Bialystok Legal Studies
Białostockie Studia Prawnicze
2021 vol. 26 nr 3

sciendo

DOI: 10.15290/bsp.2021.26.03.02

Received: 19.12.2020
Accepted: 25.03.2021

Rafał Rejmaniak
University of Białystok, Poland
r.rejmaniak@uwb.edu.pl
ORCID ID: https://orcid.org/0000–0003-1908–5844

Bias in Artificial Intelligence Systems

Abstract: Artificial intelligence systems are currently deployed in many areas of human activity. Such systems are increasingly assigned tasks that involve taking decisions about people or predicting future behaviours. These decisions are commonly regarded as fairer and more objective than those taken by humans, as AI systems are thought to be resistant to such influences as emotions or subjective beliefs. In reality, using such a system does not guarantee either objectivity or fairness. This article describes the phenomenon of bias in AI systems and the role of humans in creating it. The analysis shows that AI systems, even if operating correctly from a technical standpoint, are not guaranteed to take decisions that are more objective than those of a human, but those systems can still be used to reduce social inequalities.
Keywords: AI discrimination, AI fairness, algorithmic bias, artificial intelligence

Introduction

Technological solutions based on artificial intelligence (AI) are being used more and more widely in various spheres of human activity. AI systems are deployed in both the private and the public sectors. The widespread use of such solutions is motivated by the potential benefits, which are hard to overestimate – from making production processes more efficient or analysing large quantities of data at speeds far exceeding human capabilities to forecasting future events. One of the frequently cited properties of AI systems, said to give them an advantage over humans in performing certain types of tasks, is the greater objectivity of their 'decisions' and

their insusceptibility to the influence of subjective feelings and emotions.[1] There is no doubt that from a technical point of view, in the case of tasks requiring precision, repeatability and the processing of large quantities of data in a short time, AI systems will generally outperform humans. This does not mean, however, that such systems are guaranteed to carry out the tasks entrusted to them in a way that can be regarded as appropriate from a social perspective.

From the growing number of studies concerning this problem, it is becoming clear that even AI systems can be subject to bias, and in the longer term this may lead to discrimination against individuals or even entire social groups.[2] The problem is acknowledged by various bodies seeking to establish legal and ethical frameworks for the development of artificial intelligence, both nationally and internationally.[3] The aim of this article is to describe the phenomenon of bias in AI systems and to show that AI systems, even when biased, can be useful for reducing social inequalities. For the purposes of this work, the term 'bias' is taken to mean simply a deviation from the norm,[4] understood as a commonly accepted and agreed standard, making it a broader concept than 'discrimination'. It is worth noting that these very standards may show a discriminating nature on their own, having roots in beliefs and prejudices found in society or being politically motivated.

1. The Notion of Artificial Intelligence

'Artificial intelligence' is a notion that does not yet have a single generally accepted definition. For the purposes of this work, artificial intelligence will be understood as proposed by the High-Level Expert Group on Artificial Intelligence – as 'software (and possibly also hardware) systems designed by humans that, given a complex goal, act in the physical or digital dimension by perceiving their environment through data acquisition, interpreting the collected structured or unstructured data, reasoning on the knowledge or processing the information, derived from this data and deciding the best action(s) to take to achieve the given goal. AI systems can either use symbolic rules or learn a numeric model, and they can also adapt their behaviour by analysing

1 P. Hacker, Teaching Fairness to Artificial Intelligence: Existing and Novel Strategies against Algorithmic Discrimination under EU Law, "Common Market Law Review" 2018, vol. 55, pp. 1143–1144.

2 See R. Rodrigues, Legal and Human Rights Issues of AI: Gaps, Challenges and Vulnerabilities, "Journal of Responsible Technology" 2020, vol. 4, p. 3.

3 See J. Fjeld, N. Achten, H. Hilligoss, A. Nagy and M. Srikumar, Principled Artificial Intelligence. Mapping Consensus in Ethical and Rights-Based Approaches to Principles for AI, Cambridge 2020, pp. 47–52.

4 D. Danks and A.J. London, Algorithmic Bias in Autonomous Systems, "Proceedings of the 26th International Joint Conference on Artificial Intelligence (IJCAI 2017)", p. 2, https://www.cmu.edu/dietrich/philosophy/docs/ london/IJCAI17-AlgorithmicBias-Distrib.pdf (accessed 06.02.2021).

how the environment is affected by their previous actions'.[5] The authors of this definition specify the meaning of a 'decision' as 'any act of selecting the action to take', and this 'does not necessarily mean that AI systems are completely autonomous. A decision can also be the selection of a recommendation to be provided to a human being, who will be the final decision maker.'[6]

In contrast to ordinary algorithms, which involve the sequential completion of predefined steps, a fundamental feature of AI is the ability to 'learn'. In this process, known as machine learning, external empirical data are used to create and update rules for the improved handling of similar data in the future, and to express these rules in a comprehensible, symbolic form.[7] It is not the aim of this article to present the techniques of machine learning,[8] but it is necessary to make two remarks to enable understanding of the problems of AI bias that are to be discussed below.

First, machine learning may take the form of supervised, unsupervised or reinforcement learning.[9] In the first case, the data used to train the AI system are labelled. The system analyses the input data and determines relationships between them. If it makes an incorrect classification, it is informed of that fact and will modify its hypotheses.[10] Unsupervised learning uses a pool of unlabelled training data; the task of the AI system is to find, independently, non-trivial relationships in the data. In such cases, as a rule, the trainers do not have knowledge of the final outcome of the learning process. In reinforcement learning, on the other hand, for every correct classification the system receives a 'reward' (for example, its goal is to earn as many points as possible, and for each correct identification of data it is awarded points, while for an incorrect identification it has points taken away). Some AI systems are brought into use after their training is complete, whereas others continue to learn for the whole time that they are in use. An example of the latter type is Google Translate.[11]

Second, while various techniques are used for training AI systems, one of the most popular currently is deep learning, based on multiple layers of artificial neural networks. An artificial neural network is a simplified mathematical model

5 High-Level Expert Group on Artificial Intelligence (appointed by the European Commission in June 2018), A Definition of Artificial Intelligence: Main Capabilities and Scientific Disciplines, Brussels 2019, p. 6.

