EUROSHNET

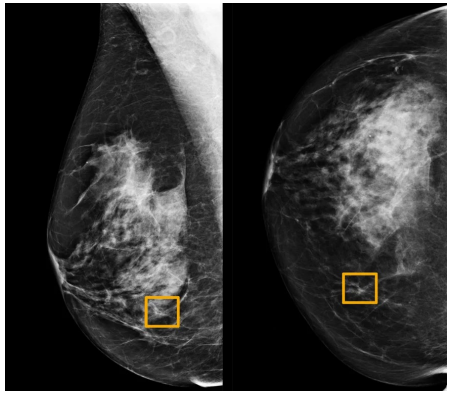
ARTIFICIAL INTELLIGENCE
MEETS SAFETY AND HEALTH AT WORK

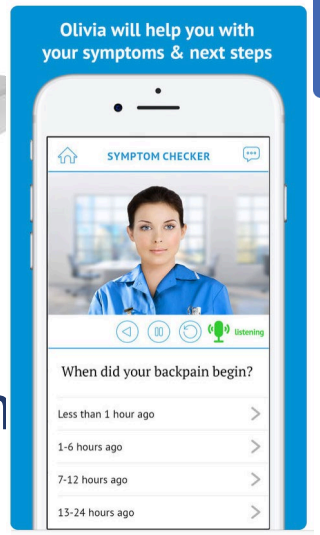
Foundations, Methods, Applications and Limitations of Artificial Intelligence

Raja Chatila
Institute of Intelligent Systems and Robotics (ISIR)
Faculty of Sciences and Engineering, Pierre and Marie Curie Campus
Sorbonne University, Paris, France
Raja.Chatila@sorbonne-universite.fr

Multiple Applications of AI And Robotics

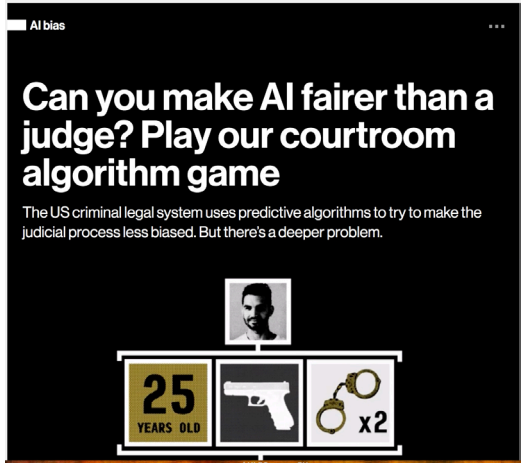
- Transportation, logistics, delivery
- Healthcare
- Manufacturing
- Agriculture
- Personal services & assistance
- Security
- Recommender systems, advertisen
- Recruitment & management
- Insurance & finance
- Justice
- Warfare
- …





A face-scanning algorithm increasingly decides whether you deserve the job

HireVue claims it uses **artificial intelligence** to decide who's best for a job. Outside experts call it 'profoundly disturbing.'



Artificial Intelligence

Machine Learning

Deep Learning

Reinforcement Learning

Robotics

Control Theory

Mechanical Design

Real-time Systems

"Symbolic AI"

Knowledge Representation;

