Trustworthy AI

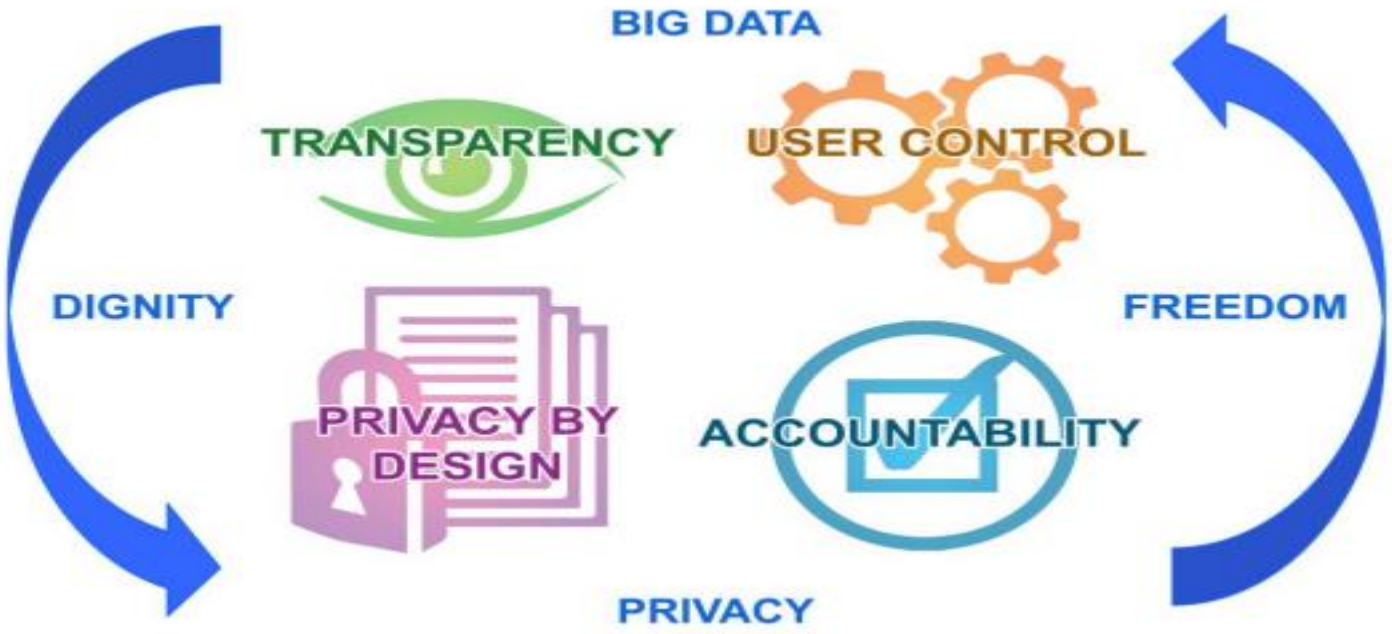
Lawful AI + Ethical AI + Robust AI

FOSCA GIANNOTTI ISTI-CNR, PISA ITALY

TAILOR ECAI WORKSHOP 2020

The GDPR

- In force on 25 May 2018
- Introduces important novelties
  - New Obligations
  - New Rights



Right of Explanation

Since 25 May 2018, GDPR establishes a right for all individuals to obtain **meaningful explanations of the logic involved** when “automated **(algorithmic)** individual decision-making”, including profiling, takes place.

