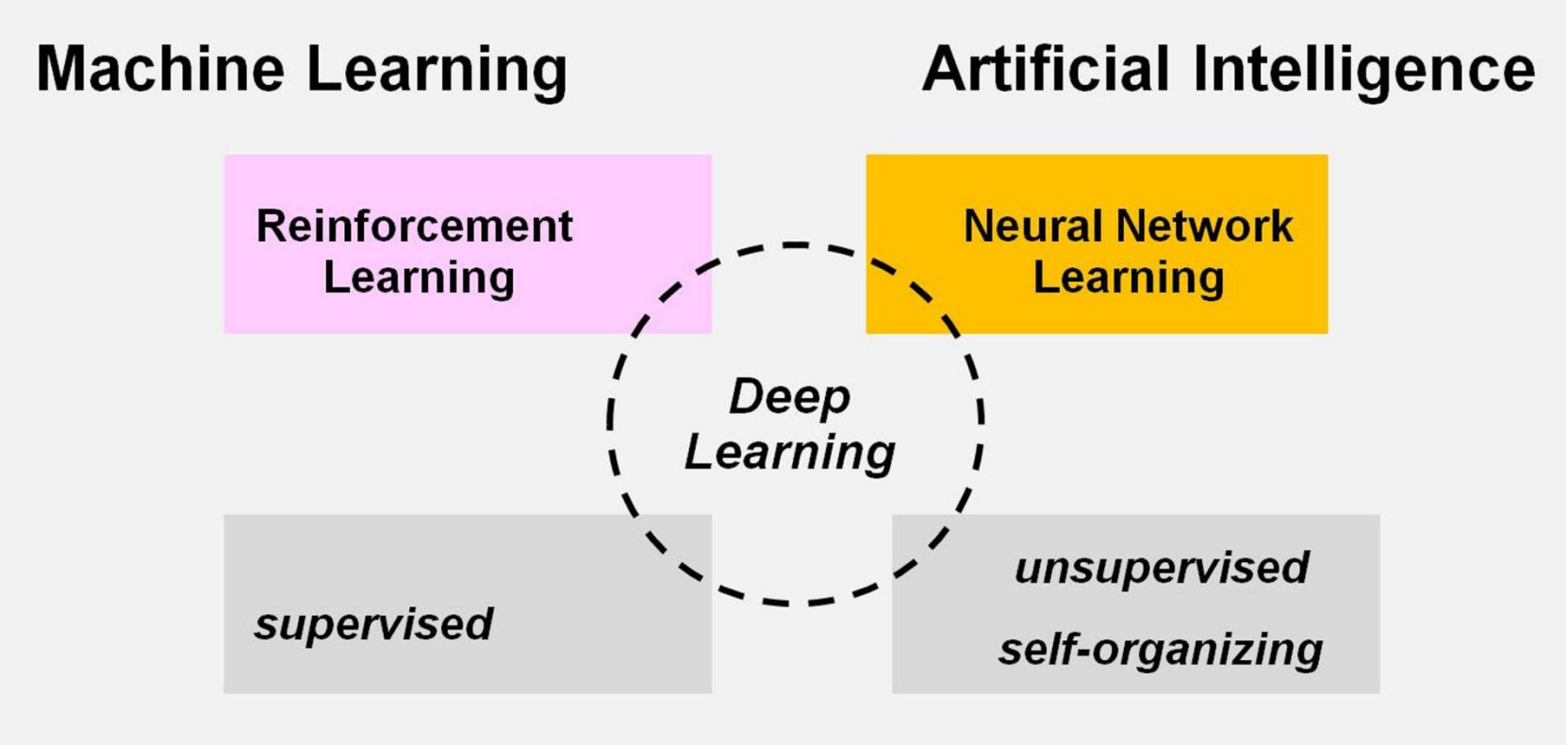
- What is Artifical Intelligence? A brief summary of definitions, methods of implementation, levels of autonomy, human *versus* machine
- The Assessment List for Trustworthy AI (ALTAI) of the European Commission: rules for due procedure, but who assesses what, where, when?
- The European AI Act: towards an international regulatory framework for the development and deployment of trustworthy AI to the service of citizens: progress, loopholes, unresolved issues
- Ethics of AI *versus* Legal Framework for AI: perspectives, limitations and barriers
- Z-inspection: an international initiative towards mindful, holistic, and inclusive validation of AI despite limitations and barriers
- The #first case study on AI-assisted emergency call management during the covid-19 pandemic at the Copenhagen City Hospital
- Trustworthy AI for medicine and surgery: field-specific barriers to developing inclusive procedures for assessment?
- Conclusions followed by Questions & Answers

Université de Strasbourg

Artificial Intelligence

Synthetic system, in principle but not systematically inspired by the functional properties of a living intelligent system, conceived by humans to learn information in an unsupervised way, ideally respecting the laws of generalization and consolidation of intelligent learning, thereby becoming able to process new information with a view to an interpretation, triggering or not a decision.

