Kitti Mezei
Budapest University of Technology and Economics, Faculty of Economics and Social Sciences, Business Law Department, Centre for Social Sciences, Institute for Legal Studies;

Nóra Bán-Forgács
Milton Friedman University, Centre for Social Sciences, Institute for Legal Studies;

Discrimination in the age of algorithms

Abstract
The application of artificial intelligence (AI) is expanding into more and more areas of life (e.g., it can improve healthcare, help law enforcement authorities fight crime more effectively, make transport safer, or even help detect fraud and cybersecurity threats, etc.). It is therefore undoubtedly one of the biggest challenges of our time, both from an economic and regulatory perspective. Not least because the European Commission has published a White Paper on Artificial Intelligence in 2020, which will form the basis for specific regulation of AI developments and applications at EU level.

The most important step forward in the regulation of AI is the publication in April 2021 of the Commission's proposal for a draft Artificial Intelligence Act (EU AI Act), which contains important restrictions on AI systems used in or in connection with the EU. The use of AI with specific characteristics may adversely affect a number of fundamental rights enshrined in the Charter of Fundamental Rights of the European Union. Therefore, the proposal aims to ensure a high level of protection of these fundamental rights. Closely related to this is one of the fundamental rights most at risk: the right to equal opportunities and the prohibition of discrimination. The focus of this study is the regulation of data and datasets used to train AI applications.

Keywords: Artifical Intelligence, Draft Artificial Intelligence Act, discrimination in artificial intelligence, data protection and artificial intelligence.

Introduction

The application of artificial intelligence (AI) is expanding into more and more areas of life (e.g. it can improve healthcare, help law enforcement authorities fight crime more effectively, make transport safer, or even help detect fraud and cyber security threats, etc.). It is therefore undoubtedly one of the biggest challenges of our time, both from an economic and regulatory perspective. Not least because the European Commission has published a White Paper on Artificial Intelligence in 2020, which will form the basis for specific regulation of AI developments and applications at EU level. It sets out that AI can have a significant impact on our society and that it is necessary to build trust and confidence in it, and that it is crucial that the AI sector is based on fundamental rights and values such as human dignity and privacy. Human-centred AI presupposes technology that people trust because it is in line with the values that underpin human societies. Ethical principles play a crucial role in establishing trust, assessing risks and managing regulation. In the overall design of AI regulation, four main ethical directions

should be highlighted: respect for human autonomy: do not control/manipulate people, do not compromise democratic processes; prevention of harm: including resistance to unintended external influences that may result in harm; fairness: the development, deployment and use of AI systems should be equitable; and explainability: means transparency of operation (trusted AI systems can be traced and their decisions explained, in particular users should be informed that they have been exposed to an AI system and also how the AI system works, what its capabilities are, how and with what reliability it uses the datasets provided to it).

The most important step forward in the regulation of AI is the publication in April 2021 of the Commission's proposal for a draft Artificial Intelligence Act (hereinafter EU AI Act) , which contains important restrictions on AI systems used in or in connection with the EU. The use of AI with specific characteristics (e.g. opacity due to the black box effect, complexity, dependence on data, autonomous behaviour) may adversely affect a number of fundamental rights enshrined in the Charter of Fundamental Rights of the European Union (hereinafter the Charter). Therefore, the proposal aims to ensure a high level of protection of these fundamental rights and to address the different sources of risk through a clearly defined risk-based approach. However, the White Paper, the accompanying Commission report on the responsibility and safety of AI, and the draft Regulation mention several times an area at the intersection of law and AI that has hardly been analysed from a legal perspective and which is the focus of this study: the regulation of data and datasets used to train AI applications.