6 *Ibidem*, p. 3.

7 D. Michie, Methodologies from Machine Learning in Data Analysis and Software, "The Computer Journal" 1991, vol. 34, no. 6, p. 562.

8 See e.g. M. Flasiński, Wstęp do sztucznej inteligencji, Warsaw 2020; L. Rutkowski, Metody i techniki sztucznej inteligencji, Warsaw 2012.

9 M.A. Boden, Sztuczna inteligencja. Jej natura i przyszłość, trans. T. Sieczkowski, *Łódź* 2020, pp. 59–60.

10 *Ibidem*, p. 60.

11 G. Massey and M. Ehrensberger-Dow, Machine Learning: Implications for Translator Education, "Lebende Sprachen" 2017, vol. 62, no. 2, p. 301.

of the structure of the brain.[12] The artificial neurons that form such a network receive input signals, each signal being multiplied by a corresponding numerical value called a weight. If an activation threshold is exceeded, the neuron transmits a signal which becomes an input signal for neurons in the next layer.[13] In this case, learning consists of determining appropriate weights for the various input signals. A significant issue concerning deep learning is the presence of hidden layers between the input and output layers – the networks themselves lack the ability to explain the decision-making process.[14] While it is still possible to determine what weights have been assigned to particular input signals and to repeat the training in case of an unsatisfactory result,[15] it is no longer possible to establish *why* the system assigned weights as it did.

2. Types of Bias in AI Systems

The phenomenon of AI bias is a complex one, and may be caused by a variety of factors arising at different stages of the training and operation of such a system. The first group of factors relates to the data used as a basis for training or for the making of decisions or predictions. A second group is related to the construction of the system itself. The third group consists of factors affecting the user who interprets the system's decisions or predictions.

An AI system is trained by supplying it with data, which may or may not be labelled. The quality of the training data will determine how the system subsequently functions. Even at this stage, human decisions can introduce bias into the system. It is humans who select the data to be included in the training set, and so if these data are chosen in a biased manner, then the system's subsequent decisions will be similarly biased.[16] For example, if the training set for a face recognition system consists mostly of photographs of white men, then a system trained on that set will be capable of recognizing white male faces much more effectively than those of black women.[17] Lower accuracy in facial recognition does not necessarily mean that it bears a nature of bias. Only if such a system is utilized in a particular context may its use be related to partiality, especially when its operation could influence the situation of the individual who is the subject of a decision.

12 M. Flasiński, Wstęp, *op. cit.*, p. 161.

13 A. Kasperska, Problemy zastosowania sztucznych sieci neuronalnych w praktyce prawniczej, „Przegląd Prawa Publicznego" 2017, no. 11, p. 25.

14 *Ibidem*, p. 27.

15 M. Flasiński, Wstęp, *op. cit.*, p. 163.

16 M. Coeckelbergh, AI Ethics, Cambridge/London 2020, pp. 130–131; W. Barfield and U. Pagallo, Advanced Introduction to Law and Artificial Intelligence, Cheltenham/Northampton 2020, p. 25.

17 M.S. Cataleta and A. Cataleta, Artificial Intelligence and Human Rights: An Unequal Struggle, "CIFILE Journal of International Law" 2020, vol. 1, no. 2, p. 46.

The problem of data bias can take yet another form, being rooted more deeply in the inequalities existing in society as a whole. An example here is COMPAS, a criminological risk assessment system based on AI algorithms that was tested in the United States. The system achieved 70% accuracy,[18] although its code remained a trade secret.[19] An investigation by journalists from the ProPublica website, covering persons charged with offences in Florida in 2013–14, showed – using reverse engineering – that COMPAS made false positive predictions twice as often in relation to black people and false negatives twice as often in relation to whites,[20] even though its creators claimed that the system did not consider race as a relevant feature.[21] It was further shown that HART, a similar prediction system used in the United Kingdom, also took decisions of a tendentious and discriminatory nature.[22] According to Hannah Fry, this type of bias is inevitable from a statistical point of view, since in the case of certain types of offence, black citizens in the United States are arrested much more often than whites, even though the percentages of offences committed are in fact similar in both populations.[23] Here the bias in the AI system results from the prejudices existing in society itself, which are reflected in the statistical data used to train the system. Although the act of comparing the criminality of black and white people may be controversial, having in mind its controversial political background, it is established that the sheer mechanism for AI functioning does not raise any doubts. For example, if 80% of a group's individuals can be characterized by a certain feature, it is most probable that the AI system is going to attribute a high value to it, no matter what type of a feature it is.

Controversies of a more general nature can also be attributed to the use of systems such as COMPAS. Criminological prognosis, not to mention the adjudication of

18 T. Brennan, W. Dieterich and B. Ehret, Evaluating the Predictive Validity of the COMPAS Risk and Needs Assessment System, "Criminal Justice and Behavior" 2009, vol. 36, no. 1, p. 31.

