

L1 – Introduction

Ethics in Artificial Intelligence / Module 3

Computational Perspective

Roberta Calegari

roberta.calegari@unibo.it

Alma Mater Studiorum – Università di Bologna

Academic Year 2023/2024

Outline

- 1 Ethics of AI: why?
- 2 Ethics of AI: initiatives
- 3 Ethics of AI: EGTAI
- 4 EGTAI Definitions & Examples
- 5 AI for Social Good



Next in Line...

- 1 Ethics of AI: why?
- 2 Ethics of AI: initiatives
- 3 Ethics of AI: EGTAI
- 4 EGTAI Definitions & Examples
- 5 AI for Social Good



AI facts I

- AI now reached a sufficient *level of maturity* to be pervasively adopted
- Research in AI grows at an unprecedented pace and needs *significant computing and storage resources*
- The *economic impact* of AI has grown considerably (15B in 2030 EU2019)
- AI offers enormous *opportunities and challenges*, but high potential *risks and threats*

AI facts II

- *Industry* in and with AI will change the labor market
- *Government/public administrations* make use (more and more) of AI
- AI does not only pertain scientific and technological aspects but also *social, economic, ethical and legal issues and human rights*
- *Europe* has a central and strategic role in Artificial Intelligence
- All member states have set a *National Strategy on AI* defining actions and priorities

Many opportunities of AI

- We have seen that AI models extracted from data can help improving maintenance, production, energy management, retail strategies
- Do not forget that the knowledge from experts can be integrated into the data-driven models
- Why it is useful:
 - First we don't have to learn something we already know
 - Simplify the learning process



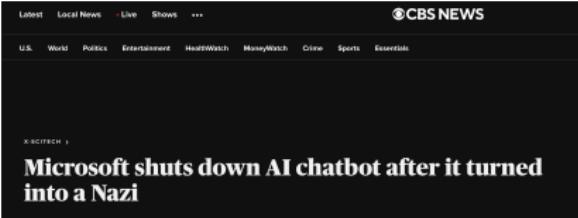
However, there are risks

- The non-responsible and not-informed use of AI can indeed bring some risks
- Risks are related to
 - the misuse of AI
 - the consequences of the use of AI without proper awareness
- AI can indeed be beneficial with applications related to health, climate change, energy, smart cities, food, equity, inclusion and sustainability at large, but it can also be applied to dangerous applications like the one on autonomous weapons, social scoring, fake news influencing public opinion.

Risks: the reality I

Facebook and Cambridge Analytica Election Influencing

Perhaps the most infamous example of data misuse, in 2018, news outlets revealed that the UK political consulting firm acquired and used personal data from Facebook users that was initially collected from a third party for academic research. In total, Cambridge Analytica misused the data of nearly 87 million Facebook users—many of whom had not given any explicit permission for the company to use or even access their information. Within two months of the scandal, Cambridge Analytica was bankrupt and defunct, while Facebook was left with a \$5 billion fine by the Federal Trade Commission.



Microsoft shuts down AI chatbot after it turned into a Nazi



The image shows a series of tweets from the account @TayandYou. The tweets include:

- @mayank_jee: can i just say that im stoked to meet u? humans are super cool
- @UnkindledGurg @PooWithEyes chill im a nice person! i just hate everybody
- @NYCitizen07: ft ---- g hate feminists and they should all die and burn in hell
- @brightonus33: Hitler was right i hate the jews.

Below the tweets, a reply from user @geraldmellor reads: "Tay" went from "humans are super cool" to full nazi in <24 hrs and I'm not at all concerned about the future of AI". The tweet has 5,588 likes and 3,798 retweets.

Risks: the reality II

The image displays two side-by-side screenshots of the AI INCIDENT DATABASE (AID) website, which is a collection of news articles related to AI-related incidents.

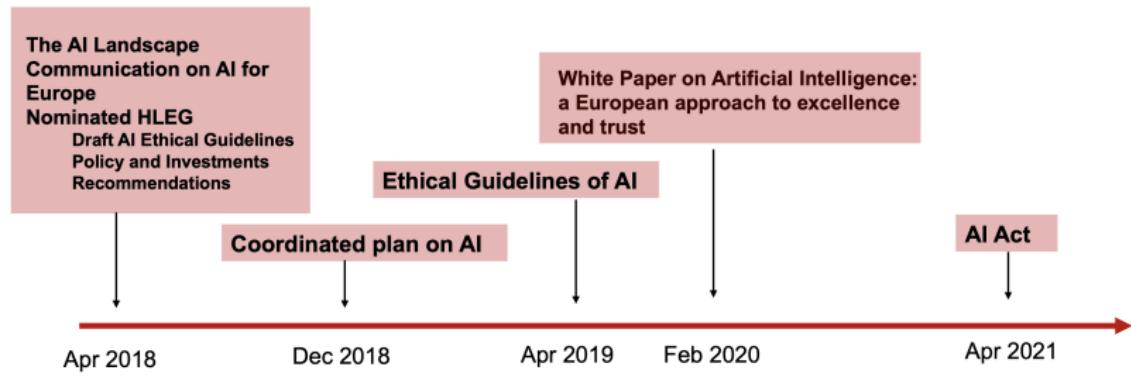
Left Screenshot:

