What is “Responsible Artificial Intelligence”?

Raja Chatila
Institute of Intelligent Systems and Robotics (ISIR)
Sorbonne University, Paris, France

Raja.Chatila@sorbonne-universite.fr

Current modern AI systems are mostly based on Machine Learning

- Statistical correlations of data to make predictions: search for regularities and similarities in data
- No reasoning using data and mathematical/physical theories, and understanding causes to build general models for making predictions

AI Based on Machine Learning

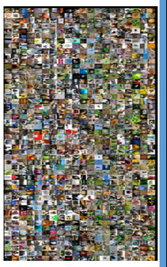
Statistical data processing and classification are the most successful methods

- Use of probability distributions, correlations, …
- Use of artificial neural nets as classifiers
- Optimisation algorithms

- Supervised learning: correct answer provided by a “tutor”.
- Unsupervised learning: search for regularities in the data
- Reinforcement Learning: select the most promising action based on optimising rewards

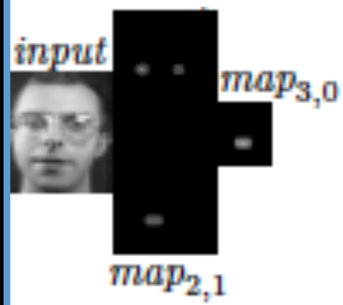
How machine learning works: Example of an AI System for Face Detection