Closely related to this is one of the fundamental rights most at risk: the right to equal opportunities and the prohibition of discrimination. The main cause of this is the incompleteness or flaw in the data set used by the AI sytem or used in the training of the AI, or the inherent bias in the system. The bias in algorithmic decision-making that can be caused by the aforementioned problems in the dataset can lead to infringement without any intentionality or human awareness behind it. AI can also produce discriminatory results in decision-making if the system learns from discriminatory training data. Distorted training data can have the following discriminative effects: the AI can be trained on biased data; problems can arise if the AI system learns from a discriminative sample; in both cases, the AI system will reproduce this bias. Increasingly, experts are exploring ways to detect and improve algorithms that may be potentially discriminatory against individuals or groups based on specific characteristics, such as gender or ethnicity. This occurs when the outcome for a particular group is systematically different from other groups, and therefore one group is consistently treated differently from others. This can occur when the data used to teach the algorithm contains information on proprietary characteristics (e.g. gender, ethnicity, religion, etc.). Furthermore, the data sometimes contain so-called „surrogate information". This could be, for example, the postcode, which may indirectly refer to ethnic origin in segregated urban areas, or more directly to the country of birth of the person. Unequal outcomes and differential treatment, particularly in relation to proxy information, should be assessed to determine whether they constitute discrimination. Discrimination may be based not only on differences in outcomes between groups, but also when the data selected for use are not neutral. This means that if the data used to build the algorithm reflects a bias, for example against one group, then the algorithm will replicate the human bias in the selection process and learn the data, i.e. discriminate against that group. The data may reflect bias for several reasons, including decisions made in the selection, collection and preparation of the data. For example, an automated image description was trained based on thousands of images described by humans. However, people do not describe images in a neutral way. Notably, an infant white baby was described as a „baby", but a black- baby was described as a „black

baby". This is biased data because it attributes additional characteristics to only one group, whereas objectively both cases should be described including skin colour, or neither. If such information is included in the training data and used to develop algorithms, the results will not be neutral. The data may be poorly selected, incomplete, incorrect or out of date. Poorly selected data may include ‚non-representative data' that do not allow generalisations to other groups. For example, if an algorithm is created based on data for a particular group of job applicants, then predictions for other groups may not be correct. In addition, an algorithm can only be as good as the data it works with, which means that the data model that makes decisions based on the analysis of the algorithm may be biased and discriminatory. In this case too, the principle of „garbage in, garbage out", as used in statistics, applies, meaning that poor quality input data will itself produce poor quality results (predictions). Therefore, algorithms can (still) disadvantage historically disadvantaged groups if they are based on negative and unsubstantiated assumptions. In this sense, data quality control and proper documentation of data and metadata are essential for high quality data analysis and the use of algorithms for decision making.

At first glance, algorithms sort, categorise and organise information in a way that eliminates human biases and prejudices. They should therefore be able to ensure the expected equal treatment by applying the same criteria and weighting, regardless of, for example, the origin of the person. In reality, however, there is no technological wizardry or mathematical neutrality: algorithms are designed by humans using data that reflect human practice. Bias and prejudice can creep into any stage of algorithm system development.

Discrimination in the criminal justice system

A notorious example of an AI system with a discriminatory effect is the system known as Correctional Offender Management Profiling for Alternative Sanctions, or COMPAS for short. COMPAS is used in the criminal justice system in some parts of the United States to predict the likelihood that offenders will re-offend. The basis for the use of this system is that COMPAS can assist the court's work by providing concrete recommendations for decisions. COMPAS can indicate three indicators for the person concerned: the risk of pre-trial release; the recidivism coefficient and the violent recidivism coefficient. Although the COMPAS may not explicitly include a racial factor, it could arguably be programmed to correlate strongly with the ethnic background of the defendant, and thus raise concerns about its use, particularly in relation to due process.

However, a study conducted in 2016 highlighted that the COMPAS system's risk classification reflects a bias against black people. While it correctly predicts recidivism in 61% of cases, it is almost twice as likely to result in a higher risk classification for black than for white. In fact, in their case, the system makes the opposite mistake by being more likely to classify them as lower risk. Furthermore, black-skinned pregnant women are twice as likely as white-skinned pregnant women to be wrongly classified as higher risk for violent recidivism. And white-skinned violent recidivists were 63 percent more likely to be misclassified as low risk.

