AI IN EVERYDAY LIFE

Unit 7 – Ethics

UNIVERSITÀ DEGLI STUDI
DI TRENTO
Dipartimento di Ingegneria
e Scienza dell'Informazione

DataScientia
Unitas per Varietatem

OPEN
UNIVERSITY OF
CYPRUS

CY. center for
algorithmic
transparency

OUTLINE

- Trustworthy AI
- Ethical principles for AI and examples
- Problems of AI
- Former solutions

USEFUL RESOURCES

AI Human Impact:
https://aihumanimpact.org/principles-toolkit.html

EC Ethics Guidelines for Trustworthy AI:
https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

AI POTENTIAL

- According to the EU High Level Expert Group on Artificial Intelligence, AI it is a tool that can improve the common good.
- AI can contribute to "facilitating the achievement of the United Nations' Sustainable Development Goals".

WHY ETHICS?

Evaluation

No right/wrong, but better/worse justifications for actions in terms of values

IN A NUTSHELL: TRUSTWORTHY AI

AI systems must be human-centered

- Maximize potential benefits
- Minimize risk

Fundamental ambition of the EU – Reliable AI

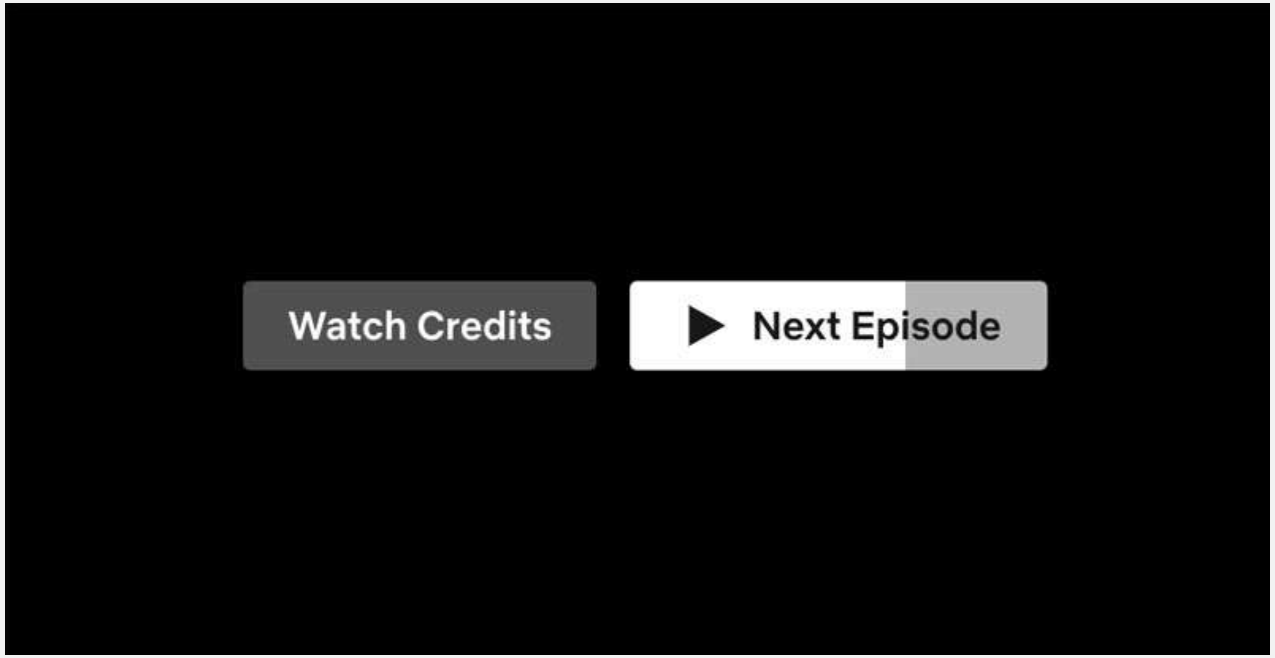
	AI Human Impact	EC Guidelines Trustworthy AI
Personal	Autonomy	Human agency and oversight
	Privacy	Privacy and data governance
Social	Fairness	Diversity, non-discrimination, fairness
	Society	Societal and environmental wellbeing
Technical	Performance	Technical robustness and safety
	Accountability	Accountability, Transparency

AUTONOMY

Autonomy is the right to self-determination and respects the individual's right to make informed decisions