Ethical principles for trustworthy AI

respect for human autonomy
- self-determination
- no-coercion
- no-manipulation

prevention of harm
- safe and secure

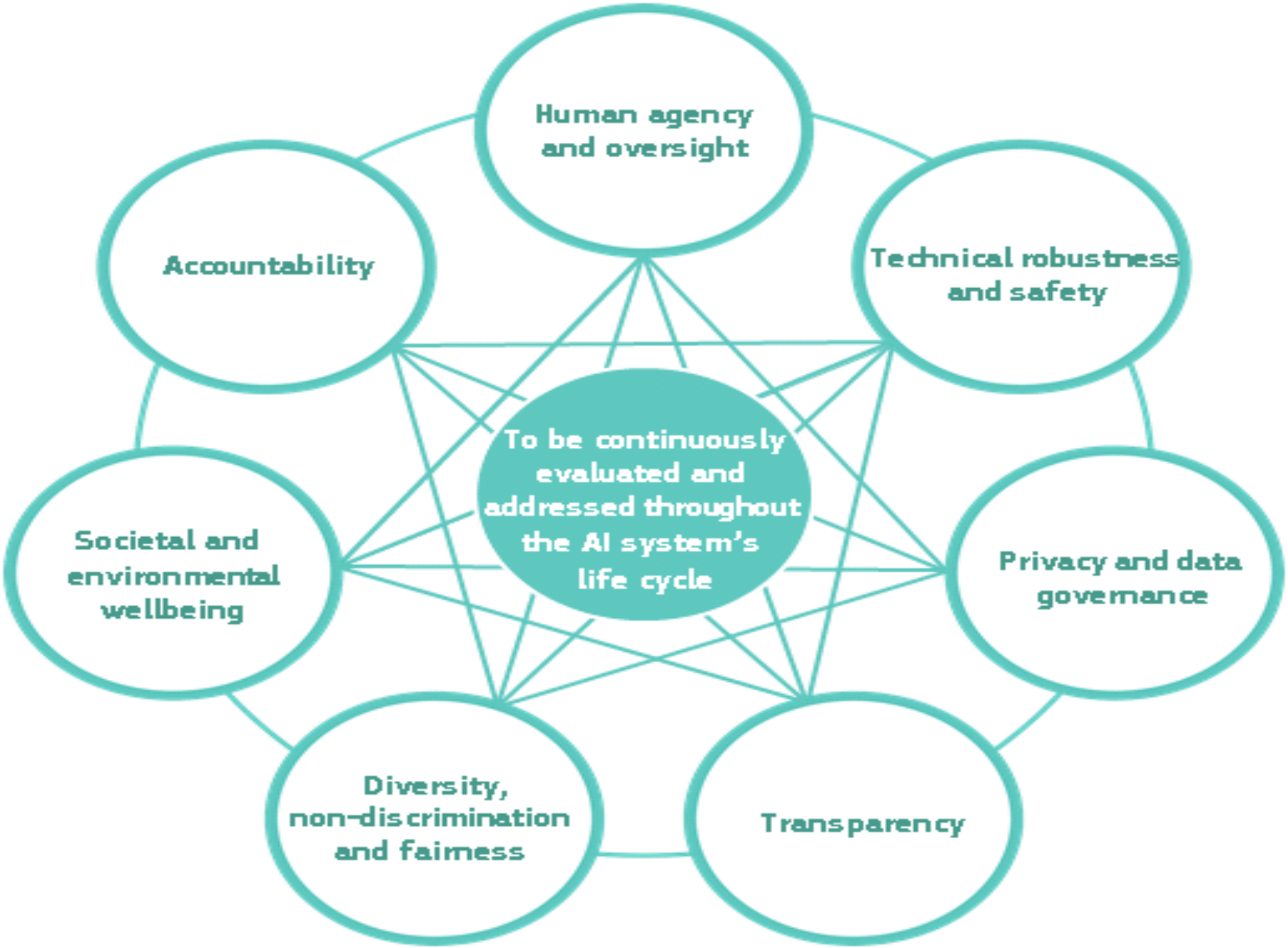
fairness
- no-discrimination (no-bias)

explicability
- User trust and transparency
- intelligibility “how does it work?”
- accountability (“who is responsible for”)

What makes an AI system trustworthy?

- **respecting the rule of law;**
- **being aligned with agreed ethical principles and values, including privacy, fairness, human dignity;**
- **keeping us, the humans, in control;**
- **ensuring the system's behavior is transparent to us, and its decision making process is explainable;**
- **and being robust and safe, meaning that the system's behavior remains trustworthy even if things go wrong.**

The 7 Requirements for T-AI



TAILOR: How to develop **foundations for** Trustworthy AI?

- **designing and developing AI systems that**
  - **incorporate the safeguards that make them trustworthy, and respectful of human agency and expectations.**
  - **Not only the mechanisms to maximize benefits, but also those for minimizing harm.**

- **D1) explainability, D2) safety, D3) fairness, D4) accountability and reproducibility, D5) privacy and D6) sustainability.**

D1: Explanability and Intellegibility

- *How can we guarantee user trust in AI systems through **explanation**? How to formulate explanations as Machine-Human conversation depending on context and user expertise?*

◆Exemplar problem:

1

ML Classifier

*C: I predict "**High Risk of Recidivism**"*

2

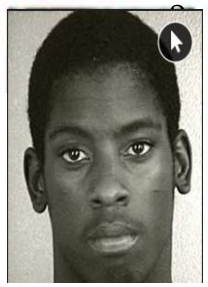
H: Why?