Methods of Implementation



Levels of Autonomy

Level 1 Human controlled - *'human on the loop'*: human agent initiates and controls all steps in the process

Level 2 Semi-autonomous - *'human in the loop'*: human has control over some of the steps in the process

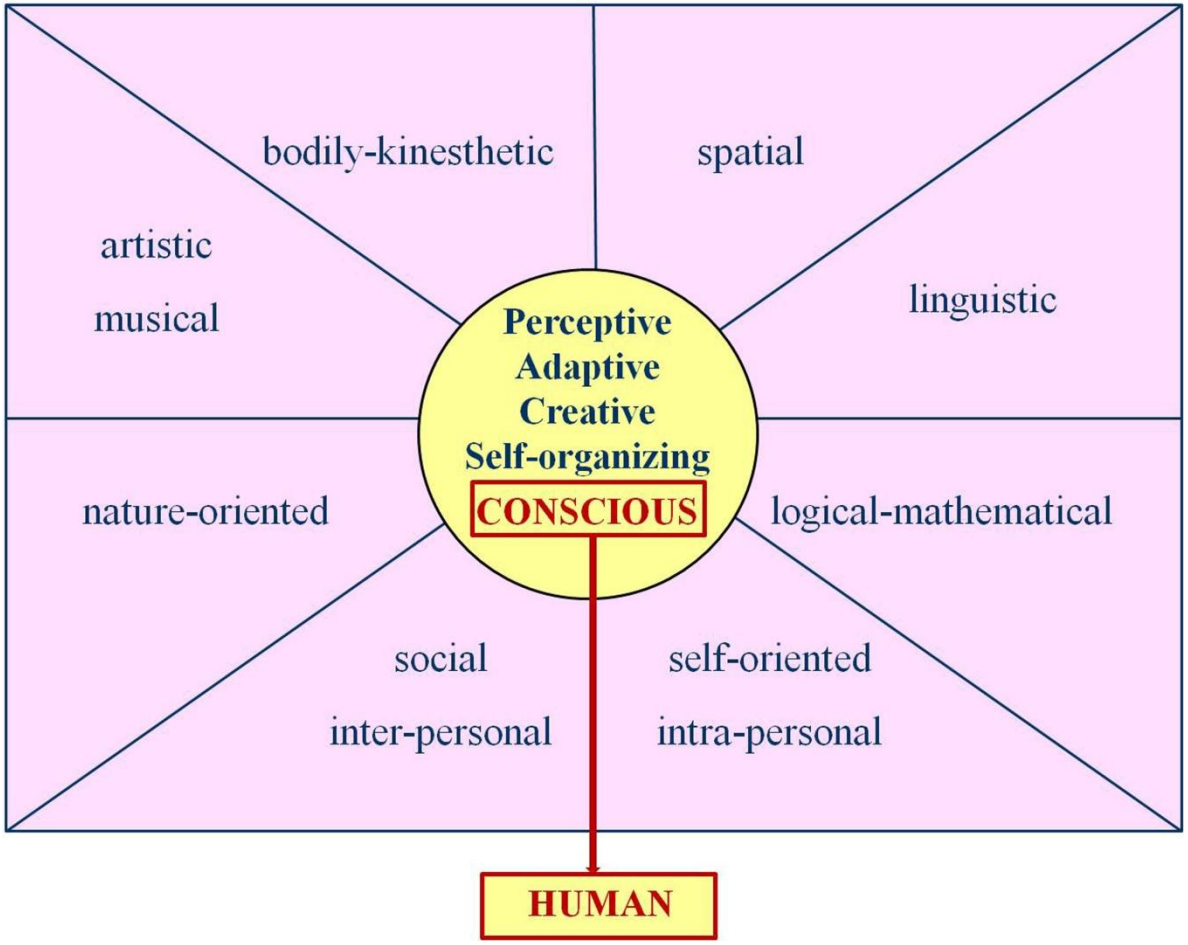
Level 3 Fully autonomous - *'no human in the loop'*: human has no control over any step in the process

Boulanin and Verbruggen. *Mapping the development of autonomy in weapon systems.* **2017;** The Stockholm International Peace Research Institute.

Université de Strasbourg

H. Gardner. *Frames of Mind : The Theory of Multiple Intelligence*, **1983;** Basic Books, New York.

Human *versus* Machine Intelligence



What about intuition?

