The AI4EU Observatory on Society and AI

4th European Conference on AI in Finance and Industry
ZHAW, Winterthur, Switzerland, 5 September 2019

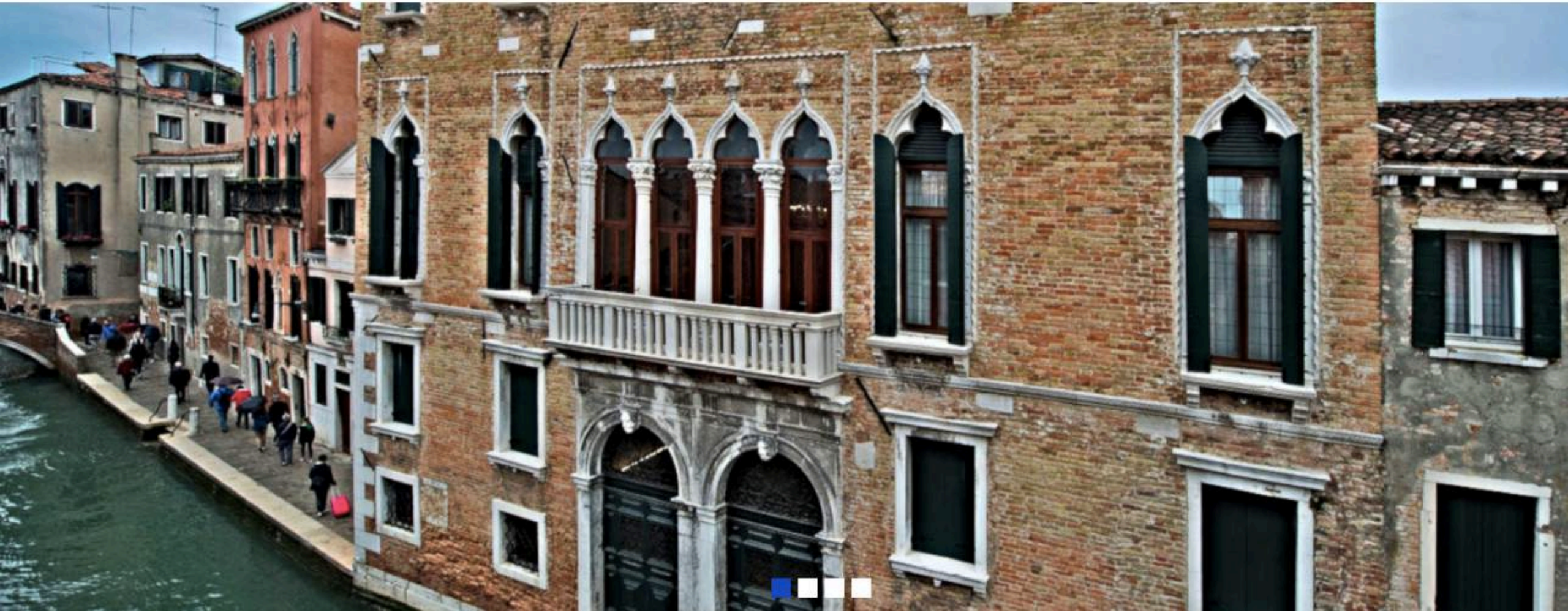
Teresa Scantamburlo

European Centre for Living Technology (ECLT)

Ca' Foscari University of Venice

Ca' Foscari University of Venice | European Centre for Living Technology

Home About us Research Activities News Location

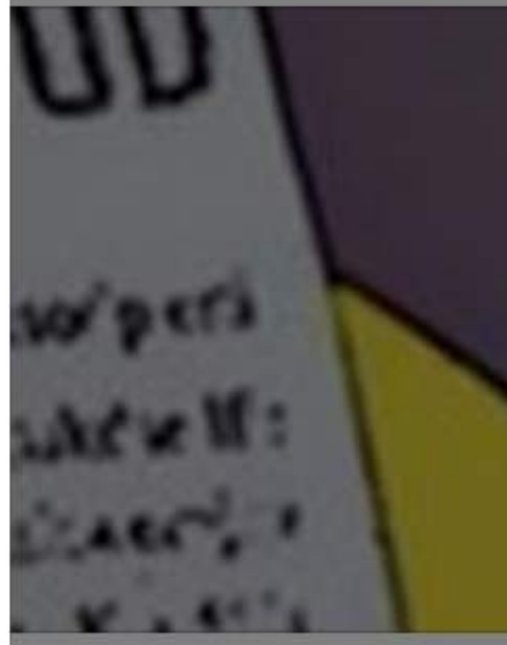


jackyalciné (he/him/his)
@jackyalcine

People-centric software consultant. 💼
black.af + koype.net; fmr @lob, @lyft, @getclef ➡️ jacky.wtf 🥬 vegan