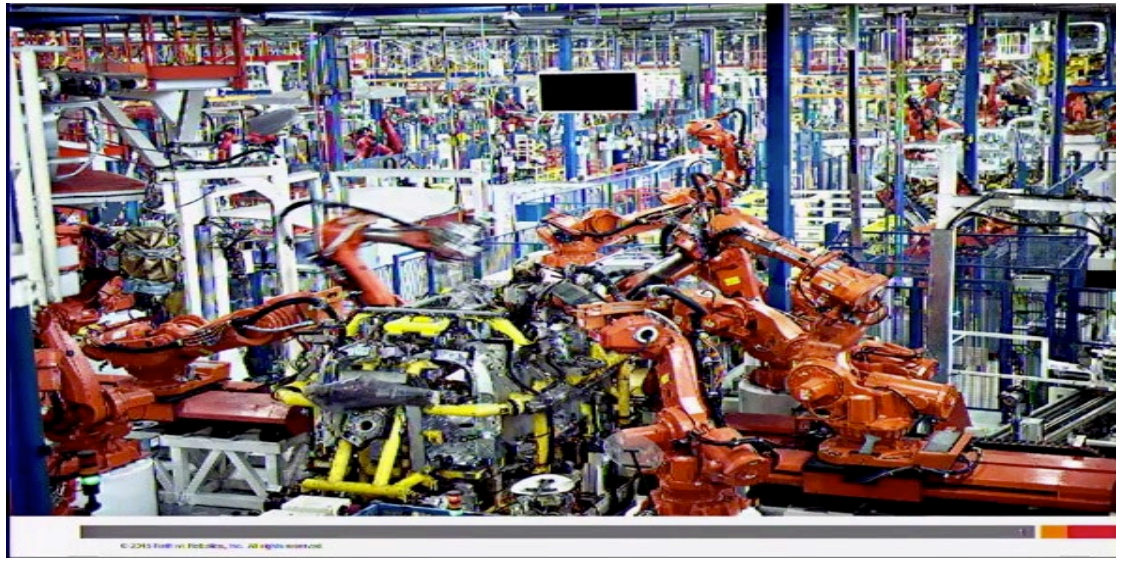
Logical inference and Probabilistic Reasoning;

Problem Solving and Search; Planning

What is an Computational "Intelligent" System?

- A computational intelligent system is a set of **algorithms designed by humans**, using data (big/small/sensed) to solve [more or less] complex problems in [more or less] complex situations.
- The system might include deductive inference, as well as machine learning processes, *i.e.*, the capability of improving its performance based on data classification to build **statistical models** from data (*e.g.*, deep learning), or on evaluating previous decisions (*e.g.*, reinforcement learning).
- Such systems could be regarded as "autonomous" in a **given domain** and for **specific tasks**, as long as they are capable of accomplishing these tasks despite environment variations within this domain.
- Difference between automated and autonomous systems is related to **complexity** of task and domain, and **importance** of variations