Université de Strasbourg

Assessment List for Trustworthy AI (ALTAI)

The European Commission, April **2019**

A result of two years of discussions between 52 members of the **High Level Expert Group on AI** (**HLEGAI**) and 350 organizations from EU member states

7 criteria for self-assessing the trustworthiness of AI

- to help organizations identify risks generated by AI
- help identify measures to avoid or minimize such risks
- specify the conditions for "trustworthiness"
- promote their being put in practice through internal guidelines or governance processes

Université de Strasbourg

The **ALTAI explicitly states** that: "Prior to self-assessing an AI system under the seven requirements stated in the Assessment List, a **Fundamental Rights Impact Assessment (FRIA)** should be performed" to ensure that AI respects human dignity - AI must not:

- negatively discriminate people on the basis of sex, race, color, ethnic or social origin, genetic features, language, religion or belief, political or any other opinion, membership of a national minority, property, birth, disability, age or sexual orientation
- violate the right of the child to protection from harm of any kind
- violate personal data protection rights
- violate a person's or group's right to freedom of expression and information

1a. Human Agency

- are end-users adequately made aware that a decision, content, advice, or outcome is the result of an algorithmic decision?
- are they informed that they are interacting with an AI system?
- is made sure that end-users do not over-rely on the AI system?
- the AI system does not affect human autonomy?
- are negative consequences for end-users minimized in case they develop a disproportionate attachment to the AI System?

minimized risk of addiction?
minimized risk of manipulation?

1b. Human Oversight

- have the humans (human-in/on-the-loop) been given specific training to exercise oversight?
- are there detection and response mechanisms for undesirable adverse effects of the AI system for the end-user or subject?
- is there a 'stop button' or procedure to safely abort an operation?
- are there specific oversight and control measures regarding the self-organizing or autonomous nature of the AI system?

2. Technical Robustness/Safety

- has the AI system been tested for resilience to attacks and security?
- have general safety risks of the AI system been assessed for each specific use case?
- is the expected level of accuracy of the AI system constant and invariant in time (consistently reliable)?
- was tested whether specific contexts or conditions need to be fulfilled to ensure the reliability of the AI system?

3. Privacy/Data Governance

- does the AI system comply with General Data Protection Regulation (GDPR) or a non-European equivalent?
- was a Data Protection Officer designated and included in the development, procurement, or deployment of the AI system?
- are there oversight mechanisms for data processing such as limiting access to qualified personnel, mechanisms for logging data access and making modifications?
- measures to achieve privacy-by-design and default encryption, pseudonymisation, aggregation, anonymisation?

4. Transparency

- are there measures that ensure traceability of the AI system during its entire lifecycle?
- are the decision(s) of the AI system explained to the users and is it ensured they understand the decision(s) of the AI system?
- do you provide appropriate communication and training material to users on how to adequately use the AI system?

5. Diversity, Non-Discrimination, Fairness

- are potential biases in the data learnt or generated by the AI system tested for and monitored during the entire lifecycle?
- are accessibility and universal design ensured by taking into account the full range of abilities in society?
- is the participation of the widest range of possible stakeholders ensured in the AI system's design and development?

6. Environmental/Societal Well-Being

- environmental impact of the AI system's development or deployment, i.e. energy used and carbon emissions?
- measures to reduce the environmental impact of the AI system throughout its lifecycle?
- were individuals/workers and/or their representatives informed/consulted on the potential impact of the AI system on their work or well-being?
- could the AI system have a negative impact on society at large or democracy?

7. Accountability

- are there mechanisms to facilitate the AI system's auditability, the traceability of the development process, the sourcing of training data, and the logging of the AI system's processes, outcomes, and positive/negative impact?

- is external guidance or a third-party auditing to oversee ethical concerns and accountability measures ensured?

- is risk assessment, training, information taking into account the potential legal framework applicable to AI systems ensured?

The European AI Act

The European Commission, December **2022**

President von der Leyen's commitment to **legislation** for a coordinated European approach on the **human** and ethical implications of **AI**; three years of discussions between members of the **Expert Group on AI and Consumer/Citizen Safety** (a sub-group of HLEGAI), the **International Consumer Safety Network (CSN)**, and organizations from EU member states

- identify risks to the physical and/or psychological safety and well-being of individuals generated by AI

- Goal: carve out the **first legal framework** for single-market regulatory policies pertaining to the development and deployment of AI systems, taking into account the ALTAI

Université de Strasbourg

Objectives of the AI Act

1. Ensure that AI placed on the market is safe and respects human rights and current EU law	2. Impose legal certainty as a condition for investment and innovation in AI
3. Reinforce existing law on product safety and human rights applicable to AI	4. The first step towards a single EU market for lawful, human-centred, and trustworthy AI

Policy Options - towards a single EU market for AI

Option 1: EU legislative, voluntary labeling scheme

Option 2: sectoral *ad-hoc* approach

Option 3: horizontal EU legislative, proportionate risk-based approach

Option 3+: Horizontal EU legislative, proportionate risk-based approach + codes of conduct for low-risk AI

Option 4: Horizontal EU legislative, mandatory requirements irrespective of the risk category