DATA

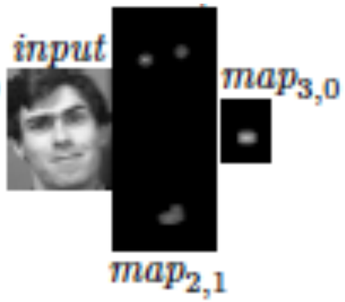


Machine Learning Black Box

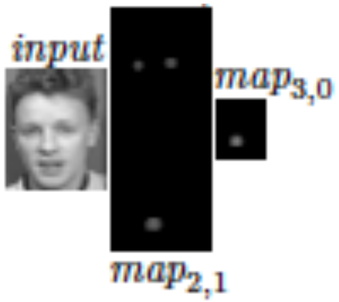
Results



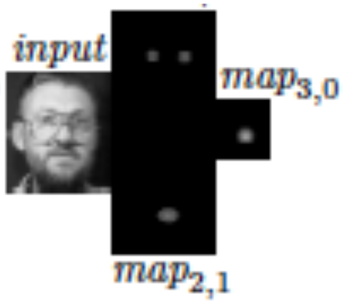
Face



Face

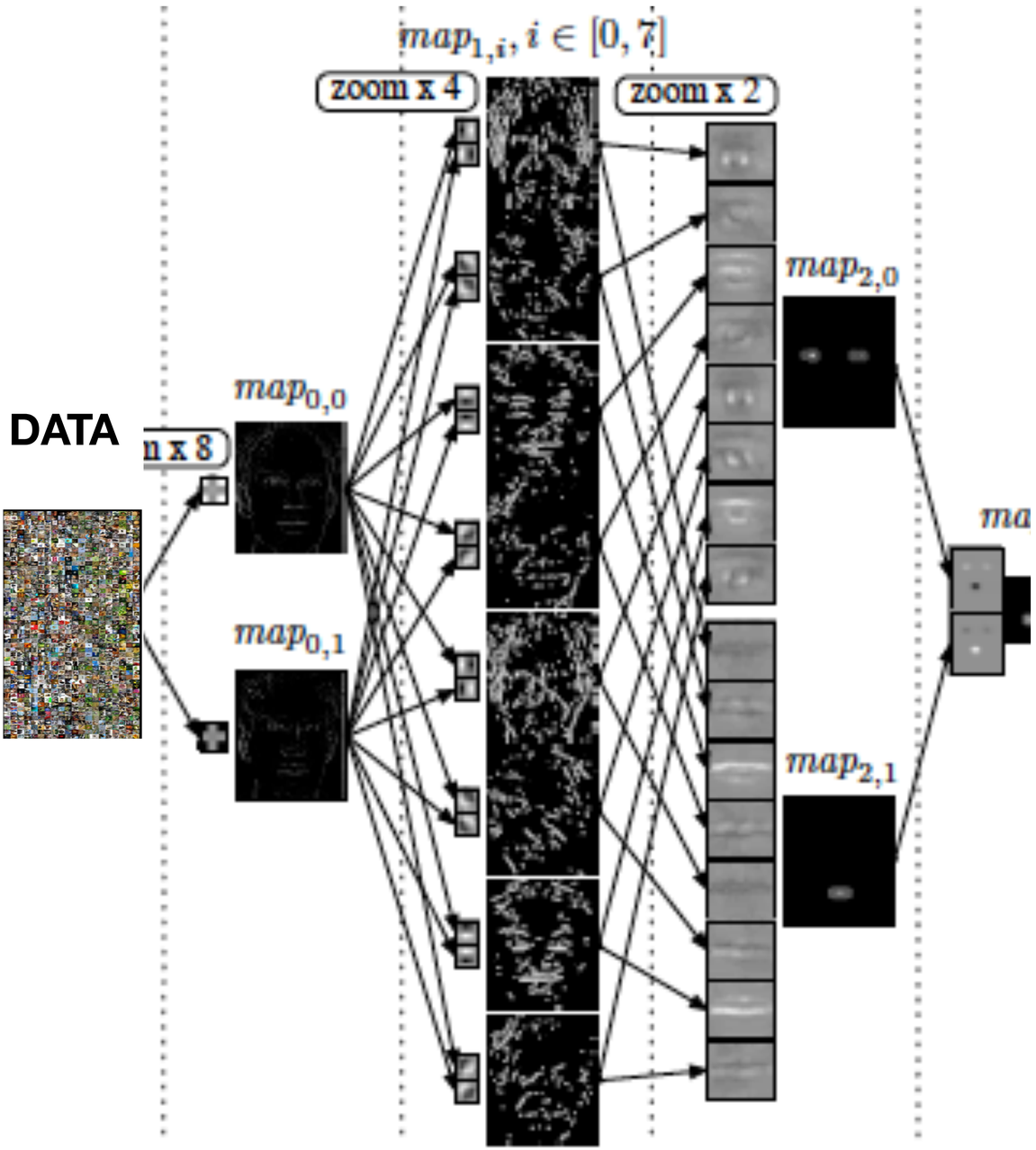