19 H. Fry, Hello World. Jak być człowiekiem w epoce maszyn, trans. S. Musielak, Krakow 2019, p. 87.

20 J. Angwin, J. Larson, S. Mattu and L. Kirchner, Machine Bias, ProPublica 2016, https://www.propublica.org/ article/machine-bias-risk-assessments-in-criminal-sentencing (accessed 19.03.2021). ProPublica journalists conducted an analysis of 10,000 accused individuals from Broward County, Florida. It has been checked whether those individuals behaved as predicted by the COMPAS system's prognosis for two consecutive years. What is more, the analysis showed that in the case of a similarity between variables such as previously committed crimes, age and sex, accused black defendants have been 45% more likely to get misclassified as higher risk than white defendants. Detailed methodology has been presented by the authors: J. Larson, S. Mattu, L. Kirchner and J. Angwin, How We Analyzed the COMPAS Recidivism Algorithm, ProPublica 2016, https://www.propublica. org/article/how-we-analyzed-the-compas-recidivism-algorithm (accessed 19.03.2021).

21 A. Yapo and J. Weiss, Ethical Implications of Bias in Machine Learning, "Proceedings of the Annual Hawaii International Conference on System Sciences" 2018, p. 5368.

22 M. Dymitruk, Sztuczna inteligencja w wymiarze sprawiedliwości?, (in:) L. Lai and M. Świerczyński (eds.), Prawo sztucznej inteligencji, Warsaw 2020, p. 283.

23 H. Fry, Hello World, *op. cit.*, pp. 92–94.

guilt and penalty, should take into consideration the circumstances of a specific case. Taking into account only the statistical models would distort the fundamental rule of criminal law – the individualization of criminal liability. Because of the above, such tools could only serve an auxiliary role for the adjudication process.

This type of bias in AI systems may lead in the end to a damaging feedback loop that petrifies or exacerbates existing inequalities. Such a system makes decisions based on previously gathered data. Those decisions are implemented, generating further strings of data that enhance the system. Thus the functioning of the system itself is generating data which is used to update its predictive model. This phenomenon is explained by C. O'Neil with an example of the PredPol system. She points out that in the case of predictive systems, a situation may be reached where the system identifies certain geographical areas as being more likely than others to experience crime. Police officers are sent to those areas, and because they happen to be there, will tend to arrest persons committing relatively minor offences. The same types of offences are not recorded in other areas, where (because of the system's predictions) officers were not sent, and therefore are not included in the police statistics. The same statistical data are fed into the AI system, which uses them to update its predictive model,[24] treating places where crimes have been recorded as potential crime areas and those where crimes have not been recorded as being less in need of the police's attention. When the system operates in this way, it produces self-fulfilling prophecies.

A predictive system on its own does not create crime. What it does is point out areas where officers' attention should be focused. It is especially important from the perspective of the optimal use of human resources, which are always limited. The problem with predictive systems does not lie in the fact that officers discover minor offences (all deviations from the criminal law norm should meet adequate state reaction) but with having certain geographical areas deemed by the system to be especially at risk of crime. It may lead to a situation where such areas could be overrated by the system due to the system's data generation, with other regions with a higher crime rate somehow neglected.

Moreover, the use of such systems may also have negative social consequences. Disproportionate police surveillance carried out in poor districts may lead to inhabitants' loss of trust towards the officers and also towards each other, and it is worth noting that trust is crucial in such places.[25] Such operation of the system may result in social exclusion based on domicile and stereotypes connected with inhabitants of particular districts dubbed as high-crime areas. Such prejudices could impact the life of an individual in many ways, e.g. during a job search or the possibility of receiving a loan.

24 C. O'Neil, Broń matematycznej zagłady. Jak algorytmy zwiększają nierówności i zagrażają demokracji, trans. M.Z. Zieliński, Warsaw 2017, pp. 128–129.

25 M. Coeckelbergh, AI Ethics, *op. cit.*, p. 128.

The above does not mean that the use of predictive systems to fight crime should be entirely discontinued. Crime forecasting is not a new phenomenon, and criminology experts have made attempts, with varying degree of success, to predict the future shape of crime rates and types. The use of a predictive system could help identify areas particularly vulnerable to crime. It may be especially important for crime categories which are related to area and infrastructure, such as burglary, the consumption of alcohol in public places, property damage, etc. Directing officers to these places may allow the creation of hot spots and in the future the implementation of relevant infrastructural solutions (e.g. city lighting or CCTV monitoring). In a case like this, it would make a predictive system one of the integrated components of a larger crime prevention system.

Another solution would be to limit the use of the AI system to only forecasting minor offences, leaving out more serious ones, which usually occur less often. It may seem, though, that no matter how the system is implemented, it would be necessary to periodically verify its accuracy and further update predictive model data with information gathered from other areas, e.g. obtained via periodic analogous intensification of patrolling in random sectors of the city.

Even if a society has overcome the problem of discrimination against a particular group, this does not mean that an AI system will be free of data bias. Such systems are trained using large quantities of data (big data), some of which are historical. Thus, if there is bias in the historical data, a system trained on those data may still end up biased. This phenomenon is known as *historic bias*.[26] Theoretically it is possible to train a system on current data, omitting the defective historical data, but in practice this approach may leave too small a training set.[27] Decisions of the AI system may still be biased. This could be attributed to the limited validity of the model, a result of it being based on a small data pool.

In the case of systems whose learning ends before they are brought into use, the data used for training are to some extent subject to control by the people responsible for the training process. However, other systems continue to learn while they are in use. This enables the system to acquire new data and to modify its behaviour continuously so as to perform its tasks in an optimum manner. The data obtained by such systems may also prove defective. There are known cases where users deliberately fed discriminatory data into the system. One of the best-known examples is the Tay bot, launched by Microsoft in 2016, which was supposed to simulate a lively, happy teenage girl on Twitter. The bot was designed to create its own tweets, learning from interactions with other users. After a few hours of being deliberately fed controversial

26 F. Lattimore, S. O'Callaghan, Z. Paleologos, A. Reid, E. Santow, H. Sargeant and A. Thomsen, Using Artificial Intelligence to Make Decisions: Addressing the Problem of Algorithmic Bias. Technical Paper, Australian Human Rights Commission, Sydney 2020, pp. 33–34.

