ETHICAL STANDARDS FOR PUBLIC SECTOR DIGITALIZATION

Dragoș BÎGU [a*], Dominic-Ștefan GEORGESCU [b]

[a, b] Bucharest University of Economic Studies, Romania

ABSTRACT
In this paper, we examine from an ethical perspective the AI tools used in the public sector. After a presentation of the categories of AI applications used in the developed countries in the public institutions, We briefly examine some AI ethical documents of the most important international organizations in order to identify the main ethical aspects relevant for this field, In the following three sections, we examine the central ethical themes identified in the ethical documents analyzed: fairness and non-discrimination, transparency and accountability, safety and privacy. For all these themes, we present and discuss ethical standards for the use of AI tools in the public sector. In the last section, we focus on two specific problems relevant for the use of algorithms in public decision-making, which would need clarification in ethical guidelines.

KEYWORDS: *AI ethics, fairness, privacy, public sector, transparency.*

DOI: 10.24818/IMC/2023/05.03

1. INTRODUCTION

In recent years, in many economically developed countries, AI tools have started to be used in the daily activity of public institutions, to help public servants in their decisions. AI algorithms are used to identify cases with a high risk of fraud in benefits claims, for the prevention of sham marriage or for the identification, through facial recognition, of the persons being pursued. Although such algorithms bring benefits in the provision of public services, their use faces a series of ethical problems, which we will discuss in this paper. The goal of the paper is to examine from an ethical perspective the conditions that must be met by AI tools used in public services.
We will start with a general presentation of the way in which AI instruments are used in public decision-making. Then we will carry out an analysis of the most important documents (guidelines, good practice principles) developed by international institutions to guide the use of the AI instruments, especially in the public sector. In the following three sections, we will discuss the main ethical problems relevant for the use of AI in the public sector, as resulted from our document analysis. In the last section, we will present and examine some detailed ethical aspects that are not captured in the documents analyzed, but are necessary for a complete and adequate analysis of the criteria that should be met by AI tools used in the public sector.

2. HOW TO USE AI TOOLS IN THE PUBLIC SECTOR

In this section, we will present how the AI tools are used in the public services. AI consists of using digital technology in order to create systems that have the ability of performing human-like tasks,

* Corresponding author. E-mail address: *dragos.bigu@man.ase.ro*

such as understanding, replying to spoken or written language, analyzing data, making recommendations. Many AI tools involve using statistical instruments to analyze large amounts of data to find patterns, useful for various objectives, such as making predictions or spotting anomalies. At the same time, AI involves performing repetitive tasks without constant human guidance. Machine learning is a form of artificial intelligence that uses algorithms that learn from experience to improve their performance in specific fields.

Machine learning algorithms use large quantities of training data to produce accurate results. The larger volume of the data is, the more accurate the result will be. The model improves in time, as the algorithm is exposed to more data. There are two approaches to machine learning: supervised and unsupervised (Delua, 2021). In supervised learning, training data are labeled, which means that the algorithm is informed on relations between inputs and outputs. For instance, in order to train a recognition algorithm, we will use labeled data in which we ascribe categories to various patterns of points, for example the algorithm is trained what animals are represented in a set of images. Based on these training data, the algorithm will give the output that in another image a certain animal is represented. In unsupervised learning, training data are unlabeled. For instance, the algorithm receives a large volume of data and has the task to find anomalous data. Data are not labeled, which means that the algorithm is not informed that certain data points are normal or anomalous. The algorithm will find that most data points satisfy certain conditions, and they will be labeled as normal, and the others are anomalous.

Most AI algorithms have two tasks: prediction and classification (Kozyrkov, 2023). The main statistical tool used for prediction is regression analysis. The purpose of regression analysis is to model the relationship between a variable that is called outcome variable and a set of other variables called predictors. For this purpose, the algorithm will find the line that fits best the data and, based on it, will predict a new value of the outcome variable. The main statistical tool for classification is cluster analysis. The purpose of cluster analysis is to group a set of data points based on similarities and differences. The criteria of similarity are not initially given, but extracted from data by algorithm.

In the public sector, AI and especially machine learning tools are used for many purposes (Office for Artificial Intelligence, 2020). First, AI tools can provide more accurate information and predictions, which could lead to better results. For instance, AI tools can help us make predictions about the phenomenon of food security. Secondly, AI provides tentative solutions for some challenging social problems, for example governance of dense urban areas (transit services, infrastructure management, urban planning). Thirdly, AI tools can simulate complex situations, systems and processes, which allow policy makers to do experiments to anticipate outputs of various policies in a rapid way and without harming individuals and communities, in the case of bad decisions. Public budget simulations are a good example of how using computer simulations can contribute to better government decisions. Fourthly, AI can increase the efficiency and effectiveness of public services, by providing service providers and policy makers with more reliable and accurate information. For instance, UNICEF Office of Innovation uses AI tools to map all schools in the world, in order to help national governments improve their education systems. Fifthly, AI applications can help public institutions detect possible frauds. For instance, the UK Home Office uses a tool to detect potential sham marriages, used to gain the right to enter and reside in the UK (Seddon, 2021). Finally, AI applications are used to perform repetitive and time-consuming tasks. For instance, in many countries. public institutions use chatbots to interact with public service customers.