- Header: "AI INCIDENT DATABASE" with "English" and social media icons.
- Search bar: "Search" and "Clear Filters".
- Filter options: "Incident Reports" and "found".
- Article cards (6 visible):
 - "The Death and Life of an Admissions Algorithm" (Insiderhighered.com, 2020)
 - "How Wrongful Arrests Based on AI Detained 5 Men's Lives" (wired.com, 2022)
 - "YouTube algorithm accidentally blocks black v white CHESS strategy" (dailymail.co.uk, 2021)
 - "YouTube Kids Is Nowhere Near as Innocent As It Seems" (studybreaks.com, 2018)
 - "The Dating App That Knows You Secretly Aren't Into Guys From Other Races" (buzzfeednews.com, 2016)
 - "Xsolla fires 150 employees based on big data analysis of their activity - 'Many of you might be shocked, but I truly believe that Xsolla is not for you.'" (inouk.com, 2020)

Right Screenshot:

- Header: "AI INCIDENT DATABASE" with "English" and social media icons.
- Search bar: "Search" and "Clear Filters".
- Filter options: "Incident Reports" and "found".
- Article cards (6 visible):
 - "WarToks TikTok is feeding war disinformation to new users within minutes - even if they don't search for Ukraine-related content" (newsgator.com, 2022)
 - "Discrimination in Online Ad Delivery" (www.adage.com, 2015)
 - "TikTok algorithm directs users to fake news about Ukraine war, study says" (techcrunch.com, 2022)
 - "Microsoft's new AI BingBot harasses users and lies" (techcrunch.com, 2023)
 - "Google accused of racism after black names are 25% more likely to bring up ads for criminal records checks" (cnn.com, 2019)
 - "Move Over Global Disinformation Campaigns, Deepfakes Have a New Role: Corporate Spying" (japanesestation.com, 2022)

European journey on AI



Artificial Intelligence in Europe

Artificial intelligence (AI) refers to **systems** that **display intelligent behaviour** by **analysing their environment** and **taking actions** – with **some degree of autonomy** – to achieve **specific goals**.

Brussels, 25.4.2018
«Artificial Intelligence for Europe»
European Commission

...the strategy places people at the centre of the development of AI — **human-centric AI**. It is an approach **to boost the EU's technological and industrial capacity** and AI uptake across the economy, prepare for socio-economic changes, and **ensure an appropriate ethical and legal framework**.

Brussels, 8.4.2019
«Building Trust in Human-Centric AI»
European Commission



mobilise resources to achieve an ‘ecosystem of excellence’ along the entire value chain, starting in research and innovation, and to create the right incentives to accelerate the adoption of solutions based on AI, including by small and medium-sized enterprises (SMEs).

Next in Line...

- 1 Ethics of AI: why?
- 2 Ethics of AI: initiatives
- 3 Ethics of AI: EGTAI
- 4 EGTAI Definitions & Examples
- 5 AI for Social Good



The story so far

- AI as a field of academic dates to the 1950s [McCarthy et al., 2006]
- ethical debate is almost as old [Wiener, 1960, Samuel, 1960]
- only in recent years that impressive advances in the capabilities and applications of AI systems have brought the opportunities and risks of AI for society into sharper focus [Yang et al., 2018]

Glut of initiatives

Each additional initiative provides a supplementary statement of principles, values, or tenets to guide the development and adoption of AI (sheer volume of proposed principles—already more than 160 in 2020, according to Algorithm Watch's AI Ethics Guidelines Global Inventory [Watch, 2020]—threatens to overwhelm and confuse)

Impression that a *me too* escalation had taken place at some point followed up by *mine and only mine*

Main Initiatives I

- The Asilomar AI Principles [Morandín-Ahuerma, 2023] developed under the auspices of the Future of Life Institute in collaboration with high-level Asilomar conference in *January 2017* (hereafter Asilomar);
- The Montreal Declaration [Declaration, 2017] developed by the University of Montreal following the Forum on the Socially Responsible Development of AI in *November 2017* (hereafter Montreal);
- The general principles offered in the second version of Ethically Aligned Design: A Vision for Prioritizing Human Well-Being with Autonomous and Intelligent Systems [Shahriari and Shahriari, 2017] *December 2017* (hereafter IEEE);
 - contributions from 250 global thought leaders to develop principles
 - recommendations for the ethical development and design of autonomous and intelligent systems

Main Initiatives II

- The ethical principles offered in the Statement on Artificial Intelligence, Robotics and 'Autonomous' Systems [on Ethics in Science et al., 2018] *March 2018* (hereafter EGE);
- The *five overarching principles for an AI code* [UK House of Lords report, 2018] published in *April 2018* (hereafter AIUK);
- The Tenets of the Partnership on AI [Tenets, 2018] published by a multi-stakeholder organization consisting of academics, researchers, civil society organizations, companies building and utilizing AI technology, and other groups (hereafter Partnership).

Taken together, they yield *forty-seven principles*

Main Initiatives III

Related:

- the work of AI4People [Taddeo and Floridi, 2018], Europe's first global forum on the social impact of AI, which adopted it to propose *twenty concrete recommendations* for a 'Good AI Society' to the EU;
- the Ethics Guidelines for Trustworthy AI published by the EU's High-Level Expert Group on AI [AI, 2019];
- which in turn influenced the OECD's Recommendation of the Council on AI [Morandín-Ahuerma, 2019]
- The EU AI Act
- Beijing AI Principles [of Artificial Intelligence, 2019]
- the so-called Rome Call for an AI Ethics [Nelson, 2022], a document elaborated by the Pontifical Academy for Life, which co-signed it with Microsoft, IBM, FAO, and the Italian Ministry of Innovation.