Face

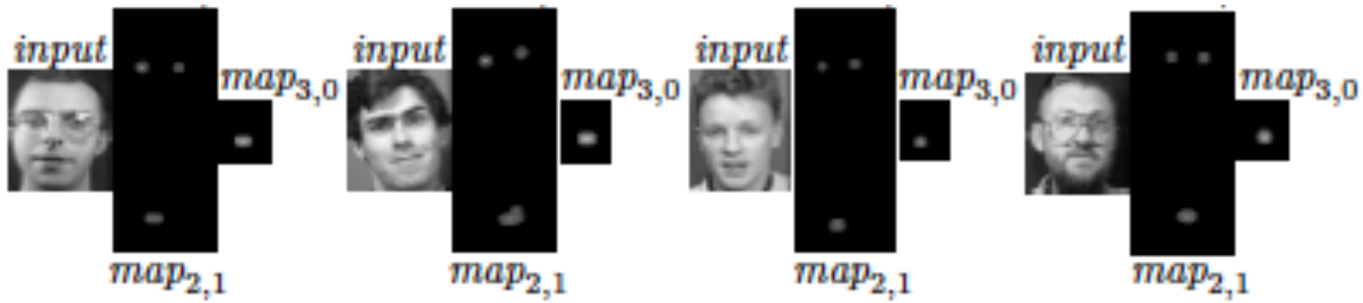


Face

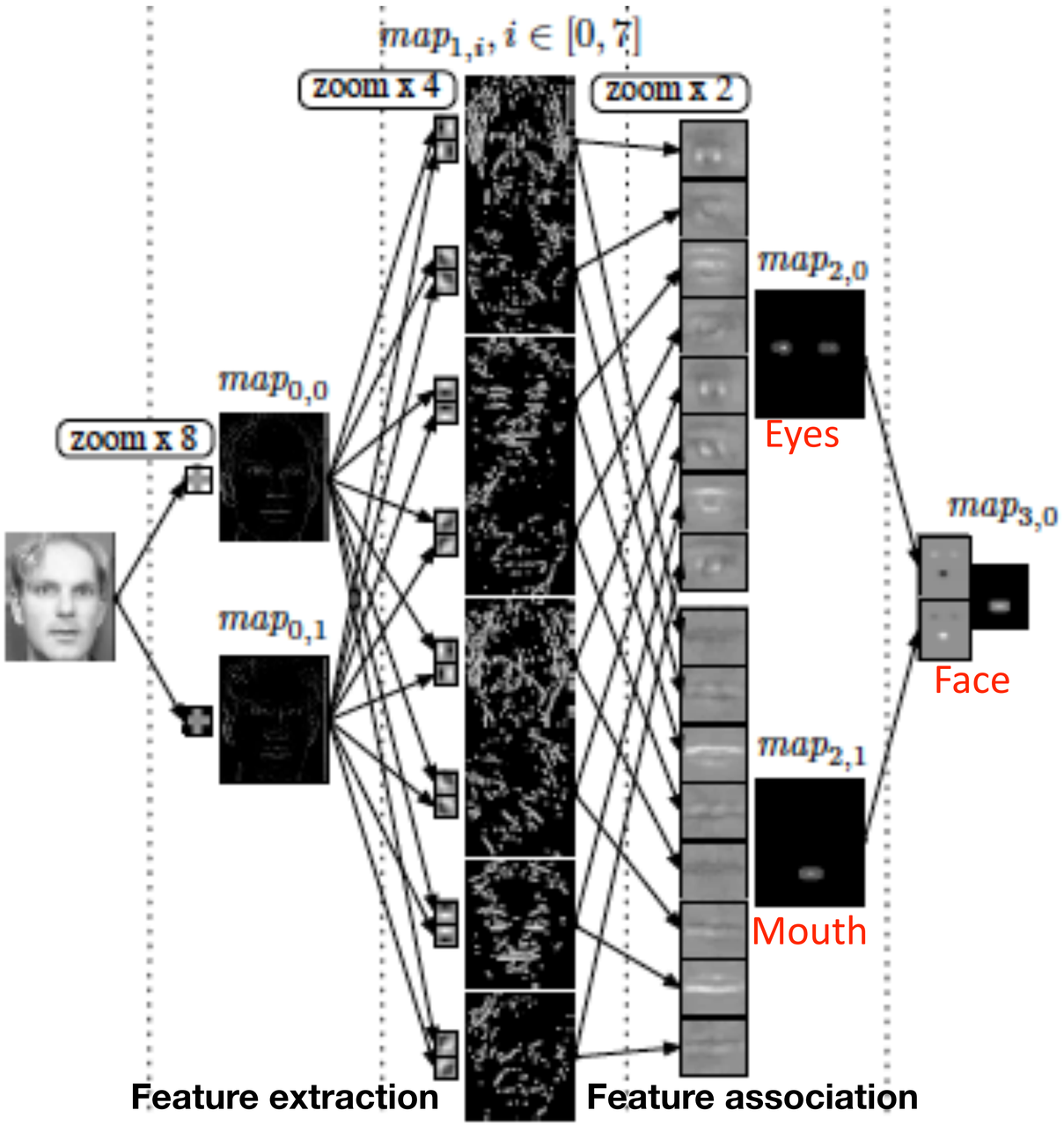
How machine learning works: Example of an AI System for Face Detection



Results



What influences the learning process?