In order to provide just a small glimpse of the variety of IA tools that are used in the public sector, we will continue with some examples of applications that are used in European countries. Since 2018, the Register of Enterprises of Latvia has used a chatbot to answer the questions regarding the process of companies' registration (Van Noordt & Misuraca, 2019). In the Swedish municipality of Upplands-Bro, the robot Tengai is used to interview job candidates for public jobs (European Commission. Joint Research Centre., 2020). Pacific Northwest National Laboratory, a Department of Energy national

laboratory, has developed an AI tool called Transportation State Estimation Capability (TranSEC). which uses Uber data to analyze traffic to optimize traffic flow (Hede, 2020). These AI applications and many others are undoubtedly useful for increasing public service quality and over time they will become more and more developed and widely adopted by public institutions. Nevertheless, some AI tools used in the public services face ethical problems, which we will discuss in the following sections.

3. GUIDELINES ON AI ETHICS: A BRIEF DOCUMENT ANALYSIS

The aim of this section is to conduct an analysis of the guidelines issued by international institutions on AI ethics. The landscape of AI ethics guidelines issued by public institutions is very rich. Many international organizations issued in the last years such guidelines: United Nations Educational, Scientific and Cultural Organization (UNESCO), United Nations (UN), Council of Europe (CoE), World Bank, Organisation for Economic Co-operation and Development (OECD), European Commission. At the same time, many developed countries have their own guidelines on AI ethics, issued by national public institutions. *Artificial Intelligence. Australia's Ethics Framework. A discussion Paper*, issued by Department of Industry Innovation and Science from Australia, *Work in the age of artificial intelligence. Four perspectives on the economy, employment, skills and ethics*, issued by the Ministry of Economic Affairs and Employment, from Finland; *How can humans keep the upper hand? Report on the ethical matters raised by AI algorithms*, issued by French Data Protection Authority, from France; *A guide to using artificial intelligence in the public sector*, issued by Office for Artificial Intelligence, from the United Kingdom are just four examples of such documents. Finally, many general AI strategies have a developed ethical component. We will analyze the documents issued by the five international and European organizations, in order to identify the main ethical aspects and values relevant for using AI in the public service.

In 2021, UNESCO issued *Recommendation on the Ethics of Artificial Intelligence*, which includes four values – respect and protection of human rights, environment and ecosystem flourishing, diversity and inclusiveness, and living in peaceful, and just and interconnected society – and some ethical principles, summarized under ten rubrics: proportionality and non-harming, fairness and non-discrimination, safety and security, sustainability, privacy and data protection, human oversight, transparency and explainability, responsibility and accountability, awareness and literacy and multi-stakeholder collaboration (UNESCO, 2021).

In 2022, Inter-Agency Working Group on Artificial Intelligence – an interagency working group on AI (IAWG-AI), co-led by UNESCO and *The* International Telecommunication Union *(*the United Nations agency for information and communication technologies) issued *Principles for the Ethical Use of Artificial Intelligence in the United Nations System* (United Nations, 2022) The ten principles developed in this document are: non-harming; defined purpose, necessity and proportionality; safety and security, fairness and non-discrimination; sustainability; privacy and data protection; human autonomy and oversight; transparency and explainability; responsibility and accountability; inclusion and participation.

Ad hoc Committee on Artificial Intelligence (CAHAI) Policy Development Group – a working group created under the auspices of CoE to examine the feasibility and potential elements of a legal framework for artificial intelligence – issued a draft document called *Artificial Intelligence in Public Sector*, which lists some areas of risk the use of AI tools in the public sector. Some of the areas of risk mentioned are: data governance, unemployment caused by the use of AI, biases and discrimination, explainability, accountability, transparency, liability, trustworthiness and legitimacy (CAHAI, Policy Development Group, 2021).