As promising as these systems are, the inherent bias and discrimination in their data sources, the „black box" problem inherent in the algorithms, is present. Hence, misinterpretations and inferences from data analysis have quickly triggered huge debates among policy makers, practitioners and academics. The consequences of this were illustrated in a recent case, State v. Loomis, in which the Wisconsin Supreme Court upheld a lower court ruling based on the COMPAS risk

assessment system and dismissed the defendant's appeal alleging a violation of his right to due process. In addition, it is worth noting that the law enforcement also use AI systems for predictive policing, which involves automated predictions of who will commit a crime, when and where. Similarly, predictive policing systems can replicate or even amplify existing discrimination.

Discrimination and online advertising

Algorithmic decision-making can also have discriminatory effects in the private sector. A good example of this is the increasingly common case of automated decision-making in recruitment. For example, the hiring of a new employee can be ‚outsourced' to an analytics software that imports and transforms CVs and automatically extracts, stores, analyses, sorts and reviews the information submitted, possibly using other data sources such as the applicant's social media accounts. One prominent example is Amazon's CV filtering software, which was trained on distorted historical data, resulting in a bias towards male candidates, as in the past Amazon has more often hired men as software engineers than women, and the algorithm was trained on this data. Amazon is also reported to have stopped using the AI system to screen job applicants because it was discovered that its new system was not assessing applicants for software engineering and other technical jobs in a gender-neutral way. It is therefore recommended that sources of human bias such as gender, race, ethnicity, religion, sexual orientation, age and information that could indicate membership of a protected group be removed from the system and dataset.

The European Union's draft Artificial Intelligence Regulation

The draft EU AI Act aims at a minimum of horizontal regulation, using a risk-based approach, classifying AI applications into risk classes. The draft distinguishes a fully prohibited category (Title II), which includes the prohibition of facial recognition programmes (with exceptions) in public places; subliminal manipulation; mass surveillance or the unlawfulness of a social point system (similar to the one used in China). In addition, it defines high-risk AI applications (Title III), for which it establishes binding rules, and other applications that are less risky (Title IV) but still deserve some attention (it addresses the risks associated with these applications by supporting transparency provisions), and finally AI applications that do not fall into either category, which it leaves to codes of conduct, i.e. self-regulation.

Of most interest to us in this topic is the regulation of high-risk AI. An AI system is considered high-risk if it is either a safety component of an already tightly regulated group of products (listed in Annex II, from toys to craft to medical devices), or because it is used in an area where human rights are particularly affected. The latter list includes two dozen specific applications in eight areas, such as AIs for biometric identification of natural persons, AIs for the control of critical infrastructures (transport, gas, water, electricity), and some other AIs (such as recruitment, university admissions, credit assessment and advice to judges).

Indeed, the draft regulation states that AI systems used in the context of employment, management of workers and access to self-employment, in particular recruitment and selection of persons, decisions on promotion and dismissal, and the allocation of tasks to persons with a contractual employment relationship, as well as the monitoring or evaluation of such persons, should be considered as high risk, as they may have a significant impact on the future career prospects and livelihood of these persons.

Actions by law enforcement authorities involving the use of certain AI systems are characterised by a significant imbalance of power and can lead to the surveillance, arrest or deprivation of liberty of a natural person, as well as other adverse effects on the fundamental rights guaranteed by the Charter. In particular, if an AI system is not trained with good quality data, does not meet adequate standards of accuracy or stability, or is not properly designed and tested before being placed on the market or otherwise put into service, it may select people in a discriminatory or otherwise unfair or unjust manner. It may also hinder the enforcement of important fundamental procedural rights, such as the right to an effective remedy and to a fair trial, as well as the rights of the defence and the presumption of innocence, if such AI systems are not sufficiently transparent, explained and documented.

The AI systems used in migration management, asylum and border management affect people who are often in a particularly vulnerable situation and whose lives are affected by the outcome of actions taken by the competent authorities. The accuracy, non-discriminatory nature and transparency of the AI systems used in this context are therefore of particular importance in ensuring respect for the fundamental rights of the persons concerned, namely their rights to free movement, non-discrimination, privacy and protection of personal data, international protection and due process.