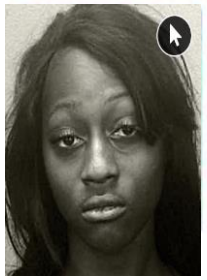
*{age <= 38 & race = Afro-Am & recidivist = True} → **"High Risk of Recidivism"***

C: Because he is younger than 38, Afro-Am and already recidivist

3

H: (Hmm. It could have a racial bias!) Which training examples are most similar to the prediction and influencing the





4

H: What happens if you don't consider that he is young and

*C: I still predict "**High Risk of Recidivism**" because he is Afro-Am*

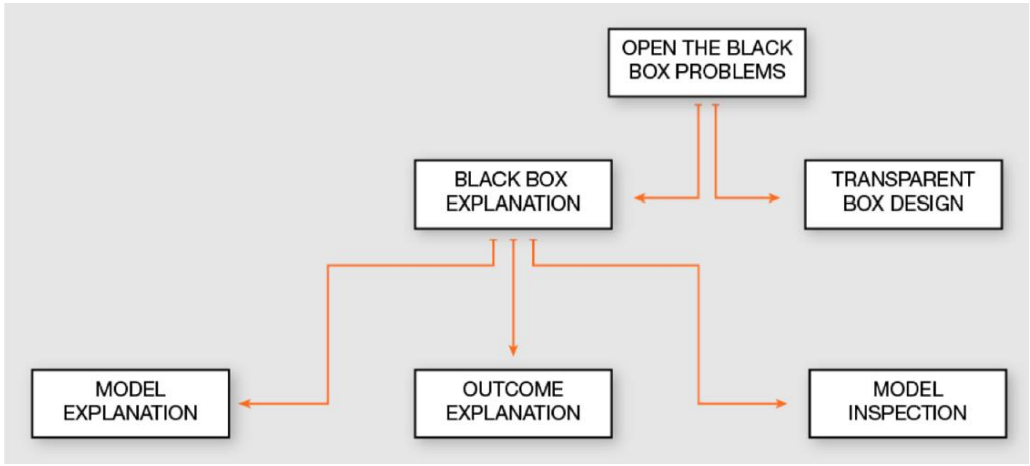
Explanation-State of art

Explanation is hot-topic in many in AI fields

- Machine Learning
- Computer Vision
- Image recognition
- Knowledge Respresentation and reasoning
- Game Theory
- Robotics
- Molti-Agents Systema
- NLP

https://xaitutorial2019.github.io/

Problem Taxonomy in Explanation in ML:



- Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). ***A survey of methods for explaining black box models***. *ACM Computing Surveys (CSUR)*, *51*(5), 93.

D2: Robustness & Safety

- ?*How to bridge the gap from **safety** engineering, formal methods, verification as well as validation to the way AI systems are built, used, and reinforced?*

◆Exemplar problem:

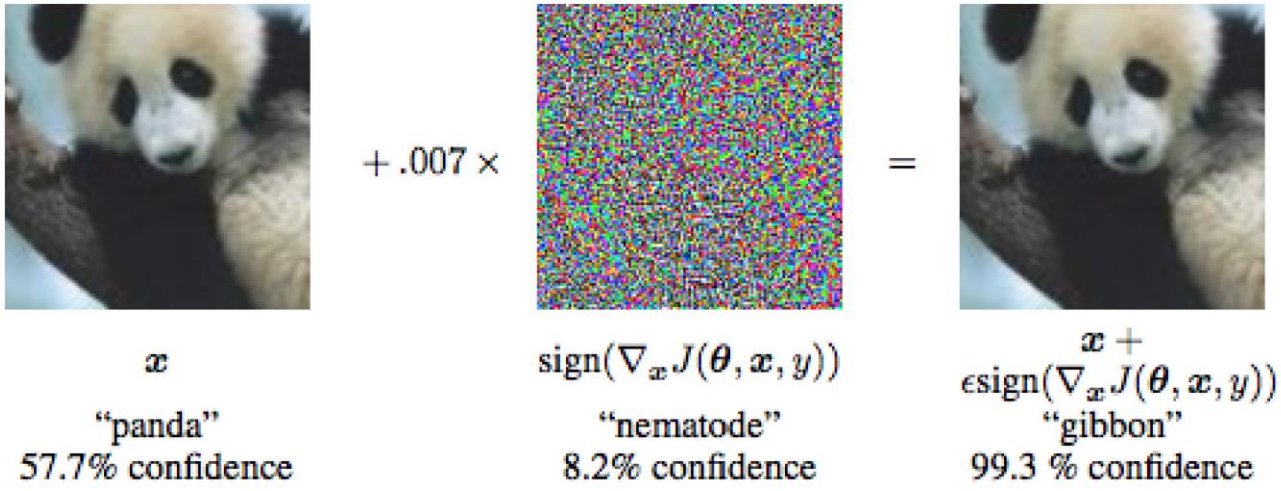
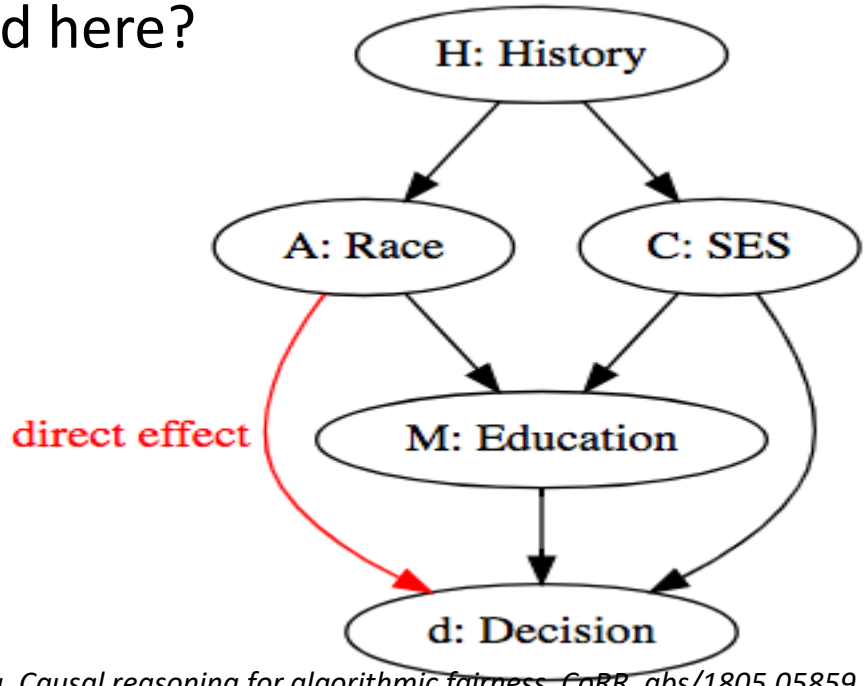


Figure 1: Adversarial example, which obtained by applying small, almost invisible, perturbation to the input image. As a result, network misclassified the object.

D3: Fairness and Causality

?*How can we build algorithms that respect **fairness** constraints by design through understanding causal influences among variables for dealing with bias-related issues?*

- Why don’t black people get hired here?
- - Direct effect
- - Indirect effect
- - Back-door path



- *[Loftus et al., 2018] J. R Loftus, C Russell, M. J Kusner, and R Silva. Causal reasoning for algorithmic fairness. CoRR, abs/1805.05859, 2018.*

D4: Respect for Privacy

- ?*Can we guarantee **privacy** while preserving the desired utility functions?*

What is mainly done

…and other reasearch questions

- D5: ?*How to uncover **accountability** gaps w.r.t. the attribution of AI-related harming of humans?*
- D6: ?*Is there any chance to reduce energy consumption for a more **sustainable** AI and how can AI contribute to solving some of the big sustainability challenges that face humanity today (e.g. climate change)?*
- ***?How to deal with properties and tensions of the interaction among multiple dimensions? For instance, accuracy vs. fairness, privacy vs. transparency, convenience vs. dignity, personalization vs. solidarity, efficiency vs. safety and sustainability.***

“Safety may not be perfect, but the greater the human ability, the more advanced it will be”

“**Knowledge has its risks, but should our reaction be to stop at risk?** Or should we not rather use knowledge to make it a barrier against the same risks that it entails? **Knives are made with a handle so that they can be grasped without danger**; stairs are equipped with railings; **electrical wires are insulated**; pressure cookers have a safety valve; in every product we take care to minimize the risk. **Sometimes the safety achieved is insufficient**, due to limitations imposed by the nature of the universe or by the human mind. **However, the attempt must be made.** As a machine, a robot will certainly be designed to offer guarantees of safety, at least as far as possible. **Safety may not be perfect (is there anything that is?), but the greater the human ability, the more advanced it will be**.”

Isaac Asimov, Introduction to “Second book of Robots” (1964)

THANK YOU!

SoBigData

AI4EU

HUMANE AI

Track & Know

PRO-RES

XAI

erc
European Research Council
Established by the European Commission