27 *Ibidem*, p. 39.

content, the bot began to publish tweets of a racist, sexist and antisemitic nature, and Microsoft therefore decided to shut it down.[28]

Sometimes the bias in an AI system may be a consequence of the way the system itself is constructed. According to David Danks and Alex John London, this situation may be reached when data are processed using a statistically biased estimator.[29] In some cases the use of such estimators may be justified: for instance, to increase the accuracy and reliability of the results when a system is trained on a small amount of data.[30]

There are also solutions that deliberately produce a given type of bias in an AI system (statistical bias) in order to counteract other biases.[31] This means that the system's decisions are intended to reflect reality not as it currently is, but as it should be.[32] In such cases it is a human who decides what vision of the world the AI system is to promote. Should it reproduce the world as it is with maximum accuracy based on collected data, or should it be a tool to correct the world's imperfections by taking decisions that have been somehow 'enhanced'?

A system may prove biased in yet another way, when it finds correlations between certain features of the input data that give a simplified picture of reality. System designers and trainers have to decide which data are significant for the system's purposes and which are to be ignored.[33] Moreover, in building a predictive model, an AI system may assign too great a weight (from an anti-discrimination perspective, say) to features – such as race or sex – that should not be decisive or should not be taken into account at all in the making of particular decisions, for instance in making criminological predictions or hiring employees. As a rule, simply removing a given feature from the database used as the system's training set will not solve this problem. The AI system may take account of the feature indirectly,[34] since in individual cases it will often have an influence on other features that are correlated with it (redundant encodings).[35] For example, from a database containing data obtained from individuals' Facebook profiles, but not including information on their

28 See G. Neff and P. Nagy, Talking to Bots: Symbiotic Agency and the Case of Tay, "International Journal of Communication" 2016, no. 10, pp. 4920–4922.

29 D. Danks and A.J. London, Algorithmic Bias, *op. cit.*, p. 3.

30 S. German, E. Bienstock and R. Doursat, Neural Networks and Bias/Variance Dilemma, "Neural Computation" 1992, vol. 4, no. 1, p. 15.

31 D. Danks and A.J. London, Algorithmic Bias, *op. cit.*, p. 3.

32 F. Lattimore et al., Using Artificial Intelligence, *op. cit.*, p. 29.

33 S. Barocas and A.D. Selbst, Big Data's Disparate Impact, "California Law Review" 2016, vol. 104, no. 2, p. 688.

34 D. Roselli, J. Matthews and N. Talagala, Managing Bias in AI, "Companion Proceedings of the 2019 World Wide Web Conference, San Francisco, CA, USA", May 2019, pp. 2–3.

35 E. Ntoutsi, P. Fafalios, U. Gadiraju, V. Iosifidis, W. Nejdl, M.-E. Vidal, S. Ruggieri, F. Turini, S. Papadopoulos, E. Krasanakis, I. Kompatsiaris, K. Kinder-Kurlanda, C. Wagner, F. Karimi, M. Fernandez, H. Alani, B. Berendt, T. Kruegel, C. Heinze, K. Broelemann, G. Kasneci,

sexual orientation, it is possible to predict their orientation relatively accurately by analysing the types of people who appear as their friends.[36]

What is more, in seeking correlations between data, an AI system may ascribe significance to incidental features that are of no importance in practice, but are nonetheless present in the dataset given. This mechanism is well illustrated by an experiment conducted by Ribeiro, Singh and Guestrin concerning the training of an AI system for image recognition. The system was supposed to distinguish photographs of wolves from photographs of husky dogs, which indeed it did with a high degree of accuracy. However, deeper analysis showed that the key criterion being used by the system was not any of the animals' features, but the presence or absence of snow in the photograph. If snow appeared, the system decided the picture was of a wolf; if not, it was deemed to show a dog.[37] Although this accidental correlation did in fact hold true for the collection of photographs used, a wolf is not a wolf merely because there is snow around it.[38]

Paradoxically, the experiment shown above can be used as an argument for utilizing artificial intelligence systems in real life. If the system is found to be biased through assigning inappropriate weight to certain features, then this bias can be detected and the system redesigned or simply retrained to meet relevant criteria. When it comes to decisions made by humans, the detection of bias could be much more complicated, for a seemingly objective substantiation may be backed with deep-seated prejudice, emotions or even certain fixed states, such as time of day or even hunger, felt while making a decision.[39] Taking into consideration the above, humans are much less 'fixable' than AI systems.

Humans themselves may be the source of bias in an AI system. As noted above, it is humans who design and train the system, and in doing so take decisions that will ultimately affect how the system operates. These may concern the selection of training data, the identification and labelling of significant features of the data and the construction of the system itself, including the use of deliberately biased estimators to eliminate other types of bias. It may therefore happen that human decisions are the original cause of the types of bias presented above. However, human involvement is not limited to those developing the system.

T. Tiropanis and S. Staab, Bias in Data-Driven Artificial Intelligence Systems – An Introductory Survey, "WIREs Data Mining Knowledge Discovery" 2020, vol. 10, no. 3, p. 4.

36 See C. Jernigan and B.F. Mistree, Gaydar: Facebook Friendships Expose Sexual Orientation, "First Monday" 2009, vol. 14, no. 10; F. Zuiderveen Borgesius, Discrimination, Artificial Intelligence and Algorithmic Decision-Making, Strasbourg 2018, p. 13.

37 See M.T. Ribeiro, S. Singh and C. Guestrin, "Why Should I Trust You?" Explaining the Predictions of Any Classifier, '22nd ACM SIGKDD International Conference 2016, San Francisco', pp. 8–10, https://www.kdd.org/ kdd2016/papers/files/rfp0573-ribeiroA.pdf (accessed 06.02.2021).