OECD identifies five values central for AI ethics – inclusive growth, sustainable development and well-being; fairness and respect for human values and rights; transparency; security, safety and

robustness; accountability – and formulates five principles correlated to these values (OECD, n.d.).
A World Bank team drew up the paper *Artificial Intelligence in the Public Sector*, which includes a list of eight principles: privacy and data protection, accountability, safety and security, transparency and explainability, fairness, human control of technology, professional responsibility, promotion of human values (World Bank, 2020).
In 2018, an independent expert group on artificial intelligence, set up by the European Commission issued a document entitled *Ethics Guidelines for Trustworthy AI*. The document formulates four ethical principles for the use of AI tools: respect for human autonomy, prevention of harm, fairness, and explicability. The document also formulates seven requirements for AI tools: human agency; technical robustness and safety; privacy; transparency; diversity; non-discrimination and fairness; societal and environmental wellbeing and accountability (European Commission, 2019).
Some values are found in all or almost all documents (sometimes with different words): fairness and non-discrimination, transparency and explainability, accountability, human control and oversight, safety, privacy. A similar list of values was found in a much wider research on the 84 documents containing ethical principles for AI (Jobin et al., 2019). Transparency and related values like explainability and interpretability were found in 73 of 84 documents. Justice and fairness (and related values like non-discrimination, equity, diversity were found in 68 documents. Non-maleficence and related values like safety and security were found in 60 documents. Responsibility and related values like accountability and liability were found in 60 documents. Privacy was found in 47 documents. Other values, like sustainability and beneficence, were found in less than a half of documents
We will examine these ethical aspects in the following three sections: fairness and non-discrimination in the first one, transparency, explainability and accountability in the second one, and safety and privacy in the third one.

4. FAIRNESS AND NON-DISCRIMINATION IN AI MODELS. THE COMPAS CASE

AI tools used in the decision-making process should be fair. This is even more true in the public sector, where institutions have a further duty to equally treat all citizens and groups. Fairness involves two dimensions: individual fairness – referring to how individuals are treated – and group fairness – referring to how groups, or categories, are treated: race, gender, age categories, urban vs rural categories. Also, at the national level, AI tools should not discriminate against some countries (UNESCO 2021, 20).
Many AI applications used in the public services were criticized for being biased against some protected categories. A famous example is COMPAS (Correctional Offender Management Profiling for Alternative Sanction), an application created by the Michigan-based company Northpointe, which is used for assessing a defendant's risk of becoming a recidivist, relevant for decision on pre-trial detention and parole. Based on around 100 factors, COMPAS ascribes a score from 1 to 10 to each defendant and classifies risk of recidivism as high (scores from 8 to 10), medium (5-7) and low (1-4), based on a very large number of variables (Corbett-Davies et al, 2016). In spite of the proprietary secrecy over the algorithms used by any AI tool, we can assume that COMPAS, like other algorithms, uses a regression model to assess the risk level and to predict offenders' reoffending outcome.
An analysis carried out by ProPublica, a nonprofit organization in the field of investigative journalism, shows that the number of false positives (high-risk defendants that did not reoffend in two years) is much higher among the black offenders that did not reoffend than among the white ones (45 vs 23 percent). At the same time, the rate of false negatives (low-risk defendants that reoffend) is much higher among the white offenders that reoffend than among the black ones (48 vs 28 percent) (Larson et al., 2016).
The company Northpointe replies to ProPublica's objection by showing that this result is the consequence of the fact that on average the black offenders have a higher rate of reoffending than the

white ones (Dieterich, William et al., 2016, pp. 3–4). It argues that actually the rate of reoffending should be calculated relative to the number of offenders with a certain score. If we do like this, the result is that the white and black offenders with the same risk level have similar rates of reoffending. Even more, in most cases, the two criteria of fairness – that used by ProPublica, relative to the number of defendants that did not reoffend. and that used by Northpointe, relative to the number of offenders with a certain risk of reoffending – cannot be satisfied together (Washington Post, 2016).

The following numerical example can help us understand why the two criteria of fairness cannot be satisfied together. Let's suppose that, in a small sample, 5 white offenders and 20 black offenders are labeled as high-risk and that the rate of successful prediction is 80% for both categories, i.e. 80% of the high-risk offenders will reoffend. As a result, the number of false positives (high-risk offenders that did not reoffend) is 1 white and 4 black offenders. Therefore, 4 of 5 the five false positives are black (80%) and only one is white (20%), in spite of the fact that the rate of successful predictions is the same for white and black offenders.

The two criteria are both satisfied in two situations: when the prediction is perfect and consequently the rate of false negatives is 0, and when the rate of reoffending is similar for both categories. We will not be especially interested in the debate upon which of the two criteria is more adequate, but it should be stressed that it shows that the criteria of fairness are not obvious and that statistical, or, more generally, mathematical tools are not enough to settle it.