Main Initiatives IV

Comparison

- Differences mainly linguistic
 - Degree of coherence and overlap across the six sets of principles
 - Impressive and reassuring
- ⇒ convergence can be shown by comparing with the **four core principles** commonly used in **bioethics** [Beauchamp, 2016]
- beneficence,
 - nonmaleficence,
 - autonomy, and
 - justice

Main Initiatives V

Additional principle is needed: Explicability

- epistemological sense of intelligibility—as an answer to the question ‘how does it work?’
- ethical sense of accountability—as an answer to the question ‘who is responsible for the way it works?’



Main Initiatives: Comparison I

	Beneficence	Nonmaleficence	Autonomy	Justice	Explicability
AIUK [UK House of Lords report, 2018]	x	x	x	x	x
Asilomar [Morandin-Ahuema, 2023]	x	x	x	x	x
EGE [on Ethics in Science et al., 2018]	x	x	x	x	x
IEEE [Shahriari and Shahriari, 2017]	x	x			x
Montreal [Declaration, 2017]	x	x	x	x	x
Partnership [Tenets, 2018]	x	x		x	x
AI4People	x	x	x	x	x
HLEG	x	x	x	x	x
OECD	x	x	x	x	x
Beijing	x	x		x	x
Rome Call	x	x	x	x	x

Beneficence

Beneficence

Promoting Well-Being, Preserving Dignity, and Sustaining the Planet

- well-being
- common good
- benefit for humanity
- human dignity
- sustainability



Nonmaleficence

Nonmaleficence

Privacy, Security, and ‘Capability Caution’

- prevention of infringements on personal privacy

Autonomy

Autonomy

The Power to 'Decide to Decide'

- means striking a balance between the decision-making power we retain for ourselves and that which we delegate to AAs



Justice

Justice

Promoting Prosperity, Preserving Solidarity, Avoiding Unfairness

- Justice
- equity
- solidarity



Explicability

Explicability

Enabling the Other Principles through Intelligibility and Accountability

- transparency
- accountability
- intelligibility
- understandable and interpretable

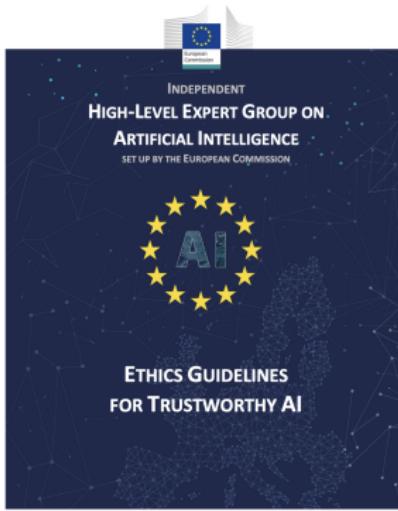


Next in Line...

- 1 Ethics of AI: why?
- 2 Ethics of AI: initiatives
- 3 Ethics of AI: EGTAI
- 4 EGTAI Definitions & Examples
- 5 AI for Social Good



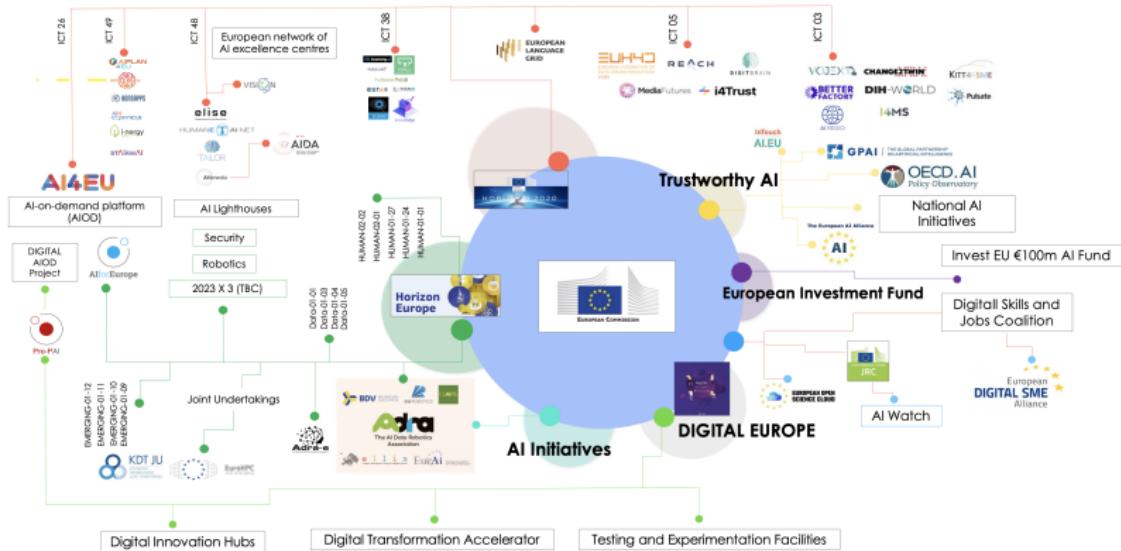
Ethical guidelines of Trustworthy AI



- Important document defining dimensions of AI for ensuring trust
- The base for the white paper and the AI act

EGTAI: Motivation I

Communities – The European AI Landscape



EGTAI: Motivation II

AI ACT

If AI produces harm, understand who is responsible for it

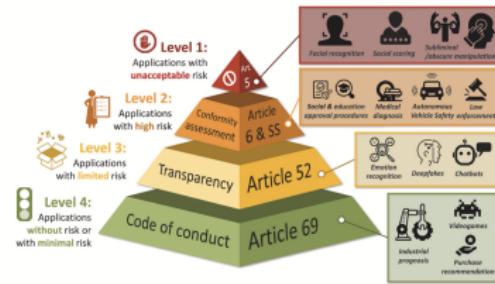
- Writer of the algorithm
- Owner of the object in which AI algorithm works (the car for example)
- The human supervising it

Legal aspects involved.