«AI systems should empower human beings, allowing them to make informed decisions and fostering their fundamental rights» (EU Trustoworthy AI)

EXAMPLE: AUTONOMY VS AUTOPLAY



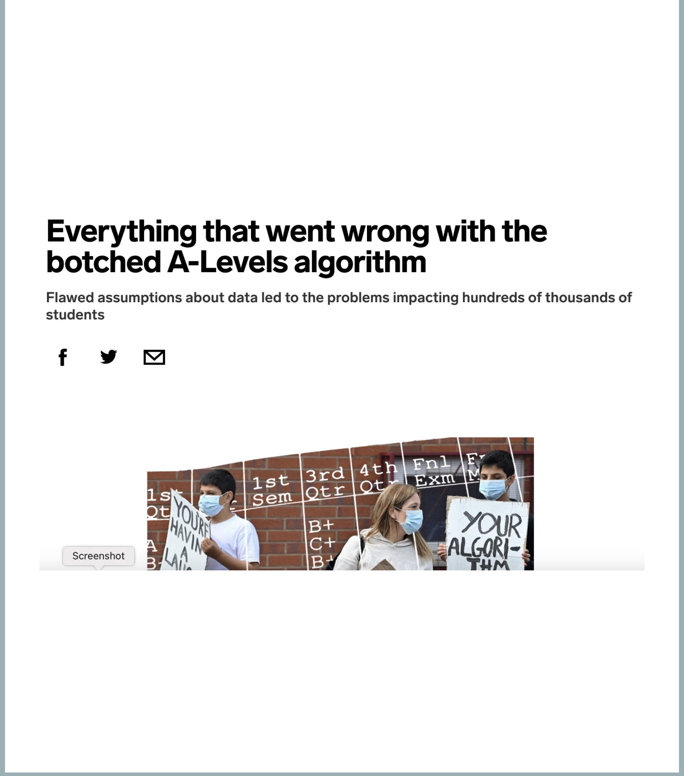
PRIVACY

Control over access to one's own personal information along the entire data process.

MULTIPLE IDENTITIES?

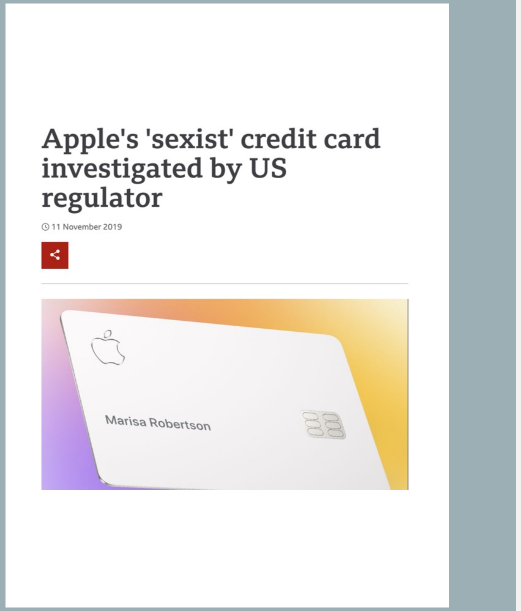
“You have one identity, The days of you having a different image for your work friends or co-workers and for the other people you know are probably coming to an end pretty quickly. […] Having two identities for yourself is an example of a lack of integrity.”

Mark Zuckerberg, interview with David Kirkpatrick («The Facebook Effect», pag. 222)



Everything that went wrong with the botched A-Levels algorithm

Flawed assumptions about data led to the problems impacting hundreds of thousands of students