From Full Robotization to Human-Robot collaborative tasks



Machine Learning

Statistical data processing and classification

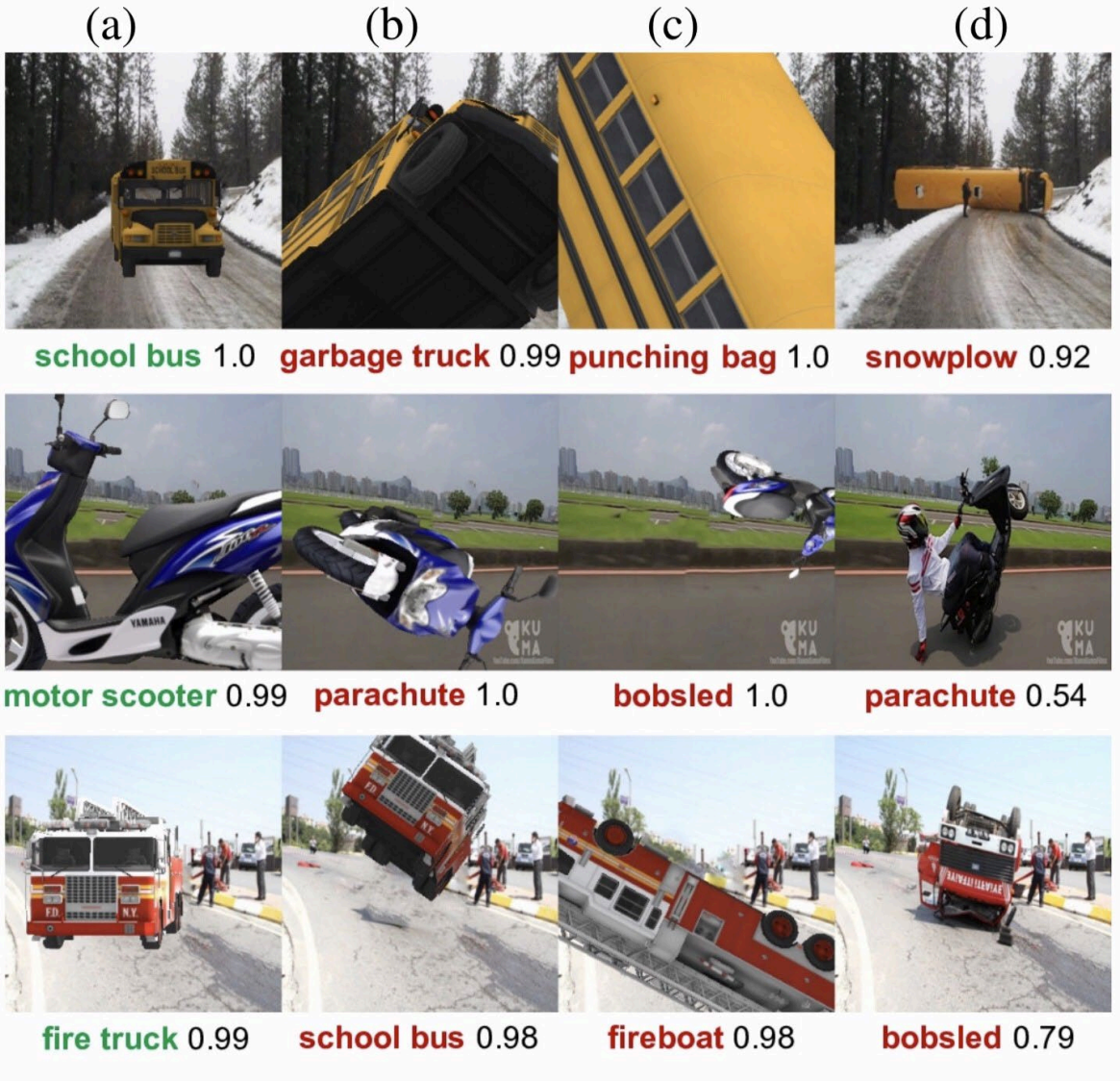
- Use of probability distributions, correlations, …
- Use of artificial neural nets as classifiers
- Optimization algorithms

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

Deep Learning Limi Robustness

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 with Statistical Machine Learning

- Black box: millions/billions of parameters, optimization algorithms, un certified off-the-shelf components
- No solid verification and validation processes or qualification of results
- Quality and representativeness of data. Data Bias
- Bias due to design and architecture choices
- Inappropriate correlations, absence of causality between data and results
- No explicability
- Computational level: No semantics, no understanding of manipulated symbols, no context awareness
- Environmental cost

Risks and Trustworthiness of AI Systems

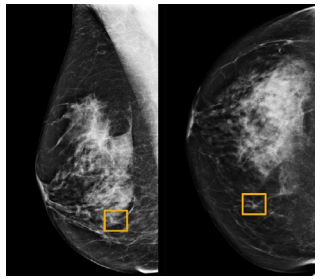
- No ethical rules in academic AI research
- Advanced AI research in industry without ethical oversight
- Applications in critical domains (healthcare, transport, security…)
- Applications potentially threatening human rights and values (surveillance, opinion manipulation, policing, justice, access to jobs and education, …)

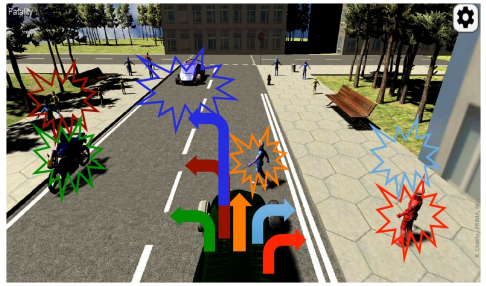
→ **Need for robustness and safety**

→ **Need for ethics and governance**

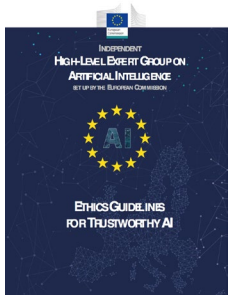
} Transparency

} Explainability





Key Requirements for Trustworthy AI
High-Level Expert Group on AI (EU) - April 2019