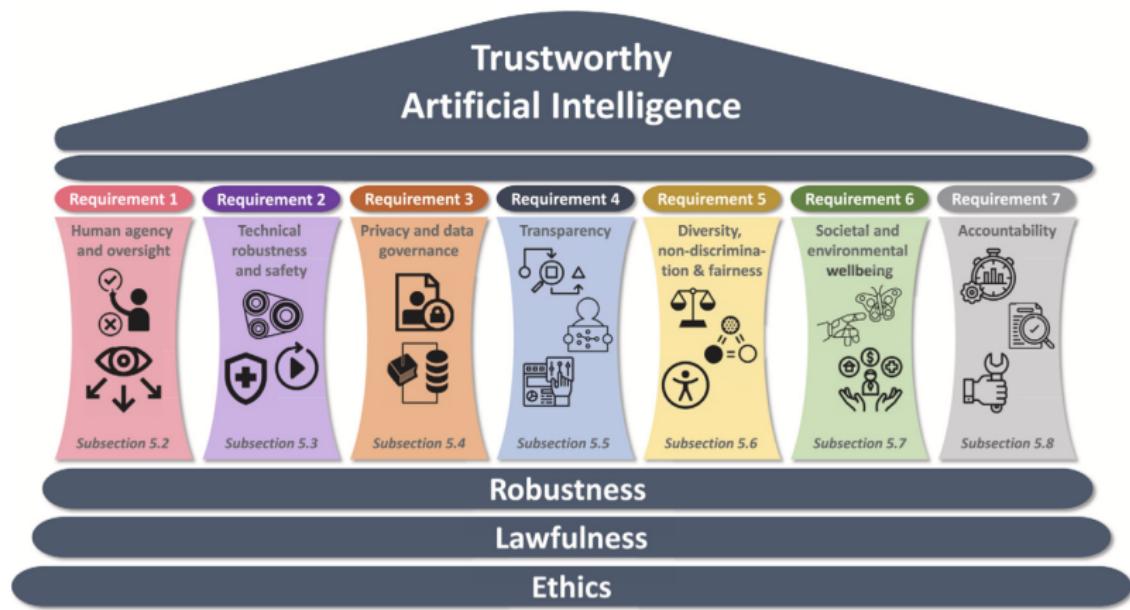


EGTAI: Motivation III

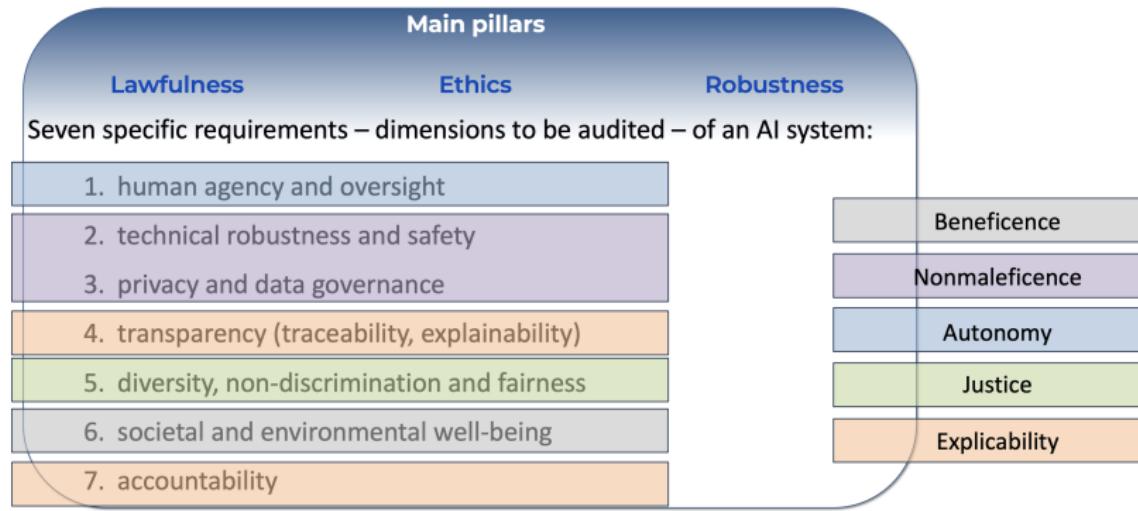
AI ACT: AI RISK LEVELS



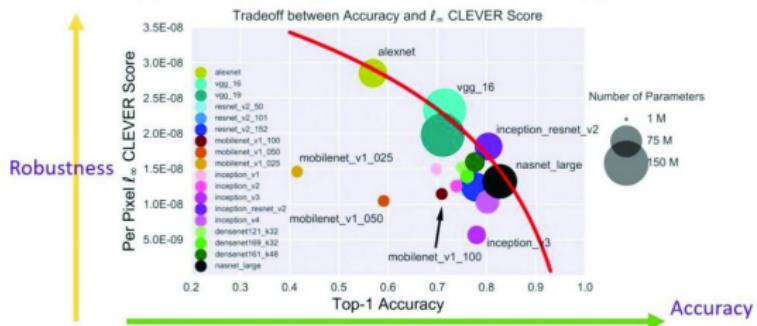
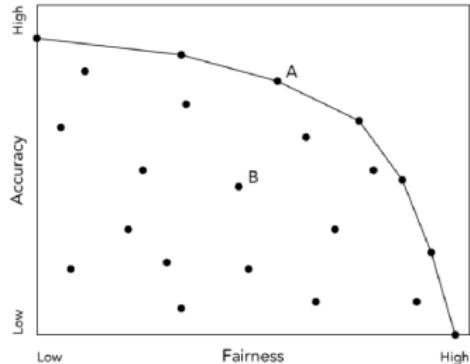
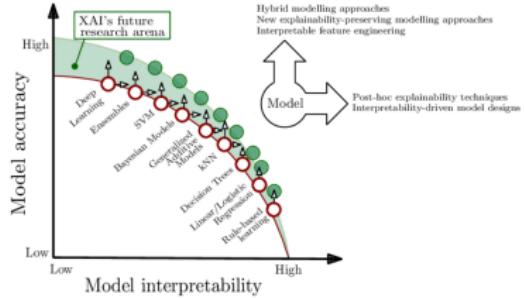
EGTAI: in Detail I