There is also a second argument for the fact that some AI tools may be biased. Taking the example of COMPAS, if black people are incarcerated more often than the white ones even when the crime rates are similar, the historical data on crime and reoffending rates will be biased. Therefore, input historical data upon which the result of the algorithm is based is biased, and the bias will be reinforced and amplified by the algorithm, since people base their decisions on its results and have a scientific ground for their biased decision, which would not have happened without the algorithm. This is a serious cause of algorithmic bias, which cannot be easily avoided, since it is difficult to estimate the size of the bias, in order to incorporate it. It is worth mentioning that this type of bias is not the result of a certain algorithm, but rather of biased input data, which cannot be avoided by a better algorithm.

The discussion on AI algorithmic fairness is mainly focused on discrimination of some categories, but this is not the only dimension of fairness. Individual fairness is also relevant: an offender can consider herself unfairly treated if she is not released before the trial because it is wrongly identified by an AI application as having a high risk of reoffending. If this is the case, a minimum rate of accuracy is a necessary condition of fairness. Anyway, when a choice of using an AI tool in the decision-making process is made it should not be forgotten that human decision is not error-free either.

COMPAS is not the only application criticized for its unfair results. The UK Home office introduced in 2015 an application to detect sham marriages. It flags – based on eight variables, such as the age difference between husband and wife, and travel history – of some couples, which will be further investigated by the Home Office. This application was also criticized for being discriminatory. Nevertheless, the situation is different from that in the COMPAS case. In the COMPAS case, studies compare prediction on reoffending (the risk score translates in a probabilistic prediction about reoffending) with real rates of reoffending, rates that can be known, since it is relatively easy to check whether a certain defendant reoffended or not. It is not the same in the case of the algorithm used to flag sham marriages, since we cannot have access to evidence that a certain marriage is actually sham. So, there is no way to compare predicted and real figures, as in the COMPAS case Public Law Project, a legal NGO, launched an action against using the algorithm.

One important piece of evidence brought against using this algorithm is that four nationalities have a significantly higher rate of flagging, between 20% and 25%: Bulgaria, Greece, Romania and Albania (Milliken, 2021). Nevertheless, there is no information about the real rate of sham marriages in this country, and it might be true that these four countries have a higher rate of sham marriages. That is

not to say that there are no arguments against using the algorithm for detecting sham marriages, but only that the high rate of flagged marriages is not necessarily discriminatory. Other general arguments against using AI tools in public decision-making, which will be further discussed, are applicable to this tool as well.

5. TRANSPARENCY OF AI SYSTEMS AND AI ACCOUNTABILITY

One of the objections most often brought against the use of algorithms for public decision-making is that it strongly violates the requirement of transparency: in many cases citizens are not even aware that public institutions use such algorithms, even when they know that an AI tool is used, they do not know how it works, and – an even more serious problem – public servants that have the final decision do not understand how the result is obtained. Transparency is a fundamental value in the public sector. While companies have mainly a duty to pursue the interests of shareholders, public institutions have a duty to all citizens. This is why the activity of public institutions should be open to public scrutiny. Public institutions in all countries, even in the most developed ones, suffer from democratic deficit. This is the consequence of the fact that citizens believe that they do not have control over the decisions that influence their lives. Even if citizens elect their representatives, more and more decisions are taken by public servants, who are not elected by citizens and are not directly accountable to them. In a way, the AI tools used for public decision-making worsen this democratic deficit: the decisions are taken with the help of AI algorithms protected by trade secrets and, consequently, citizens do not know how the algorithm works, what factors it takes into account, and what data are used for training. There is an obvious tension between protection of trade secrets and transparency: complete transparency can only be achieved Trade secret rights are essential for protecting companies' valuable information that provide a competitive advantage. In their absence, companies will not have the incentive to innovate and to invest in research. However, transparency is equally important. Without a sufficient degree of transparency, citizens cannot trust that public institutions are truly working in their best interest. At the same time, transparency is the only one that can ensure the possibility of accountability: errors can be sanctioned only when the activity of public institutions is transparent. The transparency requirement refers to three dimensions: transparency towards the public, transparency towards the public employees who actually use the respective algorithm, transparency in courts, if someone challenges an AI-generated decision.

The first dimension of transparency is important for ensuring citizens' trust in how algorithms are used to make decisions. First, without prejudice to trade secrets, institutions should be obliged to inform the public about the algorithms used in decision-making and, more specifically, the decisions in which they are used. Citizens also have the right to know when a decision affecting them is made with the help of an AI tool. Second, the companies that develop those algorithms should provide an overview of how they work in a way that is comprehensible to non-specialists. Compliance with these two requirements is necessary to ensure transparency towards the public.