38 D. Roselli, J. Matthews and N. Talagala, Managing Bias, *op. cit.*, p. 4.

39 H. Fry, Hello World, *op. cit.*, p. 103.

AI systems that could be placed in the category 'general artificial intelligence', meaning a system capable of performing any task requiring intellect at a human level or higher, do not currently exist. Existing systems represent 'narrow artificial intelligence' and are designed to serve specific purposes. An AI system providing a virtual chatbot has different tasks than a system controlling a driverless vehicle or making criminological predictions. The fact that an AI system properly performs the tasks for which it was designed does not mean that it can operate with similar accuracy and confidence in other domains. Moreover, even when a system is used for its intended purpose, bias may be introduced if the conditions are different from those anticipated by its designers. This phenomenon is known as *transfer context bias*. For example, an AI system used to control a driverless vehicle designed for a right-hand traffic environment will not function correctly in a situation where the traffic is on the left.[40] This type of bias may also be related to cultural differences between the countries in which an AI system is used (cultural bias).[41]

Some AI systems are designed to play an advisory role, helping humans to take the right decision. These systems, after analysing the input data, present recommendations or suggestions that the user can accept or reject; any erroneous decision is the user's responsibility. An example of such cooperation between humans and AI can be found in medical diagnostics.[42] The AI system can collect and process data – for example, in the form of medical publications or the medical history of large numbers of patients – with the goal of proposing a diagnosis. Assessing the accuracy of the diagnosis and deciding whether to administer a particular treatment will be the responsibility of a human. Nevertheless, problems may arise in practice owing to the temptation to treat such a system as infallible – as a kind of 'moral buffer'[43] apparently shielding from responsibility a user who is incapable of processing such large sets of data or who lacks the time or skills to take a proper decision.[44] Overconfidence in the results output by an AI system may also be due to failure to understand the principle on which it works. In building a predictive model, the system only seeks correlations between data, that is, the co-occurrence of particular features and the directions of dependence. It does not attempt to explain the identified relationships in terms of cause and effect.[45] Of course, the fact that particular features co-occur does not mean that one feature is the cause of the other and does not provide any explanation for the relationship.

40 D. Danks and A.J. London, Algorithmic Bias, *op. cit.*, p. 3.

41 M. Coeckelbergh, AI Ethics, *op. cit.*, pp. 128–129.

42 T. Davenport and R. Kalakota, The Potential for Artificial Intelligence in Healthcare, "Future Healthcare Journal" 2019, vol. 6, no. 2, pp. 95–96.

43 M.L. Cummings, Automation and Accountability in Decision Support System Interface Design, "The Journal of Technology Studies" 2006, vol. 32, no. 1, p. 26.

44 F. Zuiderveen Borgesius, Discrimination, *op. cit.*, p. 8.

45 F. Lattimore et al., Using Artificial Intelligence, *op. cit.*, p. 20.

An AI system presents its output data only with a certain degree of likelihood – it does not offer certainty.[46] Users must be aware of this, as they are usually the final decision-makers (unless the decision is taken fully automatically by the system, as with the calculation of credit scores, for example). Most often, then, it depends on a human being whether a decision proposed by a biased AI system has an actual effect on the lives of the people the decision concerns. Placing excessive trust in the objectivity and infallibility of AI systems may lead to unequal treatment of people in similar situations. This may be a result of bias in the system or in the data, but the decisive role is played by the person who interprets the result the system generates. A clear example is the above-mentioned COMPAS system, which was designed to make criminological predictions and to justify resocialization decisions taken with regard to specific individuals. In practice, however, judges in many American states used the system to determine offenders' sentences.[47]

The possibility cannot be excluded either that people might deliberately provoke an AI system to make biased decisions. The data collected for testing such a system may be manipulated, resulting in a distorted picture of reality (for example, by overrepresenting or underrepresenting certain features or groups). A system can also be designed deliberately to discriminate against individuals with certain features or against entire social groups. Moreover, it might serve as a kind of filter for identifying people with specific characteristics in order to subject them to repression. It is not difficult to imagine a situation in which an autocratic or totalitarian government might use an AI system to seek out people with features or views that deviate from those expected (based on their social media data, for example) so as to take repressive measures against such persons. In this case, however, the system itself may be functioning correctly from a technical standpoint, the problem being the use to which it is put.[48]

3. Eliminating Bias from AI Systems

Bias in AI systems is a complex phenomenon and may result from various causes occurring at different stages of the system's life cycle. This causes significant difficulty in laying down general conditions and standards to enable the reduction of bias. Moreover, not every actual bias will be of a discriminatory nature from a human rights perspective. For example, an AI system used to diagnose lung cancer may assign different weights to the same factors depending on whether they occur in men or women. This is a consequence of the fact that there exist objective differences between the sexes in terms of etiology, pathophysiology, histology, disease

46 A. Yapo and J. Weiss, Ethical Implications, *op. cit.*, p. 5366.
47 J. Angwin et al., Machine Bias, *op. cit.*
48 M.S. Cataleta and A. Cataleta, Artificial Intelligence, *op. cit.*, p. 45.

risk factors, effectiveness of therapy and survival.[49] In certain circumstances it is possible to restrict application of the right to equal treatment and the prohibition of discrimination, provided that this is done for a lawful purpose, in an appropriate form and in accordance with the principle of proportionality.[50] It would appear, then, that a proper approach is to seek appropriate solutions limited to particular types of bias or particular areas in which an AI system might be used.[51] This requires interdisciplinary studies, with the involvement of programmers, lawyers, ethicists and experts in the fields in which AI systems are to be deployed.

