Universitat Oberta
de Catalunya

Barcelona, May 19th 2023

Preventing biased results and discriminatory effects of AI use in Public Administration

uoc.edu

Agustí Cerrillo i Martínez

Bias, discrimination, and other risks of the use of artificial intelligence

- Some of the risks that can entail the use of IA in PA are due to the errors which often appear in algorithms or in the results obtained
- Errors in algorithms can have many consequences and negative effects on people and on PA
- AI can generate bias → skewed or incorrect results which do not represent reality for the people analysed
- Bias is the degree to which a reference value deviates from the truth (ISO)

Bias, discrimination, and other risks of the use of artificial intelligence

- Bias may negatively impact people’s right to equality, by generating discrimination
- When bias are not provided with an objective and reasonable cause justifying why people receive different treatment, this can lead to adopting discriminatory public decisions
- Eurobarometer 92 (2019): 36% of those surveyed were concerned that the use of AI could lead to discrimination on the basis of age, gender, or nationality

Bias, discrimination, and other risks of the use of artificial intelligence

- Bias and discrimination are not new in public administration
- The automation of public administrations has generated bias and discrimination in the past which have been strengthened by the use of AI because it can affect more people

- Gender-based discrimination
- Racial discrimination
- Functional diversity discrimination

Bias, discrimination, and other risks of the use of artificial intelligence

Principle of equality and non-discrimination

- The European Union Charter of Fundamental Rights recognises equality as one of the foundational values of the Union
- It also recognises the rights of all people to equality before the law (Article 20) and prohibits any discrimination based on gender, race, skin colour, ethnic or social origin, … (Article 21). Special mention of the guarantee of equality between men and women in all spheres (Article 23) and for people with disabilities (Article 26).
- No European regulations deal specifically with the use of AI as it relates to the principle of equality
- Current jurisprudence has not ruled specifically on this matter

Bias, discrimination, and other risks of the use of artificial intelligence

Why can AI cause bias and discrimination?

1. Bias in decisions regarding artificial intelligence

- Bias when deciding what algorithms to use, or how to use them

- Bias when defining of technology (chatbots or virtual assistants with female avatars)

- Bias when public officials are not aware of the risk of discrimination

Proposed Artificial Intelligence Act
People supervising artificial intelligence must be “aware of the possible tendency of automatically relying or over-relying on the output produced by a high-risk AI system (‘automation bias’)” (Article 14.4.b).

Bias, discrimination, and other risks of the use of artificial intelligence

Why can artificial intelligence cause bias and discrimination?

2. Algorithm design

- Bias transferred to algorithms via their designers when the models are constructed

- Most engineers and programmers are men, and intentionally or unintentionally, they may programme gender discrimination into the algorithms

Bias, discrimination, and other risks of the use of artificial intelligence

Why can artificial intelligence cause bias and discrimination?

3. Data quality

- Data availability, quality, and reusability is crucial for AI
- Sometimes the data are not high quality

Faces used to train facial recognition algorithms were 75% male and over 80% white. Because of this facial recognition errors were less than 1% in the case of light-skinned men, 7% for white women, 12% for dark-skinned men, and up to 35% for dark-skinned women (Buolamwini & Gebru, 2018).

White Paper on AI: “Without data, the development of AI and other digital applications is not possible. The enormous volume of new data yet to be generated constitutes an opportunity for Europe to position itself at the forefront of the data and AI transformation”

European Regulation of algorithmic equality

At present there are no specific regulations on the use of AI in PA, and no measures have been provided for dealing with the prohibition of algorithmic discrimination

- Declaration on Digital Rights and Principles for the Digital Decade (December 2022)

> “It should benefit everyone, achieve gender balance, and include notably elderly people, people living in rural areas, persons with disabilities, or marginalised, vulnerable or disenfranchised people and those who act on their behalf”

- Charter of Digital Rights (June 2021)

> “The right to algorithmic non-discrimination shall be guaranteed, regardless of the origin, cause or nature of the bias, with regard to decisions and processes based on algorithms”

Trustworthiness as a guarantee against algorithmic bias

"Trustworthiness is a prerequisite for people and societies to develop, deploy and use AI systems"

Independent High-Level Expert Group on Artificial Intelligence (2019)

Trustworthiness

- robustness
- legality
- human supervision
- transparency
- accountability
- prohibition of any discrimination
- governance

Trustworthiness of algorithms in the proposed Artificial Intelligence Act

1. Algorithm robustness

Fulfilment of the tasks for which it was designed, that it runs securely and reliably, and that it does not produce unwanted or unexpected results

> (50) The technical robustness is a key requirement for high-risk AI systems. They should be resilient against risks connected to the limitations of the system (e.g. errors, faults, inconsistencies, unexpected situations) as well as against malicious actions that may compromise the security of the AI system and result in harmful or otherwise undesirable behaviour. Failure to protect against these risks could lead to safety impacts or negatively affect the fundamental rights, for example due to erroneous decisions or wrong or biased outputs generated by the AI system.
> ***Users of the AI system should take steps to ensure that the possible trade-off between robustness and accuracy does not lead to discriminatory or negative outcomes for minority subgroups***

Trustworthiness of algorithms in the proposed Artificial Intelligence Act

1. Algorithm robustness

Article 15

“High-risk AI systems shall be designed and developed ***following the principle of security by design and by default*** in such a way that they achieve, in the light of their intended purpose, an appropriate level of accuracy, robustness, ***safety*** and cybersecurity, and perform consistently in those respects throughout their lifecycle. ***Compliance with these requirements shall include implementation of state-of-the-art measures, according to the specific market segment or scope of application***.”