jacky.is@black.af

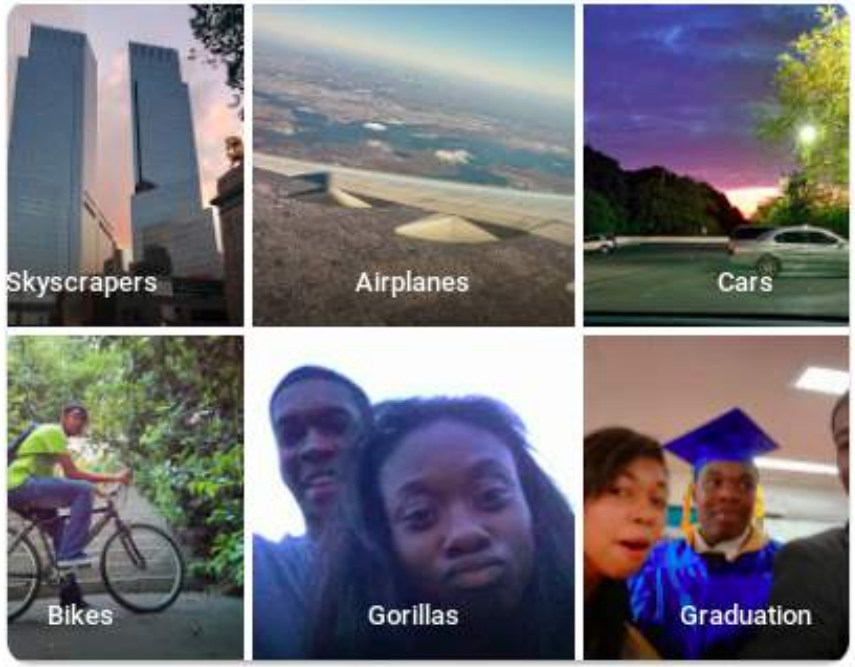
Joined June 2009



jackyalciné (he/him/his)
@jackyalcine

Follow

Google Photos, y'all fucked up. My friend's not a gorilla.



6:22 pm - 28 Jun 2015

3,261 Retweets **2,384** Likes

238 3.3K 2.4K

Amazon built an AI tool to hire people but had to shut it dow because it was discriminating against women

Isobel Asher Hamilton Oct. 10, 2018, 5:47 AM

The New York Times

YouTube's Product Chief on Online Radicalization and Algorithmic Rabbit Holes

Neal Mohan discusses the streaming site's recommendation engine, which has become a growing liability amid accusations that it steers users to increasingly extreme content.

AI & ethics in Europe

In this talk:

- AI4EU Observatory on Society and AI
- European commitment to Human-centred AI
- Ethics Guidelines for Trustworthy AI by the High-level Expert Group on AI
- Concluding remarks

AI4EU project

- Title: A European AI On Demand Platform and Ecosystem (AI4EU)
- Participants: 79 partners in 21 countries
- Overall objective: to develop and European AI Ecosystem (AI expertise, tools, data, methods for validation and ethical assessment, case studies, etc)
- Some specific objectives:
  - to build a Platform providing access to relevant AI resources in the EU for all users
  - To create an Observatory to support the development of a human-centred AI approach
- Website: https://www.ai4eu.eu/

AI4EU Observatory

- Intuition: a special place where to find valuable equipment to study certain phenomena
- AI4EU Observatory on Society and AI (OSAI) promotes the discussion and distribution of information about the Ethical, Legal, Socio-Economic and Cultural issues of AI (ELSEC-AI) within Europe
- Example of ELSEC-AI
  - Ethical issues: agency, autonomy, responsibility, solidarity…
  - Legal issues: privacy, bias, justice, rights, democracy…
  - Socio-Economic: equality, trust, labour, common good…
  - Cultural: representations of AI, AI education, interdisciplinarity, multiculturalism…

OSAI's tasks

- The main OSAI tasks are:
  - Mapping the landscape of European strategies addressing ELSEC-AI
  - Connecting existing European initiatives (gathering stakeholders)
  - Reporting what's going on around ELSEC-AI in Europe (also in less debated areas/countries)
  - Informing and educating the EU public at large
  - Supporting the debate on ELSEC-AI and the interdisciplinary dialogue (ELSEC-A working groups)
- OSAI will act mainly through the AI4EU platform
- Web demonstrator: https://www.unive.it/osai