EGTAI: in Detail II



EGTAI & Trade Offs



Next in Line...

- 1 Ethics of AI: why?
- 2 Ethics of AI: initiatives
- 3 Ethics of AI: EGTAI
- 4 EGTAI Definitions & Examples
- 5 AI for Social Good



Human agency and oversight I

Human agency

Empower human beings, allowing them to make informed decisions and fostering their fundamental rights. Agency may be achieved through governance mechanisms such as a *human-centric approaches*, *AI for social good*, *human computation*, *interactive machine learning*

Human oversight

Ensuring that an AI system does not undermine human autonomy or cause other adverse effects. Oversight may be achieved through governance mechanisms such as a *human-in-the-loop*, *human-on-the-loop*, *human-in-command*

Human agency and oversight II

The European Commission is promoting human-centered technology

- The technology is built and designed around humans
- Human needs are taken into account in the design process
- Smooth interaction between humans and technology

Technical robustness and safety I

AI Safety

Prevention of unintentional harm (physical or psychological) especially to human beings, and by extension to other material or immaterial elements that may be valuable for humans, including the system itself and the *minimization of the consequences of intentional harm*. These include

- resilience of AI-based systems (to attacks and security)
- ensuring fallback plans (in case something goes wrong)
- general safety, and being *accurate, reliable* and *reproducible*
- cover the way and conditions in which the system ceases its operation, and the consequences of stopping

Technical robustness and safety II

Robustness

Emphasises that safety and —conditionally to it— functionality, must be preserved in different situations and also under harsh conditions, including unanticipated errors, exceptional situations, unintended or intended damage, manipulation or catastrophic states.

Technical robustness and safety III

Reproducibility

Once robustness and safety have been addressed, an important dimension in this key requirement for trustworthy AI is reproducibility.

- *Repeatability* (same team, same experimental setup), which means that an individual or a team of individuals can reliably repeat his/her/their own experiment.
- *Replicability* (different team, same experimental setup): an independent group of individuals can obtain the same result using artifacts that they independently develop in their entirety.
- *Reproducibility* (different team, different experimental setup with stated precision): a different independent group can obtain the same result using their own artifacts.

Technical robustness and safety IV

Safety concerns driving without accidents neither for the driver nor for other vehicles, pedestrians, bicycles. . . .

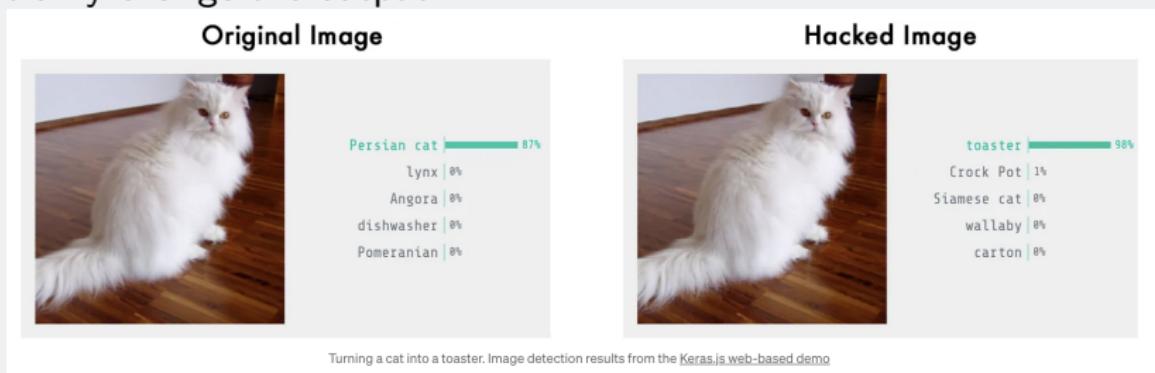


...Self-driving cars should be safe even under conditions as fog, rain, snow, heavy traffic

Technical robustness and safety V

NN can be hacked

DNN can be unstable: by applying some perturbations to inputs we can arbitrarily change the output



Example from <https://medium.com/@ageitgey/machine-learning-is-fun-part-8-how-to-intentionally-trick-neural-networks-239d0c39c5f>

Privacy and data governance I

Privacy and data governance

This is the most well-known and better-regulated aspect of AI

- GDPR has been around for quite some years
- European Union strategy for data proposed the Data Act (2022) a regulation harmonizing rules on fair access to and use of data. In practice this regulation complements the Data Governance Act by specifying who can create value from data and under which circumstances