“***Technical and organisational measures shall be taken to ensure that*** High-risk AI systems shall be resilient as regards errors, faults or inconsistencies that may occur within the system or the environment in which the system operates, in particular due to their interaction with natural persons or other systems. The robustness of high-risk AI systems may be achieved through technical redundancy solutions, which may include backup or fail-safe plans”.

“High-risk AI systems that continue to learn after being placed on the market or put into service shall be developed in such a way to ensure that possibly biased outputs due to outputs used as an input for future operations (‘feedback loops’) ***and malicious manipulation of inputs used in learning during operation*** are duly addressed with appropriate mitigation measures”.

Trustworthiness of algorithms in the proposed Artificial Intelligence Act

1. Algorithm robustness

Article 19 (**DELETED**)

“Providers of high-risk AI systems shall ensure that their systems undergo the relevant conformity assessment procedure” and that when the artificial intelligence systems comply with the requirements set out in the Artificial Intelligence Act their providers will draw up an EU declaration of conformity (Article 48) and they will affix the CE marking of conformity (Article 49).

Trustworthiness of algorithms in the proposed Artificial Intelligence Act

2. Algorithm transparency

Understanding how algorithms are designed, how they function, and what data are used to obtain a result is useful for keeping them free from bias and thus preventing discrimination.

Thanks to this, anyone can identify whether an algorithm is biased or if the results it obtains are discriminatory, and if necessary, prove the existence of discrimination

Articles 13 and 52 Proposed Artificial Intelligence Act

Trustworthiness of algorithms in the proposed Artificial Intelligence Act

3. Data and algorithm quality

“if we give these [algorithmic] systems biased data, they will be biased” (Giannandrea, 2017)

“if the AI system is not trained with high quality data, does not meet adequate requirements in terms of its accuracy or robustness, or is not properly designed and tested before being put on the market or otherwise put into service, it may single out people in a discriminatory or otherwise incorrect or unjust manner”,

“Training, validation and testing data sets should be sufficiently relevant, representative and free of errors and complete in view of the intended purpose of the system”
(Recital 38).

Trustworthiness of algorithms in the proposed Artificial Intelligence Act

3. Data and algorithm quality

“High data quality is essential for the performance of many AI systems, especially when techniques involving the training of models are used, with a view to ensure that the high-risk AI system performs as intended and safely and it does not become the source of discrimination prohibited by Union law”

Biases can for example be inherent in underlying datasets, especially when historical data is being used, introduced by the developers of the algorithms, or generated when the systems are implemented in real world settings. Results provided by AI systems are influenced by such inherent biases that are inclined to gradually increase and thereby perpetuate and amplify existing discrimination, in particular for persons belonging to certain vulnerable or ethnic groups, or racialised communities.

Negative bias should be understood as bias that create direct or indirect discriminatory effect against a natural person The requirements related to data governance can be complied with by having recourse to third-parties that offer certified compliance services including verification of data governance, data set integrity, and data training, validation and testing practices.

(Recital 44).

Trustworthiness of algorithms in the proposed Artificial Intelligence Act

3. Data and algorithm quality

Data governance

“High quality training, validation and testing data sets require the implementation of appropriate data governance and management practices” (Recital 44).

These criteria are stated in Article 10.

Trustworthiness of algorithms in the proposed Artificial Intelligence Act

4. Human supervision

The need for people to supervise them in action, and thus help to avoid the errors and discrimination algorithms may generate.

Human supervision may take very different forms

- the decisions of public administrations should always be made by a person, although they may be based on the results provided by an algorithm
- participation by interested parties in the various phases of the algorithm life cycle (from design, to assessment of the algorithm and its results, to legal monitoring of the algorithm's decisions)

The intensity of the supervision must be linked to the type of decision made, how much it affects personal rights, and the degree of automation of the process

Trustworthiness of algorithms in the proposed Artificial Intelligence Act

4. Human supervision

Article 14.4.b of the proposed Artificial Intelligence Act which state that the measures for human oversight adopted shall enable the individuals to whom human oversight is assigned to remain aware of the possible tendency of automatically relying or over-relying on the output produced by a high-risk artificial intelligence system (‘automation bias’).

The person supervising the algorithm in use must have the data, the expertise, and sufficient capacity to be able to challenge the proposal of the algorithm and to identify or eliminate any bias.

Trustworthiness of algorithms in the proposed Artificial Intelligence Act

5. Artificial intelligence governance

Strengthen governance of artificial intelligence in public bodies.

Some experiences (ie., Spanish Agency for the Supervision of Artificial Intelligence (AESIA) to minimise the possible risks to people’s security, safety, and fundamental rights from the use of artificial intelligence)

Problems that can arise in their functions (ie. New York).

Proposed Artificial Intelligence Act

> “National competent authorities shall be established or designated by each Member State for the purpose of ensuring the application and implementation of this Regulation. National competent authorities shall be organised so as to safeguard the objectivity and impartiality of their activities and tasks” (Article 59).

Universitat Oberta
de Catalunya

✉ acerrillo@uoc.edu

Thank you for your atention