Some AI systems designed to administer justice and manage the democratic process should be considered high risk, given their significant impact on democracy, the rule of law, individual freedoms and the right to an effective remedy and to a fair trial. In particular, in order to address the risk of possible distortions, errors and opacity, AI systems which aim to assist judicial authorities in researching and interpreting factual and legal elements and in applying the law to specific facts should be considered as high risk. However, this classification should not cover AI systems intended for purely ancillary administrative activities which do not affect the actual administration of justice in individual cases, such as the anonymisation or pseudonymisation of court decisions, documents or data, staff communications, administrative tasks or the allocation of resources.

The requirements for high-risk AI in the EU AI Act (Chapter 2) are: risk assessment systems must always be established, implemented, documented and maintained (Article 9); they must be operated in conjunction with appropriate data governance systems; and the data used for teaching, validating and testing must be „clean" (Article 10). High-risk AIs should be accompanied by detailed documentation and event logging systems (Articles 11-12). Systems of this type should operate transparently and always retain human oversight and intervention (Articles 13-14). They should also meet the requirements of accuracy, robustness and cyber security (Article 15).

It is worth taking a closer look at the provisions in Article 10 on instructive data, which define a governance regime for instructive data that includes comprehensive requirements for the entire lifecycle of such data sets when used to teach high-risk AI applications. The draft regulation goes on to define three important sets of specific quality criteria for high-risk systems. Firstly, Article 10(3) states that the training data should be „relevant, representative, error-free and complete", which reflects, but does not elaborate on, several data quality requirements found in the IT literature discussed above. Second, the training data must have appropriate statistical properties, including with respect to the individuals or groups of individuals to whom or to which the high-risk AI system is intended to be applied. Although the groups constituted by the protected characteristics are not explicitly mentioned in this section, the criterion does seem to include the issue of balance between members of protected groups in

the datasets. However, statistical adequacy must be met for all sufficiently distinct groups, whether defined by protected characteristics or not, which makes the provision equally broad (think for example of different socio-economic groups) and vague. It prescribes appropriate statistical characteristics for each group, but without providing any further guidance on what is meant by appropriateness in this context. Thirdly, the criterion of representativeness is further clarified in Article 10(4), which states that the training data should take into account and reflect, to the extent necessary for the purpose, characteristics or elements that are related to the specific geographical, behavioural or functional context in which the high-risk AI system is intended to be used. This provision therefore forces developers to consider the specific context of the intended use of the system. The draft Regulation presumes in Article 42(1) that the context representativeness criterion is met if the training data are derived from the intended geographical, behavioural and functional environment.

The EU AI Act provides for an important exception to the prohibition on processing sensitive data in Article 9(1) of the General Data Protection Regulation (hereinafter GDPR). Article 10(5) rightly resolves the tension between data protection and non-discrimination areas. To the extent strictly necessary to ensure the monitoring, detection and correction of bias in relation to high-risk AI systems, the providers of such systems may handle special categories of personal data, subject to appropriate safeguards for the fundamental rights and freedoms of natural persons, including technical limitations on the further use and application of state-of-the-art security and privacy measures, including pseudonymisation or, where anonymisation significantly affects the intended purpose, encryption.

The EU AI Act provides for strict sanctions (Article 71) for failure to comply with the requirements set out in Article 10 and for failure to comply with the prohibition of AI practices referred to in Article 5, with administrative fines of up to EUR 30 000 000 or, in the case of undertakings, up to 6% of the total annual worldwide turnover of the preceding financial year, whichever is the higher. Where the AI system does not comply with the requirements or obligations under this Regulation, other than those laid down in Articles 5 and 10, it may be subject to an administrative fine of up to EUR 20 000 000 or, in the case of undertakings, up to 4 % of its total annual worldwide turnover in the preceding financial year, whichever is the higher.