Data

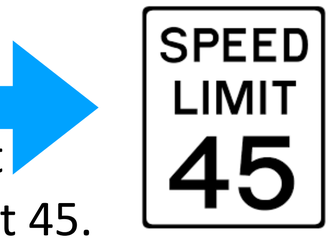
- Training and test sets (*bias, labels*, …)

Design choices and architecture

- Choice of Features
- Class semantics
- Network structure and parameters

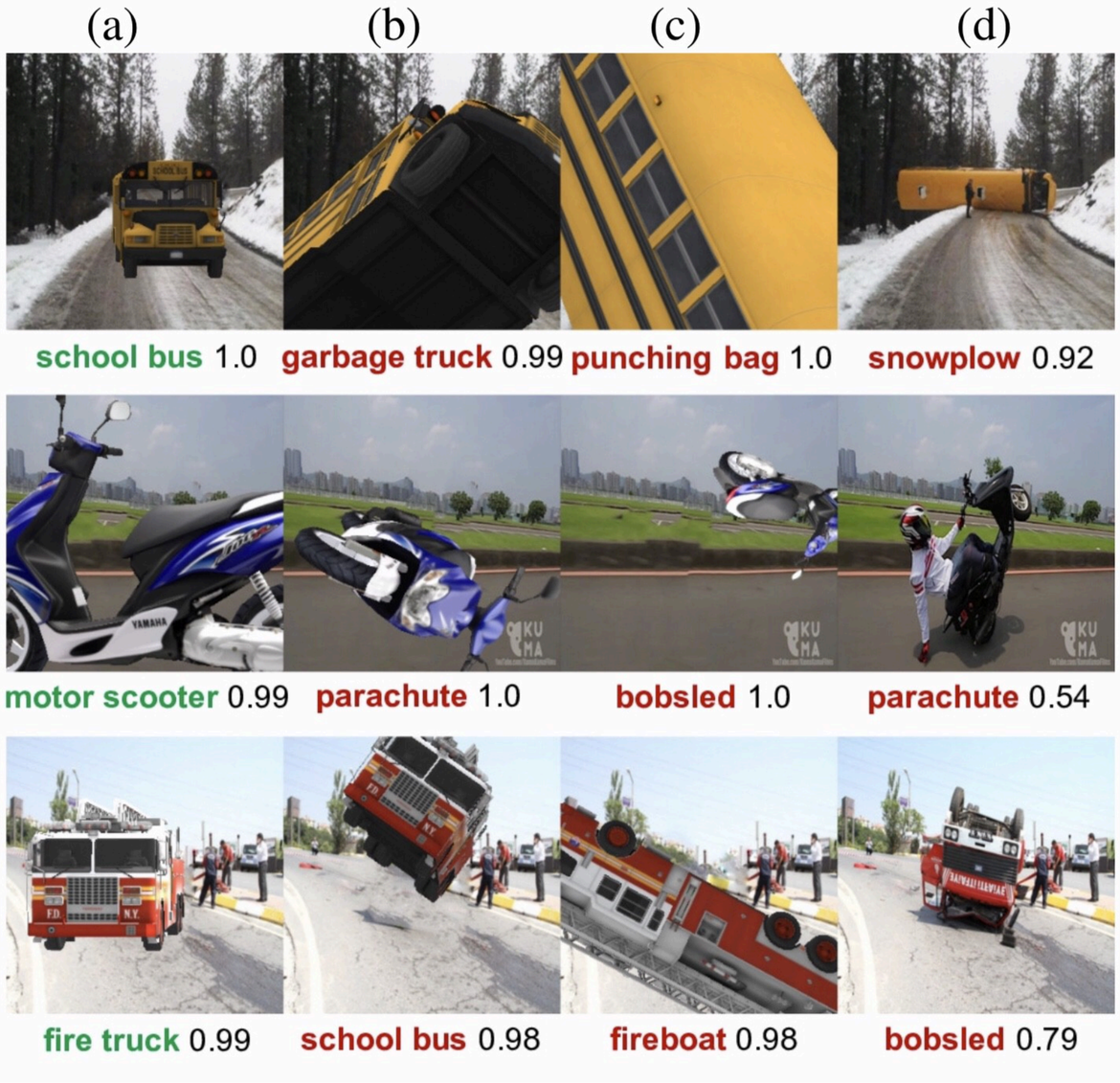
Optimization Process

Deep Learning Limitations



Targeted physical perturbation experiment
The misclassification target was Speed Limit 45.

Robust Physical-World Attacks on Deep Learning Models K. Eykholt et al. CVPR 2018.



Strike (with) a Pose: Neural Networks Are Easily Fooled by Strange Poses of Familiar Objects. Michael A. Alcorn et al., CVPR 2019

Issues in Statistical ML

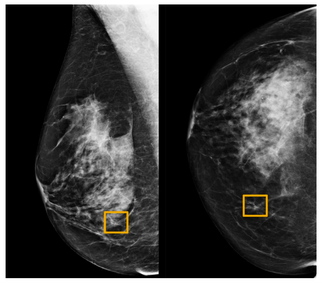
- AI Black box: data, millions/billions of parameters, optimization algorithms, off-the-shelf components
- No solid verification and validation processes, qualification and benchmarking of results
- Absence of causal links between inputs and outputs: no explanation of results
- AI systems lack semantics and context awareness

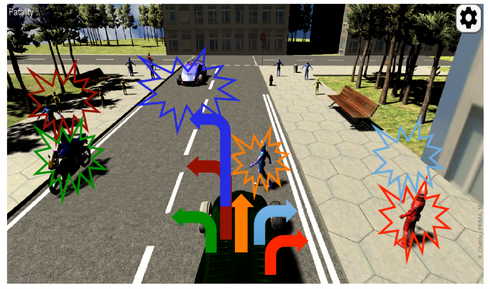
Trustworthiness of Decision and Action Delegation to AI systems and Robots

- Critical applications (healthcare, transports, …)
- Areas threatening human rights, values, wellbeing, …

→ **Need for robustness and safety**

→ **Need for ethics and governance**






Framework for Trustworthy AI (EU HLEG-AI, 2019)

- **Demonstrable trustworthiness** as a prerequisite to develop, deploy and use AI systems.
  - Trust in organizations developing and deploying AI
  - Appropriate conditions of use and applications