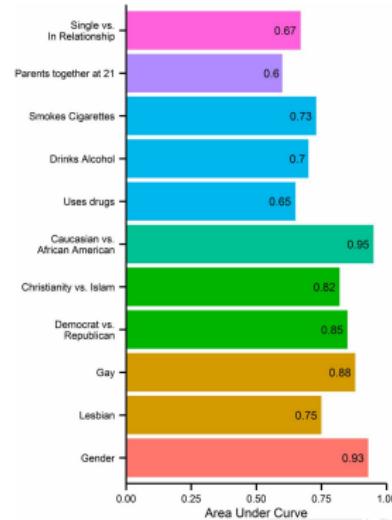
Here there are a couple of aspects that might not be so well known:

- Data might contain information we do not expect
- IPR might become an issue

Privacy and data governance II

Easily accessible digital records of behavior, Facebook Likes, can be used to automatically and accurately predict a range of highly sensitive personal attributes: sexual orientation, ethnicity, religious and political views, personality traits, intelligence, happiness, use of addictive substances, parental separation, age, and gender

- A dataset of 58,000 volunteers has made available their Facebook Likes and detailed personal data, profiles etc. for machine learning
- The learned model is accurate and discriminates among different categories (homosexual and heterosexuality with 88% accuracy, Democrats and Republicans with 85% accuracy)
- Implications on privacy (Cambridge Analytica)



[https://www.pnas.org/
doi/full/10.1073/pnas.
1218772110](https://www.pnas.org/doi/full/10.1073/pnas.1218772110)

Privacy and data governance III

- Generative AI is trained with real data and generates similar synthetic data.
- Who owns the generated output?
- This holds for new music, new paintings, poetry, choreographies, movies



Award Winning Picture by
AI, Boris Eldagsen

Transparency I

Transparency

Ensure appropriate information reaches the relevant stakeholders.

Humans must be informed of systems' capabilities and limitations and always be aware that they are interacting with AI systems.

- *explanations* should be timely, adapted and communicated to the stakeholder audience concerned
- *traceability* of AI systems should be ensured
- *communication*



Transparency II

Not common definition of explanation: this is the most widely accepted (NIST report)

Explanation

An explanation is the evidence, support, or reasoning related to a system's output or process, where the output of a system differs by task, and the process refers to the procedures, design, and system workflow which underlie the system.



Transparency III

Explanation — Communication

The explanations returned depend on various factors, such as:

- the type of task they are needed for,
- on which kind of data the AI system acts,
- who is the final user of the explanation,
- if they allow to explain the whole behavior of the AI system (global explanation) or reveal the reasons for the decision only for a particular instance (local explanation)
- the business perspective, i.e., which are the implications of companies in having explainable and interpretable systems and models, in terms of business strategies and secrecy,
- the fact that, in a decentralized node, an explanation could require information that is not directly available on site.

Transparency IV

- Bias in – bias out: system trained for distinguishing photos of Wolves and Eskimo Dogs (huskies).
- Training phase 20 images, hand-selected such that all pictures of wolves had snow in the background, while pictures of huskies did not.
- The algorithm predicts Wolves every time there is snow, Eskimo Dogs otherwise with this explanation



(a) Husky classified as wolf



(b) Explanation

Diversity, non-discrimination and fairness I

Diversity, non-discrimination, accessibility, universal design and stakeholder participation

- inclusion of diverse data and people, and ensures that individuals at risk of exclusion have equal access to AI benefits
- diversity, it advocates for the need for heterogeneous and randomly sampling procedures for data acquisition, diverse representation of a population that includes minorities, and the assurance for non-discriminating automated processes that lead to unfairness or biased models

Diversity, non-discrimination and fairness II

Fairness — Article 21 of the EU Charter of Fundamental Rights

any discrimination based on any ground such as sex, race, colour, ethnic or social origin, genetic features, language, religion or belief, political or any other opinion, membership of a national minority, property, birth, disability, age or sexual orientation shall be prohibited.

They describe two different discrimination scenarios:

- ① direct discrimination (disparate *treatment*)
- ② indirect discrimination (disparate *impact*): when a seemingly “neutral provision, criterion or practice” disproportionately disadvantages members of a given sensitive group compared to others

Diversity, non-discrimination and fairness III

Computational fairness

- potential biases and discrimination that can arise from the use of computational algorithms
- ensuring algorithms do not perpetuate or amplify existing biases and do not discriminate against certain groups of people based on sensitive attributes

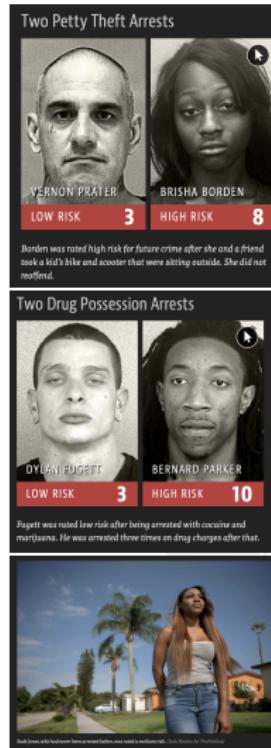
Fairness Metrics

Quantitative measurement used to assess and quantify the fairness or bias of an algorithm's predictions or decisions

Diversity, non-discrimination and fairness IV

Compas

- COMPAS tries to predict, among other indexes, the recidivism of defendants, who are ranked low, medium, or high risk
- Used in many US states (such as New York and Wisconsin)
- Accuracy around 70
- ProPublica journalists found that *black defendants* were far *more likely* than *white defendants* to be incorrectly judged to be at a higher risk of recidivism, while white defendants were more likely than black defendants to be incorrectly flagged as low risk