The second dimension of transparency is relevant to ensure that public servants who actually use these tools can effectively appreciate when there is a risk that an algorithm result is inappropriate. For this, public employees need to receive more detailed information on how the algorithm works and especially the data with which it was trained, which will help them appreciate when the results provided by the algorithm are not well founded or do not agree with some principles. This transparency towards algorithm users must provide a real possibility for them to monitor the algorithms and justifiably violate their recommendations where necessary. In the absence of relevant information, the right of public employees to change an algorithm's recommendation becomes irrelevant.

The third dimension of transparency is relevant when an individual wants to challenge an AI-generated decision. Citizen's right to contest in court such a decision should be recognized

(Stankovich et al., 2023, p. 16). Transparency is a necessary condition for contestability. The principle of contestability, which rarely appears explicitly among the fundamental principles of the ethics of artificial intelligence, states that when an algorithm has a significant impact on a person, he must have the opportunity to challenge the result of the algorithm (Zalnieriute & Gould-Fensom, 2019). These potential challenges do not have to refer to inadequacies of the algorithm, but may refer to the data with which the algorithm was trained or even to the erroneous introduction of the data regarding that individual case. The ability to challenge the use of an algorithm means that any information about the algorithm relevant to the case must be disclosed in court, even if it is considered a trade secret. In this case, the legal framework must ensure that the said trade secret will be protected, mainly by two means: limiting those to whom the algorithm information is disclosed and prohibiting them from disclosing the information. Even if the two requirements in tension – transparency and trade secret protection – cannot be completely satisfied, a compromise between them is necessary.

6. SAFETY AND PRIVACY

The last section that discusses ethical requirements of AI tools in the public sector will be focused on safety and privacy. Machine learning algorithms are vulnerable to attacks. These attacks intentionally introduce, in the training phase, inaccurate or falsified data in order to manipulate the model or, after the training phase, corrupt data to mislead the algorithm. Three of the most common attacks against machine learning models are poisoning, evasion attack and model extraction. In poisoning, incorrectly labeled data are introduced in the training phase in order to make the system wrongly classify inputs and give wrong answers. In evasion attacks, data are modified to evade detection or to wrongly classify some input as legitimate. In model extraction, an attacker tries to reconstruct the model or to extract the data the algorithm was trained on. Public institutions, by regulations, and AI companies should work together to ensure AI safety.

Attacks against machine learning systems used in the public sector can lead to significant harm, resulting from misclassifications or wrong predictions resulting from inaccurate models. But the topic of the risk posed by AI algorithms is not limited to attacks. Some technologies were accused of posing a significant risk on citizens. Facial recognition is such an example of technology that can have significantly harmful uses, for instance mass surveillance. For such technologies, a cost-benefit analysis is necessary. At the same time, public institutions should ensure that technologies used do not violate fundamental human rights and freedoms.

The last requirement for the AI tools used in the public sector is privacy. As relevant for the topic discussed here, privacy involves two main dimensions: the former related to the management of personal information and the latter to preserving a private sphere. Concerning the first dimension, AI actors should carefully manage the huge volumes of data used by AI tools. Public institutions have access to huge amounts of sensitive citizens data that should be kept private by anonymization and de-identification. The aggregation of these data makes the problem more difficult: more data on citizens are aggregated, higher the risk of identification of the person is. For this reason, gathering excessive volumes of data can pose a risk of privacy harm. For instance, using facial recognition algorithms in public spaces There is a tradeoff between privacy and efficiency: maximum efficiency can be achieved only by gathering high volumes of data, which leads to a privacy risk.

In its second dimension, privacy requires preserving a personal space free of any interference. For instance, there were developed some software tools for recognizing emotions, which can be used in various fields, for instance in the recruitment process. Apart from the concern that there is not sufficient scientific evidence that such tools give accurate results, an even more serious problem is that they interfere in a sphere generally considered personal.

In a different way, the social credit system developed in China also affects citizens' privacy. Its general idea is that huge volumes of data on individuals and businesses are gathered in order to assess

their trustworthiness, in order to penalize undesirable conduct. The social credit system is an extension of the credit score system used in many countries to assess persons' creditworthiness based on their credit history: loans, debt payment, mortgage refinancing, etc. The social credit system extends the general idea of the credit score system beyond banking to the many aspects of social life(Mac Síthigh & Siems, 2019, p. 2). Extension is in two ways. First, more elements are taken into account: minor offenses, like smoking in a non-smoking area, bad driving, smoking in non-smoking zones, posting fake news online. Secondly, inadequate behavior leads to sanctions in various spheres of social life: ability to rent, to travel, employment. It is noteworthy that the social credit system does not give an aggregate score for each citizen, but rather consists of some scores for particular fields. Anyway, after testing this system, these scores can be aggregated and the resulting system will be dangerous for citizens' freedom.