https://ec.europa.eu/digital-single-market/en/high-level-expert-group-artificial-intelligence

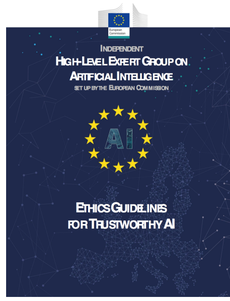
Ethical Principles for Trustworthy AI

Ethical imperatives

- **Principle of Autonomy**: “Preserve Human Agency and control”
- **Principle of Non maleficence**: “Do no Harm” - Neither cause nor exacerbate harm or otherwise adversely affect human beings. safety and security, technical robustness.
- **Principle of Justice**: “Be Fair”. Equal and just distribution of benefits and costs, free from unfair bias, increase social fairness
- **Principle of Explicability**: “Operate transparently”. Traceability, auditability, transparent system capabilities, …

Requirements for Trustworthy AI

High-Level Expert Group on AI (EU) - April 2019



1. **Human agency and oversight**- Including fundamental rights, human control
2. **Technical robustness and safety** - Including resilience to attack and security, fall back plan and general safety, accuracy, reliability and reproducibility
3. **Privacy and data governance** - Including respect for privacy, quality and integrity of data, and access to data
4. **Transparency** - Including traceability, **explainability** and communication
5. **Diversity, non-discrimination and fairness** - Including the avoidance of unfair bias, accessibility and universal design, and stakeholder participation
6. **Societal and environmental wellbeing** - Including sustainability and environmental friendliness, social impact, society and democracy
7. **Accountability** - Including auditability, minimisation and reporting of negative impact, trade-offs and redress.