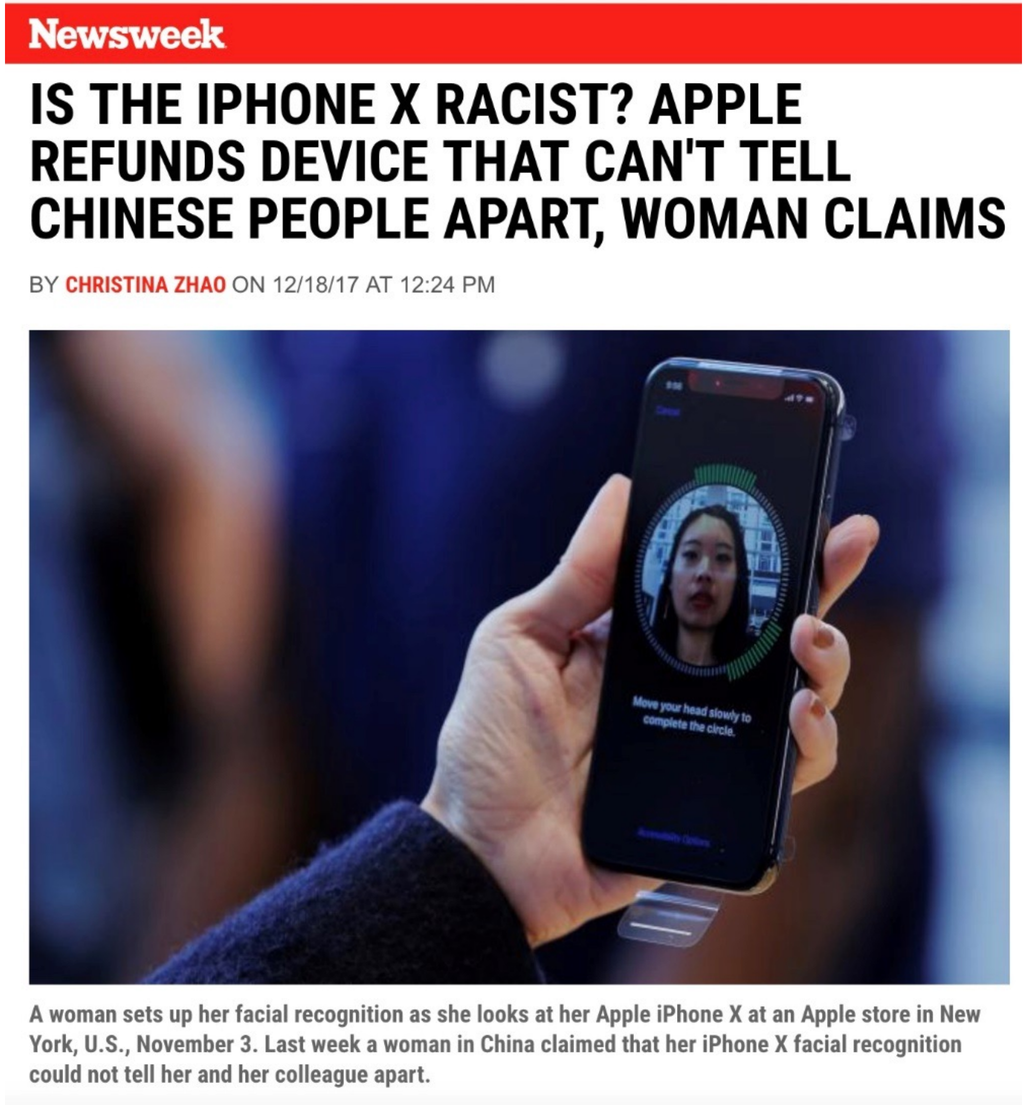
Apple's 'sexist' credit card investigated by US regulator