We will not discuss in detail the social credit system, about which there is not enough information. From our perspective, the important thing is to understand why it is considered a negative example. For instance, according to a proposal for a regulation, adopted by European Commission, social credit scoring should be banned in Europe Union (European Commission, 2021, p. 13), After all, the credit score used in many countries also penalize citizens (by decreasing their chance to take a loan) who act an undesirable way (for instance pay their debt late). The distinction is important: social credit system is much more pervasive, involves many spheres of social life and, as a result, can be seen as a tool that violates citizens' privacy, while credit score refers to an isolated field. Therefore, AI tools that involve a pervasive control of citizens' life and that affect many spheres of their lives pose a danger for citizens' privacy. At the same time, it might be argued that the social credit system uses a complex instrument, which involves a significant number of factors, which is not proportional to the aim that should be achieved. By this, the proportionality principle – which states that the method choses should be proportional to the aim pursued – is violated.

7. TWO FURTHER ETHICAL ASPECTS REGARDING THE USE OF AI ALGORITHMS IN THE PUBLIC DECISION-MAKING

The ethical documents examined in the second sections are, naturally, general, so that they cannot answer all important questions regarding the use of AI in the public sector. They should be approached in more detailed documents. Next, we will discuss two points that, in our view, should be captured in regulations or recommendations.

The first point that should be clarified is what are the acceptable purposes of an AI tool in public decision-making. For instance, probably it would not be acceptable to use an AI tool to recommend a verdict, i.e. whether a defendant is guilty or not. The main reason for this is that an AI algorithm would use, like COMPAS, the model of predicting the verdict based on a set of variables. For instance, it would use in its decision the statistical fact that most defendants that belong to a certain category are found guilty. But, regardless of the categories taken into account, such a treatment would not be just, since due process right requires an individualized verdict. It is true that a judge would have the authority to override an algorithm's verdict, but even in this case algorithm's verdicts would have an unfair influence, which is not acceptable. Therefore, we cannot accept using AI algorithms for all public sector decisions and some general rules are needed in order to clarify for what kind of decisions we can use AI tools.

The second point that needs clarification is whether it should be considered fair to use protected categories in AI algorithms. Many companies that create algorithms do not include protected categories – such gender and race – among explicit variables taken into account, since they probably interpret that as being discriminatory. But this is not obvious at all.

First, if belonging to a protected category is really relevant for a certain prediction (for instance, if black offenders are actually more likely to reoffend than the white ones), taking race into account

would increase prediction's accuracy. In this case, it would be difficult to argue that race, or other protected variables, should not be taken into account. Furthermore, not including protected categories as variables can sometimes have a discriminatory effect. For instance, Melissa Hamilton (2019) argues that COMPAS has also a gender bias against women and that including gender among variables would eliminate it (Hamilton, 2019). If this inclusion is desirable, fairness would require including race as a variable as well. Secondly, in most cases not including protected categories (race, age, gender) as variables does not lead to a different result, since many variables strongly correlated to them are included. Removing all of them is not feasible and, at least in some cases. would cancel any predictive force of the model.

In a relevant case, State vs Loomis, the defendant, accused by a drive-by shooting, argued that use of COMPAS in his case violates his due process right. The judge of the case, Ann Walsh Bradley, stated in her decision that using gender as a variable in reoffending risk assessment should not be considered discriminatory in itself, but can be used for increasing accuracy (The Harvard Law Review Association, 2016, p. 1532). Therefore, we don't think that using variables for protected categories in this situation is discriminatory, although in other cases it can be. Nevertheless, public institutions should play a central role in clarifying whether using variables for protected categories should be considered discriminatory in specific situations.

Discrimination generated by using variables for protected variables is a particular case of statistical discrimination, as theorized by Kenneth Arrow and other economists. They show that most cases of labor market discrimination can be explained by the fact that employers tend to base their decisions on statements regarding average performance of some categories, because it is very expensive and inefficient for employers to collect individual information on all applicants. For instance, since on average men have more physical strength than women, if the number of applicants is large, it is efficient for employers to refuse all women before any recruitment test. It is true that there is a risk – a low one – for the employer to lose the best applicant, but even in this case, the cost of the selection process would be, most likely, higher than the benefit obtained by hiring the best applicant instead of the best male applicant.

Furthermore, also for reasons of efficiency, employers base their decisions on characteristics of applicants that are easily observable, like race or gender, even if other characteristics, which are more difficult to detect, would be more relevant. This type of discrimination is economically efficient for employers, but, at least in the employment field, is considered illegal and unfair. The reason is easy to understand: it is not fair to sanction applicants for their belonging to a certain race or gender; in our previous example it is not fair not to hire the best applicant, who is a woman, only because she belongs to a certain category. Furthermore, members of some categories are unfairly disadvantaged in many jobs and even in other spheres of life. These categories are expressly protected by national regulations in many countries. It is important to mention that these arguments work even if the statement about average performance of the two categories is true.