Biased decisions taken by AI systems, if acted on, may go against such values as the right to equal treatment and the prohibition of discrimination. Hence, action is being taken to construct certain ethical and legal frameworks to ensure respect for the rights of the individual when AI systems are used. At European Union level, the concept of 'trustworthy artificial intelligence' is being developed. The need to avoid discrimination has been expressed in a number of documents, including the White Paper on Artificial Intelligence,[52] the Ethics Guidelines for Trustworthy AI[53] and a European Parliament resolution on a framework of the ethical aspects of artificial intelligence, robotics and related technologies.[54] In that resolution, the Parliament called on the European Commission, among other things, to draw up political solutions with regard to bias in AI algorithms, pointing out that this problem can cause real harm to individuals and society. The elimination of bias might be served by the introduction of rules on data processing that could be used to counteract unequal treatment and discrimination in certain situations and provide a driving force for equal rights and positive social changes. The European Parliament also proposes that national supervisory authorities should inspect the datasets used in AI systems and that investment should continue to be made in research, analysis, innovations, and cross-border and intersectoral knowledge transfer to allow the development of AI technologies completely free of any type of profiling, unequal treatment or discrimination. It further proposes to provide citizens with effective means of appeal that would guarantee unbiased human verification of any claims relating to breaches of their rights resulting from the use of algorithmic systems.

49 E. Ntoutsi et al., Bias, *op. cit.,* p. 8.

50 P. Hacker, Teaching Fairness, *op. cit.*, p. 1164ff.

51 F. Zuiderveen Borgesius, Discrimination, *op. cit.,* p. 39.

52 White Paper on Artificial Intelligence. A European Approach to Excellence and Trust, COM(2020) 65 final, European Commission, Brussels 2020, p. 22.

53 High-Level Expert Group on Artificial Intelligence, Ethics Guidelines for Trustworthy AI, Brussels 2019, pp. 13, 23.

54 European Parliament resolution of 20 October 2020 with recommendations to the Commission on a framework of ethical aspects of artificial intelligence, robotics and related technologies (2020/2012(INL)).

The actions mentioned above are a part of the 'ecosystem of trust' being built within the EU, where apart from the assurance of equal treatment, consideration is also given to such issues as the right to privacy, the autonomy, transparency and explicability of AI systems, and responsibility for inappropriate system operation.

In the literature on bias in AI systems, it is noted that general principles, although important in indicating directions for action, are difficult to put into practice because of their lack of precision.[55] Combating bias, however, is something that can be approached from two directions. First, it is possible to take preventive measures, aimed at preventing the creation of bias, through appropriate data selection and the 'sanitization' of biased data, and also to ensure that designers (as well as trainers and testers) evaluate system operation not only from a technical but also a social perspective (for example, in accordance with the 'fairness-by-design' concept).[56] This is an extremely difficult task, however, requiring designers to have profound knowledge of the prejudices and inequalities that may be transferred to an AI system, particularly in the case of indirect discrimination, which is often not easy to identify.[57] Moreover, AI systems are often commercial products, and their source code (being a trade secret) is not made public; this is a significant limitation on attempts to analyse a system's bias before it is brought into use.[58] Thus, for this method of eliminating AI system bias to work, it is essential to enforce code transparency.[59] It should also be noted that independent tools are being created to identify algorithmic bias, such as the AI Fairness 360 Open Source Toolkit.[60] Therefore, it seems reasonable to postulate that the design and audit teams of such systems should consist not only of technical experts but also of ethicists and lawyers, especially if those systems could be used in an area connected with the rights and freedoms of an individual. As indicated above, an AI system may be functioning properly from a technical point of view but its use may still result in some negative social consequences. Not everything that is technically possible is at the same time ethically justified.

A second approach uses the possibility of human elucidation and verification of decisions that have been made by the system. This solution is of a corrective nature, enabling the elimination of biased decisions that the system itself has taken. It may

55 E. Ntoutsi et al., Bias, *op. cit.*, p. 9; F. Zuiderveen Borgesius, Discrimination, *op. cit.*, p. 19.

56 See F. Lattimore et al., Using Artificial Intelligence, *op. cit.*, p. 55.

57 B. Berendt and S. Preibusch, Toward Accountable Discrimination-Aware Data Mining: The Importance of Keeping the Human in the Loop – and Under the Looking-Glass, "Big Data" 2017, vol. 5, no. 2, p. 145.

58 I.D. Raji and J. Buolamwini, Actionable Auditing: Investigating the Impact of Publicly Naming Biased Performance Results of Commercial AI Products, "Conference on Artificial Intelligence, Ethics, and Society" 2019, p. 1, https://www.media.mit.edu/publications/actionable-auditing-investigating-the-impact-of-publicly-naming-biased-performance-results-of-commercial-ai-products/ (accessed 06.02.2021).

59 M.S. Cataleta and A. Cataleta, Artificial Intelligence, *op. cit.*, p. 47.

60 R. Rodrigues, Legal and Human Rights Issues, *op. cit.*, p. 3.

involve assigning to the AI system the role of 'advisor' to a human decision-maker[61] or allowing the system to take its own decisions but with the possibility of appeal to a human assessor. To make verification of the system's decisions possible at all, it must fulfil the requirement of explainability – that is, the possibility of presenting the system's decision-making process in a way that a human can understand.[62]

4. AI Systems as Tools for Reducing Social Inequalities

The right to equal treatment and the non-discrimination approach are expressed nowadays as human rights both in international conventions[63] and constitutions.[64] The essence of the principle of equality is that entities in a similar situation should be treated in a similar way, and entities in a different situation in a different way, respectively.[65] However, this principle is not absolute and does not mean that the rights of all individuals are identical. It should always be related to a certain situational context in order to properly assess a case.[66]