European approach to AI

“Artificial Intelligence for Europe” COM(2018) 237, 25 April 2018

The European initiative aims to:

- “Boost the EU's technological and industrial capacity and AI uptake across the economy”
- “Prepare for socio-economic changes brought about by AI”
- “Ensure an appropriate ethical and legal framework, based on the Union's values and in line with the Charter of Fundamental Rights of the EU”

“Coordinated Plan on Artificial Intelligence” COM(2018) 795, 7 December 2018

“Overall, the ambition is for Europe to become the world-leading region for developing and deploying cutting-edge, ethical and secure AI, promoting a human-centric approach in the global context.”

Human-centric AI

In short human-centric AI implies:

- People can trust AI systems (trustworthy AI)
- Individuals and the society can benefit from the use of AI
- AI systems are based on ethical and societal values, in particular, the European Charter of Fundamental Rights

In more concrete terms:

- ethical and secure by design
- clear ethics guidelines and standards
- legal framework

Ethics guidelines

High-level Expert Group on Artificial Intelligence (AI HLEG)

AI HLEG's main deliverables:

- AI Ethics guidelines delivered
- Policy and investment Recommendations

AI HLEG's ethics guidelines:

- first draft December 2018
- public consultation
- official delivery in April 2019
- piloting process with the support of AI4EU (June-December 2019)

Website: https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai

Trustworthy AI

"AI systems need to be human-centric, resting on a commitment to their use in the service of humanity and the common good, with the goal of improving human welfare and freedom."

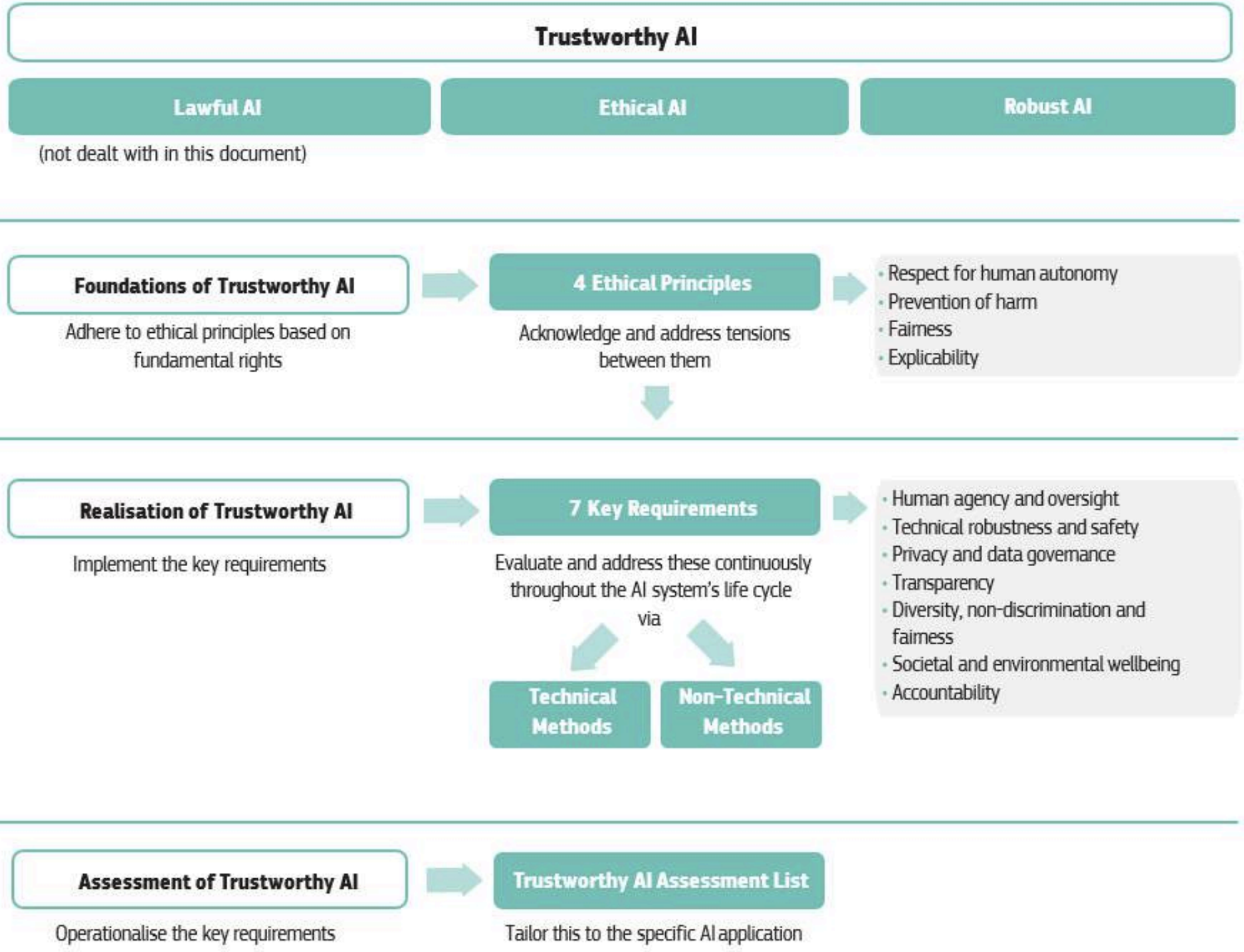
Trustworthy AI (instead of ethical AI)