The problem is that most algorithms used in the public sector discriminate in a similar way as the employer from the previous example does. For instance, COMPAS bases its prediction about reoffending on the statements about average behavior of black offenders, and it can be argued that it is not fair to use this prediction, since a particular black offender should not be sanctioned, for instance by not releasing him on parole, for their belonging to a certain category. If this is the case, using COMPAS for predicting the reoffending behavior should also be considered unfair and illegal. As a reply to this argument, it can be noticed that there is an important distinction between the two situations: the choice not to hire a woman, based on a statistical statement, and the choice not to release a black offender on parole. In the employee selection case, individual evaluation of applicants is possible, therefore the decision to base the selection on statistical data on general categories is just a solution favored for reasons of efficiency. On the contrary, in the case of the decision whether to release a defendant on parole, there is no individual test to check whether the defendant will reoffend.

This means that decisions on parole release, whether human or AI-generated, are based on general factors and statistical generalizations about the categories to which offenders belong. As we have already mentioned, a solution would still be to avoid protected variables, such as gender, race and age. This solution has two shortcomings. First, since AI algorithms used for public decision-making are based on many variables, it's likely that other variables in that set would be strongly linked to those for protected categories. Secondly, avoiding variables for protected categories would sacrifice accuracy. A large part of the AI ethics literature accepts a trade-off between accuracy and fairness. Accuracy is also correlated to efficiency, since inaccurate decisions will lead, regardless of the field in which algorithms are used, to a certain inefficiency. One of the most important advantages of using AI algorithms is the fact that statistical generalizations are obtained through an objective and systematically applied method, which is not always the case for human decisions. For this reason, the value of accuracy plays and should play a central role in the ethics of public-sector algorithms.
The discussion on the fairness of statistical discrimination is much more complex and cannot be covered here. It refers not only to discrimination in employment, but also in consumer relations. In some cases, non-discrimination is considered a central value. For instance, in most countries, insurance companies are not allowed to use gender as a factor in setting prices of their products, such as health, life and auto insurance. This could lead to a loss in accuracy in risk assessment and, consequently, a loss in efficiency, since in some cases gender can be a relevant factor (for instance, some studies show that men are involved in more accidents than women). Therefore, in the fairness-accuracy tradeoff, the latter one is sacrificed to a certain degree (probably, a low one). Public institutions should decide whether the same approach should be enforced in the case of algorithms used in public decision-making.
It is not necessary to have the same decision for all algorithms, since situations can be different in some respects. For instance, the price of health insurance varies based on age, and the decision not to do this would significantly affect the accuracy of assessing risk and it would have a harmful impact on the company and on the younger clients. For this reason, setting the health insurance price based on age is considered morally and legally acceptable, but not discriminatory.

8. CONCLUSIONS

In this article, we examined the main ethical aspects regarding the use of AI algorithms in public decision-making. We want to stress two conclusions. First, general ethical guidelines are not sufficient for solving all ethical difficulties. Apart from these necessary ethical guidelines, other more detailed discussions are necessary. Two of them are approached in the previous section, but other points also need clarification. For instance, as resulted from the third section, the test of fairness that should be applied to AI algorithms needs also clarification. The second point that we want to emphasize is that the ethical decisions on AI algorithms are filled with tradeoffs. Two of them were stressed through the article: fairness-accuracy, privacy-efficiency, but other tradeoffs can also be noticed. In a detailed ethical examination of AI tools used in the public sector, these tradeoffs need careful treatment.

ACKNOWLEDGMENT

The research presented in this article was funded within the institutional project "Integration of digitalization opportunities in improving the management process of state institutions in Romania".

REFERENCES

Ad hoc Committee on Artificial Intelligence Policy Development Group. (2021). Artificial Intelligence in Public Sector Retrieved December 3, 2023, from https://rm.coe.int/cahai-pdg-2021-03-subwg2-ai-in-public-sector-final-draft-12032021-2751/1680a1c066.