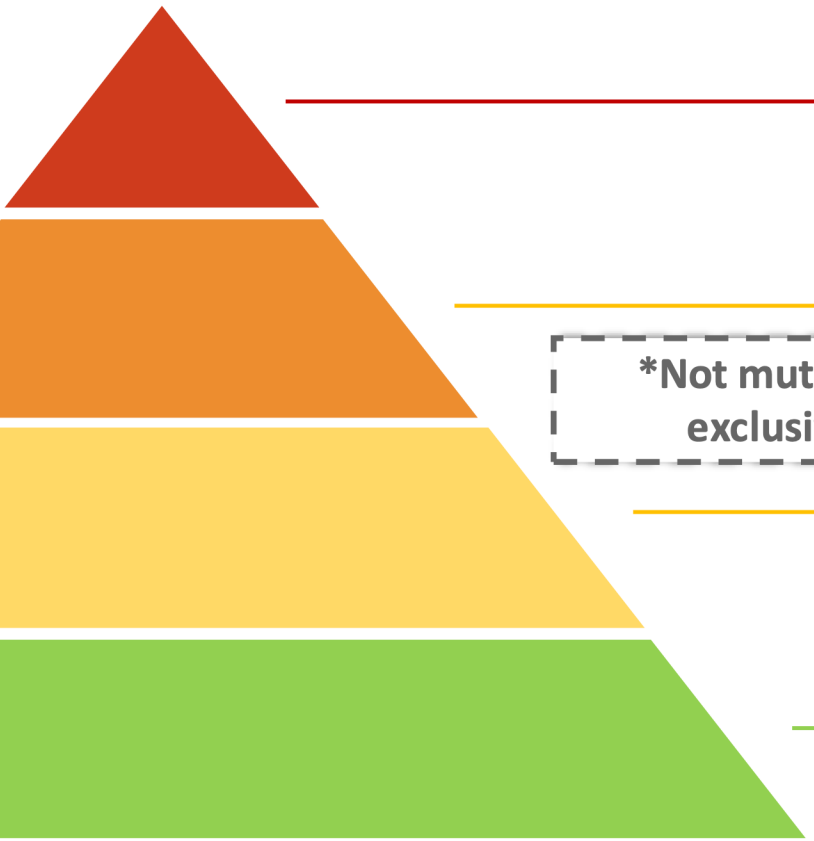
1. **Human agency and oversight**- Including respect tof fundamental rights, human control
2. **Technical robustness and safety** - Including resilience to attack and security, fall back plan and general safety, accuracy, reliability and reproducibility
3. **Privacy and data governance** - Including 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

A risk-based approach to regulation

EU Legislative proposal (21/04/2021)



*Not mutually exclusive

Unacceptable risk
e.g. social scoring

High risk
e.g. recruitment, medical devices

AI with specific transparency obligations
'Impersonation' (bots)

Minimal or no risk

Prohibited

Permitted subject to compliance with AI requirements and ex-ante conformity assessment

Permitted but subject to information/transparency Obligations

Permitted with no restrictions

Courtesy L. Sioli

European Commission

Courtesy L. Sioli

Main Takeaways

- AI and Robotics contributes ot increase productivity through physical process or software automation
- They enable to achieve tasks that are too repetitive, or were not achievable before (too dangerous, too costly, too difficult for humans) and create new services
- Exploit available massive data (images, scientific data, text, …)
- But AI is no silver bullet for many application. Avoid technical solutionism.
- AI systems using machine learning need to be made robust and resilient
- Explainability is essential to build trust in AI systems
- Appropriate design approaches, governance frameworks, auditing and certification of AI systems are necessary.