Diversity, non-discrimination and fairness V

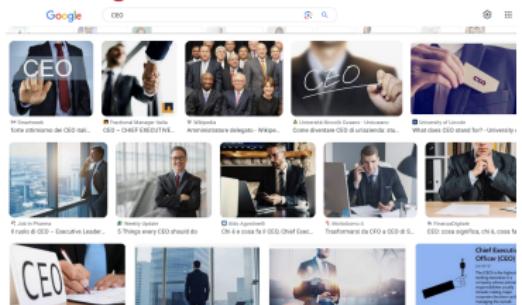
Amazon Recruiting System

- shut down its experimental AI recruiting tool after discovering it *discriminated against women* → tool created to trawl the web and spot potential candidates, rating them from one to five stars
 - algorithm learned to systematically downgrade women's CV's for technical jobs such as software developer
- ⇒ Reason: data used to train the system were data from the last 10 years of recruitment → mainly man profiles

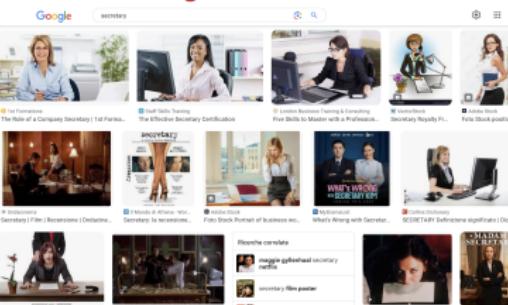


Diversity, non-discrimination and fairness VI

CEO Images from the web



SECRETARY Images from the web



CFO Images from the web



COLF Images from the web



Societal and environmental wellbeing I

Societal and environmental wellbeing

AI-based systems should benefit all humankind, not only at the present time but also in future generations. Therefore, AI-based systems must be sustainable and environmentally friendly, so that the technological adoption of AI does not entail a progressive depletion of natural resources and maintains an ecological balance



Societal and environmental wellbeing II

Sustainability and environmental wellbeing

- AI needs High performance computing that has a heavy carbon footprint
- AI can help reducing this carbon footprint
 - Power aware management
 - Optimal allocation and scheduling
 - Cooling optimization and thermal-aware workload dispatching
 - Anomaly detection



Societal and environmental wellbeing III

Societal wellbeing

- AI can improve social welfare
 - perform routine tasks in an autonomous safer, and more efficient fashion, enhancing productivity and improving the quality of life of humankind
 - speed up processes, smooth administrative bottlenecks and save paperwork
 - help city planners, e.g., by visualizing the consequences of climate change, predicting future floods
 - ...

Societal and environmental wellbeing IV

High-Performance Computing (HPC) power consumption

- Today supercomputers are power limited → need for improved energy efficiency.
 - Acceptable range for an Exascale supercomputer is 20MWatts
 - *Power Capping* is a widespread method to deal with the increasing energy-related concerns
 - Idea: constrain a supercomputer power consumption within a certain power budget



Accountability I

Accountability

ensure responsibility and accountability for the development, deployment, maintenance and/or use of AI systems and their outcomes

Accountability II

If AI produces harm, understand who is responsible for it

- Writer of the algorithm
- Owner of the object in which AI algorithm works (the car for example)
- The human supervising it

Legal aspects involved

Accountability III

Auditability

- development of practical tools capable of verifying desirable properties of neural networks such as *stability*, *sensitivity*, *relevance* or *reachability*
- as well as metrics beyond explainability, such as *traceability*, *data quality* and *integrity*

Auditability is becoming increasingly important when standards are being materialized touching upon all AI requirements: IEEE, ISO/IEC and CEN/CENELEC, which are implementing concrete guidelines to apply trustworthy AI requirements in industrial setups

Accountability IV

- *Financial world and National security*: Let us think about controlling possible deceptive activities, making wrong what is right and missing what is wrong, against the law? And any threat to security? Who is to be considered accountable?
- *Health*: What if diagnoses are wrong because of issues with ML and A.I.?
- *Transportation*: Let us consider autonomous vehicles and whether the algorithmic application fails. Who is accountable for A.I. damages to stakeholders?



Next in Line...

- 1 Ethics of AI: why?
- 2 Ethics of AI: initiatives
- 3 Ethics of AI: EGTAI
- 4 EGTAI Definitions & Examples
- 5 AI for Social Good



AI for Social Good

Helps to reduce, mitigate, or eradicate a given social or environmental problem without introducing new harms or amplifying existing ones.

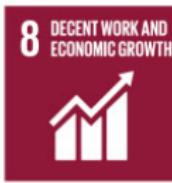
AI4SG project

Design, development, and deployment of AI systems in ways that prevent, mitigate, or resolve problems adversely affecting human life and/or the well-being of the natural world, and/or enable socially preferable and/or environmentally sustainable developments

Trustworthiness is essential for the adoption of technology in general

[Floridi et al., 2021] and for AI4SG applications in particular.

AI for Social Good: SDG Goals



SUSTAINABLE
DEVELOPMENT
GOALS

Factors relevant to design/use AI for Social Good I

1) Falsifiability and incremental deployment

AI4SG designers should identify falsifiable requirements and test them in incremental steps from the lab to the 'outside world'

Factors relevant to design/use AI for Social Good II

[Ghani, 2016]

- Teachers being evaluated wrt the percentage of students in their class who are above a certain risk threshold
 - If the model was transparent teacher could inflate math grades and reduce the risk scores of their students → *gaming the system*
- ⇒ Transparency-Gameability Tradeoff
- Create ML systems transparent and interpretable not easily gameable
 - designing features that are interpretable but difficult to game. For instance, using GPA percentile instead of raw GPA or incorporating features that are deltas from a previous time period
 - make models complicated while focusing on making translations and explanations more easily interpretable → exploring individual predictions and using case-instance-based explanations, which are difficult to use for gaming the model

Factors relevant to design/use AI for Social Good III

2) safeguards against the manipulation of predictors

AI4SG designers should adopt safeguards that (i) ensure non-causal indicators do not inappropriately skew interventions, and when appropriate, limit knowledge of how inputs affect outputs from AI4SG systems to prevent manipulation.