Corbett-Davies, S., Pierson, E., Feller, A & Sharad, G. (2016, October 17) A computer program used for bail and sentencing decisions was labeled biased against blacks. It's actually not that clear. Retrieved 10 December 2023, from https://www.washingtonpost.com/news/monkey-cage/wp/2016/10/17/can-an-algorithm-be-racist-our-analysis-is-more-cautious-than-propublicas/

Delua, J. (2021, March 12). Supervised vs. Unsupervised Learning: What's the Difference? IBM Blog. Retrieved December 10, from https://www.ibm.com/blog/supervised-vs-unsupervised-learning/www.ibm.com/blog/ supervised-vs-unsupervised-learning

Dieterich, W., Mendoza, C., & Brennan, T. (n.d.). COMPAS Risk Scales: Demonstrating Accuracy Equity and Predictive Parity. Retrieved 2 December 2023, from https://www.documentcloud.org/documents/2998391-ProPublica-Commentary-Final-070616

European Commission. (2019). Independent High-Level Expert Group Set up by European Commission. *Ethics Guidelines for Trustworthy AI*. Retrieved December 10, from https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=60419

European Commission. Joint Research Centre. (2020). AI Watch, artificial intelligence in public services: Overview of the use and impact of AI in public services in the EU. Publications Office. Retrieved December 10, from https://data.europa.eu/doi/10.2760/039619

European Commission. (2021). Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Ammending Certain Union Legislative Acts. Retrieved December 10, from https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=celex:52021PC0206

Hamilton, M. (2019). The sexist algorithm. *Behavioral Sciences & the Law*, *37*(2), 145–157. Retrieved December 10, from https://doi.org/10.1002/bsl.2406

Harvard Law Review Association (2017). Criminal Law—Sentencing Guidelines—Wisconsin Supreme Court Requires Warning Before Use of Algorithmic Risk Assessments in Sentencing. —"State v. Loomis", 881 N.W.2d 749 (Wis. 2016). (2017). *Harvard Law Review, 130*(5), 1530–1537.

Hede, K., New machine learning tool tracks urban traffic congestion. (December 2, 2020). EurekAlert! Retrieved 3 December 2023, from https://www.eurekalert.org/news-releases/882754

Inter-Agency Working Group on Artificial Intelligence Principles for the Ethical Use of AI in the UN System_1.pdf. (n.d.). Retrieved 3 December 2023, fromhttps://unsceb.org/sites/default/files/2022/09/Principles%20for%20the%20Ethical%20Use%20of%20AI%20in%20the%20UN%20System_1.pdf

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. Nature Machine Intelligence, 1(9), Article 9. Retrieved December 10, from https://doi.org/10.1038/s42256-019-0088-2

Kozyrkov, C. (2023, April 29). Classification, regression, and prediction—What's the difference? Medium. Retrieved December 10, from https://towardsdatascience.com/classification-regression-and-prediction-whats-the-difference-5423d9efe4ec

Larson, J., Mattu, J. L., Kirchner, L., & Angwin, J. (2023). How We Analyzed the COMPAS Recidivism Algorithm. ProPublica. Retrieved 2 December 2023, from https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm

Mac Síthigh, D., & Siems, M. (2019). The Chinese Social Credit System: A Model for Other Countries? *The Modern Law Review, 82*(6), 1034–1071.

OECD. (n.d.). OECD AI Principles. An Overview. Retrieved December 10, from https://oecd.ai/en/ai-principles

Office for Artificial Intelligence. (2020). A Guide to Using Artifical Intelligence in the Public Sector. Retrieved December 10, from https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/964787/A_guide_to_using_AI_in_the_public_sector__Mobile_version_.pdf

Stankovich, M., Behrens, E., & Burchell, J. (2023). Toward Meaningful Transparency and Accountability of AI Algorithms in Public Service Delivery. Retrieved December 10, from https://www.dai.com/uploads/ai-in-public-service.pdf

UNESCO (2021). Recommendation on the Ethics of Artificial Intelligence. UNESCO Digital Library. (n.d.). Retrieved 3 December 2023, from https://unesdoc.unesco.org/ark:/48223/pf0000381137.

Van Noordt, C., & Misuraca, G. (2019). New Wine in Old Bottles: Chatbots in Government: Exploring the Transformative Impact of Chatbots in Public Service Delivery. In P. Panagiotopoulos, N. Edelmann, O. Glassey, G. Misuraca, P. Parycek, T. Lampoltshammer, & B. Re (Eds.), *Electronic Participation* (Vol. 11686, pp. 49–59). Springer International Publishing. Retrieved December 10, from https://doi.org/10.1007/978-3-030-27397-2_5.

World Bank. (2020). Artificial Intelligence in the Public Sector: Maximizing Opportunities, Managing Risks. Retrieved December 10, from https://openknowledge.worldbank.org/entities/publication/2c62df59-a40b-5d54-8d7d-3ee9d6c61715

Zalnieriute, M., & Gould-Fensom, O. (2019). Artificial Intelligence: Australia's Ethics Framework Submission to the Department of Industry, Innovation and Science. *SSRN Electronic Journal, 40*. Retrieved December 10, from https://doi.org/10.2139/ssrn.3399276.