11 November 2019

Fairness

From Aristotle treats equal equally and unequal proportionately unequally

EXAMPLE 1: AUTOMATIC FACE RECOGNITION



Newsweek

IS THE IPHONE X RACIST? APPLE REFUNDS DEVICE THAT CAN'T TELL CHINESE PEOPLE APART, WOMAN CLAIMS

BY CHRISTINA ZHAO ON 12/18/17 AT 12:24 PM

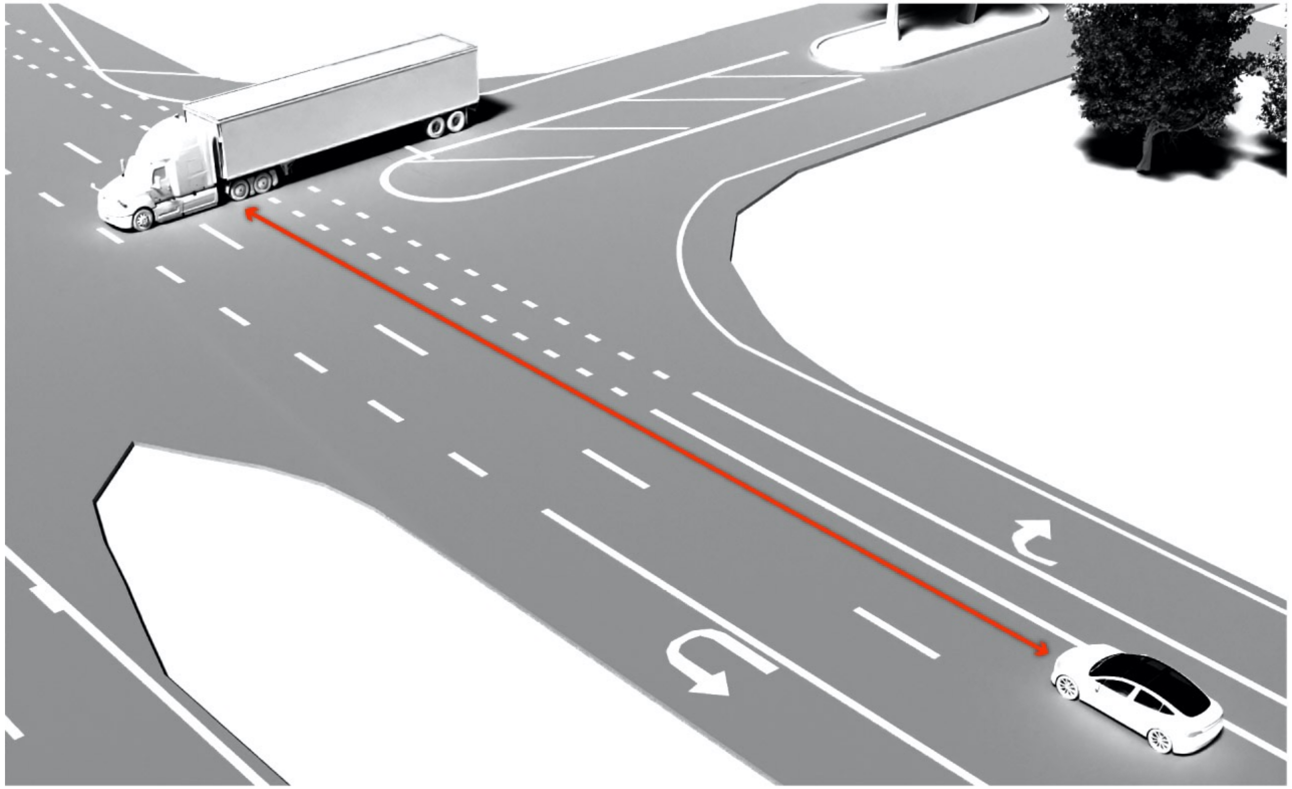
A woman sets up her facial recognition as she looks at her Apple iPhone X at an Apple store in New York, U.S., November 3. Last week a woman in China claimed that her iPhone X facial recognition could not tell her and her colleague apart.

ACCOUNTABILITY

«Explicability is crucial for building and maintaining users’ trust in AI systems. This means that processes need to be transparent, the capabilities and purpose of AI systems openly communicated, and decisions – to the extent possible – explainable to those directly and indirectly affected.»

FATAL TESLA AUTOPILOT CRASH

- **Source**: https://www.washingtonpost.com/technology/interactive/2023/tesla-autopilot-crash-analysis/



WHY AI HAS THESE PROBLEMS?

- “Data in the Wild”
  - Learning data that is not designed for our purpose
- Implicit feedback
  - E.g., we assume that a click on something is an indication of user interest but we do NOT have an explicit one
- Correlation instead of causation
  - Models that describe statistical correlations between variables but we don't know if there is a causal relationship

TOWARDS TRUSTWORTHY AI

EU REGULATION

- GDPR
- EU AI act: https://www.europarl.europa.eu/topics/en/article/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence
- Standards and ISO

SUPERVISION BY HUMANS

- Human-in-the-loop
  These are technical approaches that involve humans in every decision made by the system. Usually this is neither possible nor desirable.
- Human-in-control
  Some systems have built-in levels of human supervision. For example, in a military AI application/technology, the human operator may have the final say in the activation of system actions/implementation of system decisions.

NON-TECHNICAL METHODS FOR TRUSTWORTHY AI

- Codes of conduct
- Certification
- Education and awareness to foster an ethical mind-set
- Stakeholder participation and social dialogue

LINKS AND CONTACTS

- https://datascientiafoundation.github.io/datascientia-education-eai-2023-24-unitn
- http://knowdive.disi.unitn.it/
- 𝕏 @knowdive
- ✉ matteo.busso@unitn.it

THANK YOU!