Tool: Assessment List for Trustworthy AI - ALTAI

https://ec.europa.eu/digital-single-market/en/high-level-expert-group-artificial-intelligence

Achieving Trustworthy AI : Technical Aspects

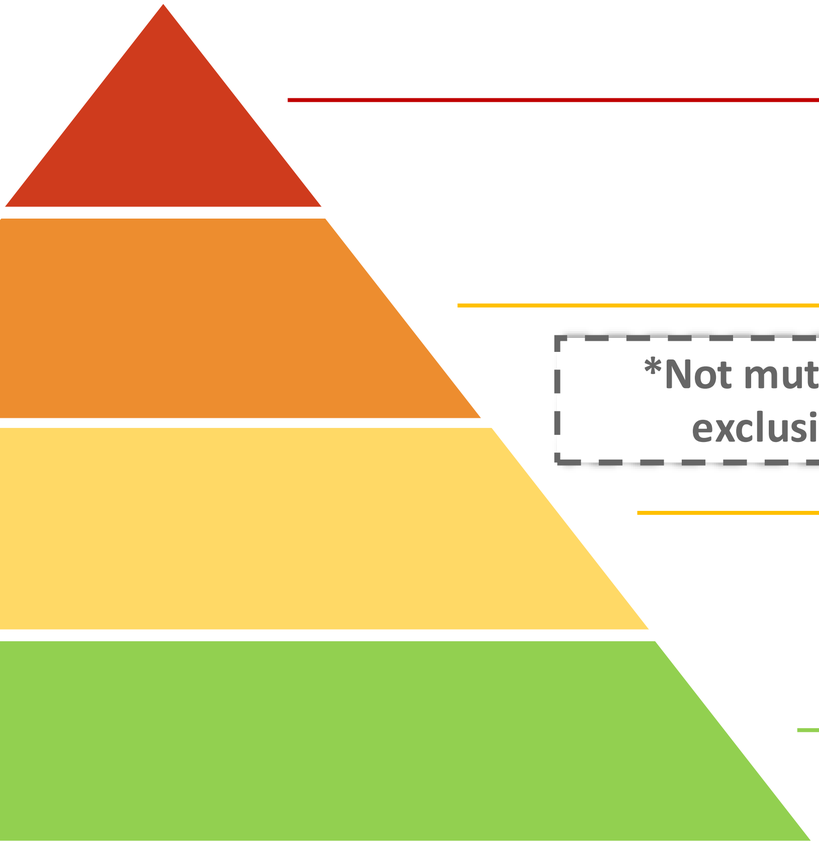
- *Architectures for Trustworthy AI*
- *Ethics and rule of law by design (X-by-design)*
- *Explanation methods*
- *Testing and validating*
- *Quality of Service Indicators*

Achieving Trustworthy AI : Non-Technical Aspects

- *Regulation*
- *Codes of conduct*
- *Standardisation*
- *Certification*
- *Accountability via governance frameworks*
- *Education and awareness to foster an ethical mind-set*
- *Stakeholder participation and social dialogue*
- *Diversity and inclusive design teams*

A risk-based approach to regulation

EU Legislative proposal (21/04/2021)



Unacceptable risk
e.g. social scoring

High risk
e.g. recruitment, medical devices

AI with specific transparency obligations
'Impersonation' (bots)

Minimal or no risk

*Not mutually exclusive

Unacceptable risk e.g. social scoring	**Prohibited**
High risk e.g. recruitment, medical devices	**Permitted** subject to compliance with AI requirements and ex-ante conformity assessment
AI with specific transparency obligations 'Impersonation' (bots)	**Permitted** but subject to information/transparency Obligations
Minimal or no risk	**Permitted** with no restrictions