The principle of equal treatment may also be subject to limitations. For example, under Polish law it is acceptable to treat similar entities differently if this is in line with the principle of social justice.[67] Assessment as to whether such a differentiation is justified or not is based on the relevance of the differentiation's character, the proportionality of the arguments for differentiation and the constitutional basis for the differentiation.[68] One example of such a non-discriminatory differentiation is the so-called compensatory privilege, i.e. the one aimed at reducing inequalities actually occurring in social life.[69]

It seems that AI systems could be used as a tool to minimize inequalities due to the fact that those systems may be biased. Firstly, utilizing such systems and subsequent analysis of their decisions may allow revealing of prejudices hidden within the society, which could be exposed using statistical data. Secondly, it could be possible to facilitate the use of estimators to introduce corrective measures to the system (although often, due to the complexity of the situation and the multitude of

61 B. Berendt and S. Preibusch, Toward Accountable, *op. cit.*, p. 146.
62 E. Ntoutsi et al., Bias, *op. cit.*, p. 8.
63 For example, Art. 14 of the Act of 4 November 1950 – Convention for the Protection of Human Rights and Fundamental Freedoms (Journal of Laws 1993, No. 61, item 284).
64 Art. 32 of the Act of 2 April 1997 – The Constitution of the Republic of Poland (Journal of Laws 1997, No. 78, item 483, as amended).
65 W. Borysiak and L. Bosek, Komentarz do art. 32, (in:) M. Safjan and L. Bosek (eds.), Konstytucja RP. Tom I. Komentarz do art. 1–86, Warsaw 2016, p. 831.
66 *Ibidem*, pp. 831–832.
67 Judgement of the Constitutional Tribunal of 24 February 1999, SK 4/98, Lex No. 36177.
68 Judgement of the Constitutional Tribunal of 3 September 1996, K 10/96, Lex No. 25751.
69 Judgement of the Constitutional Tribunal of 28 March 2000, K 27/99, Lex No. 39995.

variables, it may prove to be difficult to implement in practice). Thirdly, it would be possible to design the system to 'reward' certain features as a means to achieve compensatory privilege. In the end, paradoxically, AI systems' bias can be used to remove real social inequalities.

However, some possible problems should be highlighted. The first of these concerns would be who should decide to introduce equalization mechanisms to such a system. Usually AI systems are commercial products made by private entities. Equipping them with such authorization to influence social reality seems too far-reaching, and it seems necessary to introduce mechanisms of cooperation with the state authorities. The problem, however, increases when such a system is to be used solely by the private entity (e.g. for employee recruitment or credit risk assessment). Then the involvement of the state authority in such cases would be limited. What is more, social inequalities existing in one country do not necessarily exist in others, and even if they do, not usually to the same extent. This means that AI systems used to reduce social inequalities would have to take into account the specificity of each country in which they are to be used.

The second problem is to establish a vision of the future reality that would be achieved with these systems. It would require a diagnosis of existing inequalities and the setting up of groups of people or features impacted by those inequalities. The next step would be to determine the appropriate direction of change. In a democratic state ruled by law, it should be established by means of a social consensus based on rational premises. An arbitrarily set direction of change could lead to replacing existing social inequalities with others, e.g. through disproportionately favouring certain groups.

Regardless of whether or not AI systems will be actively used to reduce social inequalities, or whether actions aimed at ensuring equal treatment will be limited to adjusting the decisions of such a system in individual cases, human involvement in the decision-making process seems indispensable. It appears that the limitations of the AI system combined with understanding the context (a human domain) would allow us to make the most of AI capabilities. On the one hand, the bias of AI systems does not in itself prejudge their rejection; on the other hand, these systems do not reduce social inequalities on their own but may be a powerful tool in the hands of a human.

Conclusion

Like any technology, artificial intelligence in itself is neither good nor bad. It is people who impart it such a character when they decide how a system is to be used. AI is used in various areas of human life and sometimes produces spectacular results, for example by improving the diagnosis of cancer. However, we must not lose sight of the fact that AI systems are not a remedy for the stereotypes, nurtured over many

years, that do harm to people with particular characteristics or to whole social groups. These may infect the operation of AI systems in various ways, including the use of biased data, bias resulting from the way the system functions and bias being an effect of the actions of the designer of the system or the person interpreting its decisions. This does not mean that people should stop using artificial intelligence – quite the reverse. It is necessary, however, to be aware of the limitations of such systems and to take measures to overcome those limitations, and also to understand that humans' decisions have a moral character and can affect the operation of the AI systems that they design and use.

It becomes necessary in this regard to take certain difficult decisions about whether we as a society are prepared to allow AI systems to ignore certain data (on race, for instance), accepting a certain reduction in the accuracy of the system's decisions and forecasts, for the sake of ensuring equality, understood as the treatment of similar individuals in similar ways. Another question to be answered is how far it is justified to take steps to eliminate statistical bias through the deliberate introduction of other types of bias into AI systems' operation. Although solutions of this type may reduce the effects of existing prejudices, they are based on a certain predefined vision of the world and may thus serve as a way of designing the future. It is therefore necessary to act with particular vigilance and to ask ourselves, while we still have time, what kind of future world that ought to be.

REFERENCES

Angwin J., Larson J., Mattu S. and Kirchner L., Machine Bias, ProPublica 2016, https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

Barfield W. and Pagallo U., Advanced Introduction to Law and Artificial Intelligence, Cheltenham/Northampton 2020.

Barocas S. and Selbst A.D., Big Data's disparate impact, "California Law Review" 2016, vol. 104, no. 2.

Berendt B., Preibusch S., Toward accountable discrimination-aware data mining: The importance of keeping human in the loop – and under the looking-glass, "Big Data" 2017, vol. 5, no. 2.

Boden M.A., Sztuczna inteligencja. Jej natura i przyszłość, trans. T. Sieczkowski, Łódź 2020.