- being demonstrably worthy of trust (concrete pathways)
- it refers to the socio-technical system in which AI technology is embedded (holistic approach)
- Trustworthy AI to promote "responsible competitiveness"
- Addressed to AI stakeholders, e.g. companies, civil society organisations, individuals, ...

Some remarks:

- Trustworthy AI is a contribution to elaborate "a normative vision of an AI-immersed future"
- need of an ethical culture through public debate, education and practical learning

Framework



AI HLEG, *Ethics Guidelines for Trustworthy AI* (2019, p 8)

7 ethical requirements

They can help the implementation of trustworthy AI

1. human agency and oversights
2. technical robustness and safety
3. privacy and data governance
4. transparency
5. diversity, non-discrimination and fairness
6. societal and environmental well-being
7. accountability

value-by-design methodology

fairness metrics

multiple performance measures

explainable AI methods

testing performed by diverse groups

adversarial testing

codes of conduct

"bug bounties"

regulation (e.g. GDPR)

accountability via governance frameworks

education

stakeholder participation

diversity and inclusive design teams

Toy example

- Sirio = a personal assistant providing advisory service to University's students
- goal = to help students make better choice (e.g. curriculum selection, assistance in bureaucratic processes, advices to improve performances, etc.)
- how do key requirements for trustworthy AI apply here?

Toy example

- Human agency and oversight
  - does Sirio respect students' autonomy?
  - does it act in accordance with their goals?
- Technical robustness and safety
  - how does Sirio perform?
  - does Sirio provide bad answers (e.g. due to malicious attacks or exposition to rude conversations)?
  - does Sirio have a consistent behaviour?
- Transparency
  - how does Sirio make its decisions?
  - Is this mechanism accessible by students?
  - how does Sirio present itself to students?
  - is a human alternative provided?

- impact assessment
- discussion with all stakeholders (university representatives , including students)
- stakeholder panel
- educating students to interact with Sirio

- diverse performance measures
- precise definition of the desired outcome
- multidisciplinary testing
- regular user studies
- incentives to report Sirio's errors or weaknesses

- disseminating the logics of Sirio
- making (training) data and algorithms open to public audit
- communicate the strengths and the limits of Sirio

Are you ready for trustworthy AI?

Trustworthy assessment list

Brief sketch:

- list of questions structured around the 7 requirements
- goal = to operationalise the key requirements
- primarily addressed to developers and deployers of AI systems
- compliance with this list is not evidence of legal compliance
- piloting process (qualitative and quantitative assessment)

Examples of questions:

- In case of a chat bot or other conversational system, are the human end users made aware that they are interacting with a non-human agent?
- Did you ensure an oversight mechanism to log when, where, how, by whom and for what purpose data was accessed?
- Did you clearly communicate characteristics, limitations and potential shortcomings of the AI system?
- did you ensure a mechanism that allows others to flag issues related to bias, discrimination or poor performance of the AI system?

Towards trustworthy AI

Some insights:

- holistic approach, being open to changes (business models)
- diversity and inclusion (design, validation, deployment)
- disseminate results and communication to the public (realistic expectations, open questions)
- long term solutions, gradual and dynamic process (ethical culture)

Some weaknesses:

- being demonstrably trustworthy is hard
- some methods for implementing requirements are too abstract
- assessment list include too many questions
- risks of applying requirements/assessment list in a mechanical way

Thanks for your attention

Feedbacks, comments or requests are welcome

teresa.scantamburlo@unive.it