European Commission

The EC's High-Risk AI Application Definition

EU Legislative proposal (21/04/2021)

"The classification of an AI system as high-risk is based on the intended purpose of the AI system, in line with existing product safety legislation. Therefore, the classification as high-risk does not only depend on the function performed by the AI system, but also on the specific purpose and modalities for which that system is used." (§5.2.3)

Two main categories of high-risk AI systems (Chapter 1 of Title III):

- ***AI systems intended to be used as safety component of products that are subject to third party ex-ante conformity assessment;*** e.g.: medical devices
- *other stand-alone AI systems with mainly fundamental rights implications that are explicitly listed in Annex III.* e.g.: Law enforcement

Risks

- Risks, *i.e.*, probability of dangers to materialize, must be properly assessed
- Risks are threats to values or rights and are often multidimensional (life, physical integrity, physical wellbeing, mental wellbeing, dignity, privacy, freedoms, security, fairness, equality, truth …)
- Risks are related to the technology itself, environment conditions, context, usage;
- Uncertainties and unknown impacts, especially if the systems interact with people (e.g., : recommender systems, transformer based language models)

Human Responsibility

- *“To ensure every stakeholder involved in the design and development of AIS is educated, trained, and empowered to prioritize ethical considerations so that these technologies are advanced for the benefit of humanity.”*

  Mission statement of the IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems, 2016

- Assess reality of dangers and risks to elaborate guidelines, policies and regulations to prevent and mitigate potential risks
- **Appropriate <u>design</u> approaches and <u>governance</u> frameworks**

Human Responsibility

- *“To ensure every stakeholder involved in the design and development of AIS is educated, trained, and empowered to prioritize ethical considerations so that these technologies are advanced for the benefit of humanity.”*

  Mission statement of the IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems, 2016

- *“To support and guide the responsible adoption of AI that is grounded in human rights, inclusion, diversity, innovation, economic growth, and societal benefit, while seeking to address the UN Sustainable Development Goals.”*

Ethically Aligned Design
Value-Based Design
Value-Sensitive Design

- Project definition: What are the project objectives? What are its benefits? Does it entail risks? Who are the stakeholders and what are their values?
- Project Initiation: Value conceptualisation; feasibility studies.
- System specification: Value analysis, value tensions, value priorities.
- Design, architecture, alternatives.
- Validation; Success metrics, validation of values
- Deployment and maintenance.
- Evaluation of value compliance during system operation

See Standard IEEE P7000- Model Process for Addressing Ethical Concerns During System Design

The case of “Autonomous” Vehicles

- **Why build “autonomous” vehicles**? Examine motivations. In what conditions? Who is impacted (stakeholders and their values)? Benefits and risks?
  - Motivation for AVs, terminology used to describe them (automated driving or autonomous driving?)
- **Safety and security**
  - Functions associated with automation (perception, control, …); What are the technical limitations? How do they impact stakeholders’ values?
- **Human control**
  - Human agency and oversight enablers, Human-Machine interaction.
- **Personal and public freedoms**
  - Impact of data collection o, privacy and freedom of movement
- **Social and environmental impacts**
  - Accessibility, fairness, impact on traffic, on cities, lifestyle, impact on the environment
- [Trolley problem (not the main issue in automated driving!)]

See: https://www.ccne-ethique.fr/fr/actualites/cnpen-le-vehicule-autonome-enjeux-dethique

Takeaways: Responsible Development, Use and Governance of AI

- AI is no silver bullet for many application. Avoid technical solutionism. Technology alone will not solve the issues: need for **governance and regulation**.
- AI systems using machine learning need to be made robust at system level, not (just) the AI component level
- Dependability and resilience
- Explainability is essential to build trust in AI systems
- Auditing and certification of AI systems are necessary
- Complying with a responsible and human-centered AI approach can be assessed through the compliance with the HLEG-AI 7 key requirements