Borysiak W. and Bosek L., Komentarz do art. 32, (in:) M. Safjan and L. Bosek (eds.), Konstytucja RP. Tom I. Komentarz do art. 1–86, Warsaw 2016.

Brennan T., Dieterich W. and Ehret B., Evaluating the predictive validity of the COMPAS risk and needs assessment system, "Criminal Justice and Behavior" 2009, vol. 36, no. 1.

Cataleta M.S. and Cataleta A., Artificial Intelligence and Human Rights, an Unequal Struggle, "CIFILE Journal of International Law" 2020, vol. 1, no. 2.

Coeckelbergh M., AI Ethics, Cambridge/London 2020.

Cummings M.L., Automation and Accountability in Decision Support System Interface Design, "The Journal of Technology Studies" 2006, vol. 32, no. 1.

Danks D. and London A.J., Algorithmic Bias in Autonomous Systems, 'Proceedings of the 26th International Joint Conference on Artificial Intelligence (IJCAI 2017)', https://www.cmu.edu/dietrich/philosophy/docs/london/IJCAI17-AlgorithmicBias-Distrib.pdf.

Davenport T. and Kalakota R., The potential for artificial intelligence in healthcare, "Future Healthcare Journal" 2019, vol. 6, no. 2.

Dymitruk M., Sztuczna inteligencja w wymiarze sprawiedliwości? (in:) L. Lai and M. Świerczyński (eds.), Prawo sztucznej inteligencji, Warsaw 2020.

European Parliament resolution of 20 October 2020 with recommendations to the Commission on a framework of ethical aspects of artificial intelligence, robotics and related technologies (2020/2012(INL)).

Fjeld J., Achten N., Hilligoss H., Nagy A. and Srikumar M., Principled Artificial Intelligence. Mapping Consensus in Ethical and Rights-Based Approaches to Principles for AI, Cambridge 2020.

Flasiński M., Wstęp do sztucznej inteligencji, Warsaw 2020.

Fry H., Hello world. Jak być człowiekiem w epoce maszyn, trans. S. Musielak, Krakow 2019.

German S., Bienstock E. and Doursat R., Neural networks and bias/variance dilemma, "Neural Computation" 1992, vol. 4, no. 1.

Hacker P., Teaching Fairness to Artificial Intelligence: Existing and Novel Strategies against Algorithmic Discrimination under EU Law, "Common Market Law Review" 2018, vol. 55.

High-Level Expert Group on Artificial Intelligence (appointed by the European Commission in June 2018), A Definition of Artificial Intelligence: Main Capabilities and Scientific Disciplines, Brussels 2019.

High-Level Expert Group on Artificial Intelligence, Ethics Guidelines for Trustworthy AI, Brussels 2019.

Jernigan C. and Mistree B.F., Gaydar: Facebook friendships expose sexual orientation, "First Monday" 2009, vol. 14, no. 10.

Kasperska A., Problemy zastosowania sztucznych sieci neuronalnych w praktyce prawniczej, „Przegląd Prawa Publicznego" 2017, no. 11.

Lattimore F., O'Callaghan S., Paleologos Z., Reid A., Santow E., Sargeant H. and Thomsen A., Using artificial intelligence to make decisions: Addressing the problem of algorithmic bias. Technical Paper, Australian Human Rights Commission, Sydney 2020.

Massey G. and Ehrensberger-Dow M., Machine learning: Implications for translator education, "Lebende Sprachen" 2017, vol. 62, no. 2.

Michie D., Methodologies from Machine Learning in Data Analysis and Software, "The Computer Journal" 1991, vol. 34, no. 6.

Neff G. and Nagy P., Talking to Bots: Symbiotic Agency and the Case of Tay, "International Journal of Communication" 2016, no. 10.

Ntoutsi E., Fafalios P., Gadiraju U., Iosifidis V., Nejdl W., Vidal M.-E., Ruggieri S., Turini F., Papadopoulos S., Krasanakis E., Kompatsiaris I., Kinder-Kurlanda K., Wagner C., Karimi F., Fernandez M., Alani H., Berendt B., Kruegel T., Heinze Ch., Broelemann K., Kasneci G., Tiropanis T. and Staab S., Bias in data-driven artificial intelligence systems – An introductory survey, "WIREs Data Mining Knowledge Discovery" 2020, vol. 10, no. 3.

O'Neil C., Broń matematycznej zagłady. Jak algorytmy zwiększają nierówności i zagrażają demokracji, trans. M. Z. Zieliński, Warsaw 2017.

Raji I.D., Buolamwini J., Actionable Auditing: Investigating the Impact of Publicly Naming Biased Performance Results of Commercial AI Products, 'Conference on Artificial Intelligence, Ethics, and Society' 2019, https://www.media.mit.edu/publications/actionable-auditing-investigating-the-impact-of-publicly-naming-biased-performance-results-of-commercial-ai-products/.

Ribeiro M.T., Singh S. and Guestrin C., „Why Should I Trust You?" Explaining the Predictions of Any Classifier, "22nd ACM SIGKDD International Conference 2016, San Francisco", https://www.kdd.org/kdd2016/papers/files/rfp0573-ribeiroA.pdf.

Rodrigues R., Legal and human rights issues of AI: Gaps, challenges and vulnerabilities, "Journal of Responsible Technology" 2020, vol. 4.

Roselli D., Matthews J., Talagala N., Managing Bias in AI, "Companion Proceedings of the 2019 World Wide Web Conference, San Francisco, CA, USA", May 2019.

Rutkowski L., Metody i techniki sztucznej inteligencji, Warsaw 2012.

White Paper On Artificial Intelligence. A European approach to excellence and trust, COM(2020) 65 final, European Commission, Brussels 2020.

Yapo A. and Weiss J., Ethical Implications of Bias In Machine Learning, "Proceedings of the Annual Hawaii International Conference on System Sciences" 2018.

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