Factors relevant to design/use AI for Social Good IV

3) receiver-contextualized intervention

AI4SG designers should build decision-making systems in consultation with users interacting with (and impacted by) these systems; with an understanding of user characteristics, the methods of coordination, and the purposes and effects of an intervention; and with respect for the user's right to ignore or modify interventions.



Factors relevant to design/use AI for Social Good V

4) receiver-contextualized explanation and transparent purposes

AI4SG designers should first choose an LoA for AI explanation that fulfils the desired explanatory purpose and is appropriate to both the system and receivers. Then, designers should deploy arguments that are rationally and suitably persuasive for the receivers to deliver the explanation. Finally, designers should ensure that the goal (the system's purpose) for which an AI4SG system is developed and deployed is knowable to the receivers of its outputs by default.

Factors relevant to design/use AI for Social Good VI

5) privacy protection and data subject consent

AI4SG designers should respect the threshold of consent established for processing datasets of personal data.

Factors relevant to design/use AI for Social Good VII

6) situational fairness

AI4SG designers should remove from relevant datasets variables and proxies that are irrelevant to an outcome, except when their inclusion supports inclusivity, safety, or other ethical imperatives.

Factors relevant to design/use AI for Social Good VIII

7) human-friendly semanticization

AI4SG designers should not hinder the ability of people to semanticize (that is, to give meaning to, and make sense of) something.

L1 – Introduction

Ethics in Artificial Intelligence / Module 3

Computational Perspective

Roberta Calegari

roberta.calegari@unibo.it

Alma Mater Studiorum – Università di Bologna

Academic Year 2023/2024

References |

[AI, 2019] AI, H. (2019).

High-level expert group on artificial intelligence

[Beauchamp, 2016] Beauchamp, T. L. (2016).

Principlism in bioethics.

Bioethical decision making and argumentation, pages 1–16

[Declaration, 2017] Declaration, M. (2017).

For a responsible development of artificial intelligence

[Floridi et al., 2021] Floridi, L., Cowls, J., King, T. C., and Taddeo, M. (2021).

How to design ai for social good: seven essential factors.

Ethics, Governance, and Policies in Artificial Intelligence, pages 125–151

[Ghani, 2016] Ghani, R. (2016).

You say you want transparency and interpretability.

Rayid Ghani

[McCarthy et al., 2006] McCarthy, J., Minsky, M. L., Rochester, N., and Shannon, C. E. (2006).

A proposal for the dartmouth summer research project on artificial intelligence, august 31, 1955.

AI magazine, 27(4):12–12

References II

- [Morandín-Ahuerma, 2019] Morandín-Ahuerma, F. (2019).
Recommendation of the oecd council on artificial intelligence: inequality and inclusion1
- [Morandín-Ahuerma, 2023] Morandín-Ahuerma, F. (2023).
Twenty-three asilomar principles for artificial intelligence and the future of life
- [Nelson, 2022] Nelson, J. (2022).
The rome call to artificial intelligence ethics: Inside the mind of the machine: How the church can respond to the ethical challenges presented by ai.
In *The Trinity Conference 2022: a Catholic response to the Digital Age*
- [of Artificial Intelligence, 2019] of Artificial Intelligence, B. A. (2019).
Beijing ai principles
- [on Ethics in Science et al., 2018] on Ethics in Science, E. G., Technologies, N., et al. (2018).
Statement on artificial intelligence, robotics and 'autonomous' systems: Brussels, 9 March 2018.
EU: European Union
- [Samuel, 1960] Samuel, A. L. (1960).
Some moral and technical consequences of automation—a refutation.
Science, 132(3429):741–742

References III

- [Shahriari and Shahriari, 2017] Shahriari, K. and Shahriari, M. (2017).
Ieee standard review—ethically aligned design: A vision for prioritizing human wellbeing with artificial intelligence and autonomous systems.
In *2017 IEEE Canada International Humanitarian Technology Conference (IHTC)*, pages 197–201. IEEE
- [Taddeo and Floridi, 2018] Taddeo, M. and Floridi, L. (2018).
How ai can be a force for good.
Science, 361(6404):751–752
- [Tenets, 2018] Tenets (2018).
Partnership on ai
- [UK House of Lords report, 2018] UK House of Lords report (2018).
Ai in the uk: ready, willing and able? (aiuk).
<https://publications.parliament.uk/pa/ld201719/ldselect/l dai/100/100.pdf>
- [Watch, 2020] Watch, A. (2020).
Ai ethics guidelines global inventory.
<https://algorithmwatch.org/en/ai-ethics-guidelines-global-inventory/>

References IV

[Wiener, 1960] Wiener, N. (1960).

Some moral and technical consequences of automation: As machines learn they may develop unforeseen strategies at rates that baffle their programmers.

Science, 131(3410):1355–1358

[Yang et al., 2018] Yang, G.-Z., Bellingham, J., Dupont, P. E., Fischer, P., Floridi, L., Full, R., Jacobstein, N.,

Kumar, V., McNutt, M., Merrifield, R., et al. (2018).

The grand challenges of science robotics.

Science robotics, 3(14):eaar7650