In addition, the obligations on prior testing, risk management and human oversight in the draft Regulation will also help to ensure respect for other fundamental rights by minimising the risk of erroneous or biased decisions based on AI in critical areas such as education and training, employment, essential services, law enforcement and justice. If violations of fundamental rights continue to occur, ensuring transparency and traceability of AI systems and rigorous ex-post monitoring will allow for effective redress for the individuals concerned. Enhanced transparency obligations are limited to the minimum information necessary for individuals to exercise their right to effective redress and to the transparency necessary for supervisory and enforcement authorities, in accordance with their mandate, thus not disproportionately affecting the right to the protection of intellectual property [Article 17(2)]. Where public authorities and notified bodies need to have access to confidential information or source code for the purpose of verifying compliance with the relevant obligations, they are bound by a duty of confidentiality.

Summary

The legal standards on algorithmic discrimination are clear. Our societies do not and should not accept discrimination based on protected characteristics such as ethnic origin or gender. This raises the question of how to improve the enforcement of non-discrimination norms in the field of algorithmic decision making? As already mentioned, one of the main problems of AI systems is their black box nature. This opacity can be seen as a problem in itself, but opacity is also a barrier to detecting discrimination. However, appropriate legal regulation can help to make algorithmic decision-making more transparent. For example, in the European Union, the draft AI Act already sets minimum requirements for high-risk AI to be developed in a way that allows for verification and explanation, and includes specific provisions on the datasets used by the system. However, there are still unanswered questions about regulation. Furthermore, it is an important step to ensure that, in the case of high-risk AI, the competent authorities have access to the underlying code (software) and datasets of algorithmic systems in case of a serious breach, as the examination of the code may provide information on the functioning of the system. However, it can be agreed that code reviews can be most useful when there is a clearly defined question about how an algorithm operates in the controlled space and there are specific standards against which the behaviour or performance of the system can be measured.

BIBLIOGRAPHY

Angwin, Julia et al.: Machine Bias: There's Software Used Across the Country to Predict Future Criminals. And It's Biased Against Blacks, ProPublica, 2016

Borgesius, Frederik Zuiderveen: Discrimination, artificial intelligence, and algorithmic-decision making. Council of Europe, 2018.

Dastin, Jeffrey: Amazon scraps secret AI recruiting tool that showed bias against women. Reuters, https://reut.rs/3vvkHsh

Dieterich, William - Mendoza, Christina - Brennan, Tim: COMPAS Risk Scales: Demonstrating Accuracy Equity and Predictive Parity, Northpointe, 2016

Dumbrava, Costica: Artificial intelligence at EU borders. European Parliamentary Research Service, Brussels, 2021.

Hacker, Philipp: Teaching fairness to artificial intelligence: Existing and novel strategies against algorithmic discrimination under EU law. Common Market Law Review 2017.

High Level Expert Group on Artificial Intelligence: ethics guidelines for trustworhy AI. Brussels, 2019.

Köchling, Alina - Wehner, Marius Claus: Discriminated by an algorithm: a systematic review of discrimination and fairness by algorithmic decisionmaking in the context of HR recruitment and HR development. Business Research 2020/13.

Körtvélyesi, Zsolt: Coded inequalities? Coding discrimination in the age of algorithms. JTI blog, https://bit.ly/3K0Zfk5

Karsai, Krisztina: The European draft for the regulation of artificial intelligence, or the signs of the rise of algorithms in (criminal) justice. Forum: Acta Juridica Et Politica 2021/3. 189-196.

Kullmann, Miriam: Discriminating job applicants through algorithmic decision-making. https://bit.ly/3vr6NaF, European Union Agency for Fundamental Rights: Data quality and artificial intelligence - mitigating bias and error to protect fundamental rights. Vienna, 2019

Larson, Jeff et al.: How We Analyzed the COMPAS Recidivism Algorithm, ProPublica, 2016.
Berk, Richard: Criminal justice forecasts of risk: a machine learning approach. Springer, 2012
Rieke, Aaron - Bogen, Miranda - Robinson, David G.: Public scrutiny of automated decisions: early lessons and emerging methods, Upturn and Omidyar Network. 2018.