Université de Strasbourg

Risk Categories

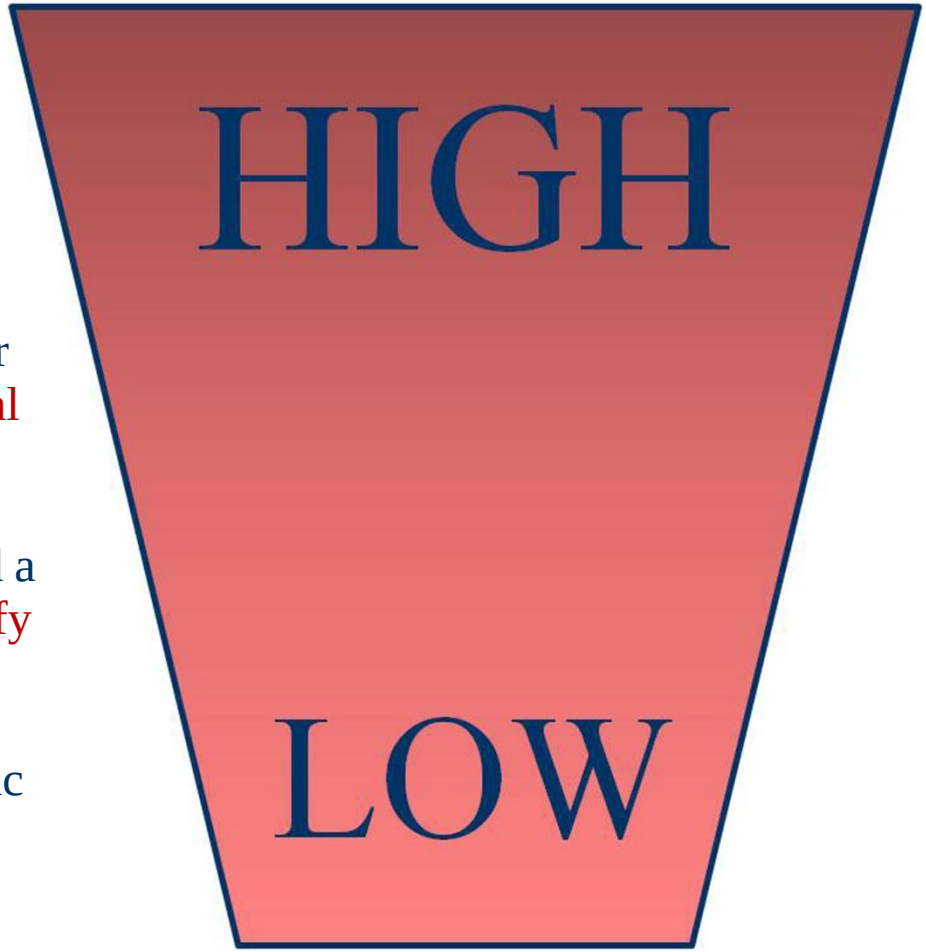
prohibited high-risk AI systems:

safety components subject to third party *ex-ante* conformity assessment and other stand-alone AI systems with fundamental rights implications

deploying subliminal techniques beyond a person's consciousness in order to modify a person's behaviour

exploiting the vulnerabilities of a specific group based on age, physical, or mental disability

'real-time' remote biometric identification in public spaces for law enforcement



Université de Strasbourg

Criteria

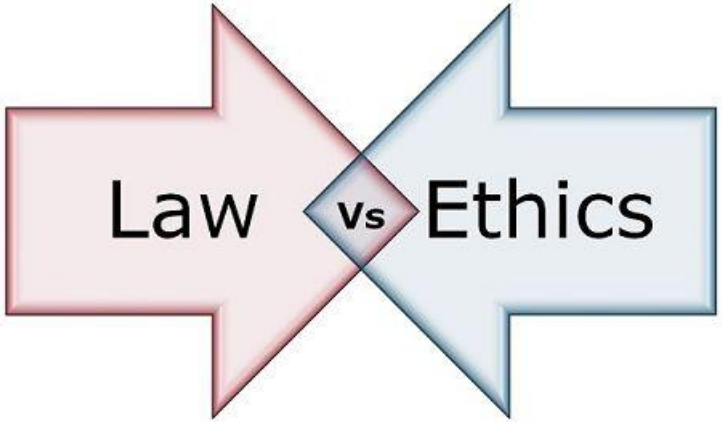
- compliance with existing law on safety and human rights
- human oversight
- quality of input and output data
- robustness/safety
- transparency/traceability
- accountability

ALTAI

Université de Strasbourg

Ethics *vs* Law Dilemma

AI Act



ALTAI

- rules and regulations enforced by government or authorities
- binding
- non-respect results in penalties

- moral principles adopted by people within a cultural context
- non-binding
- non-respect does not result in penalties

Université de Strasbourg