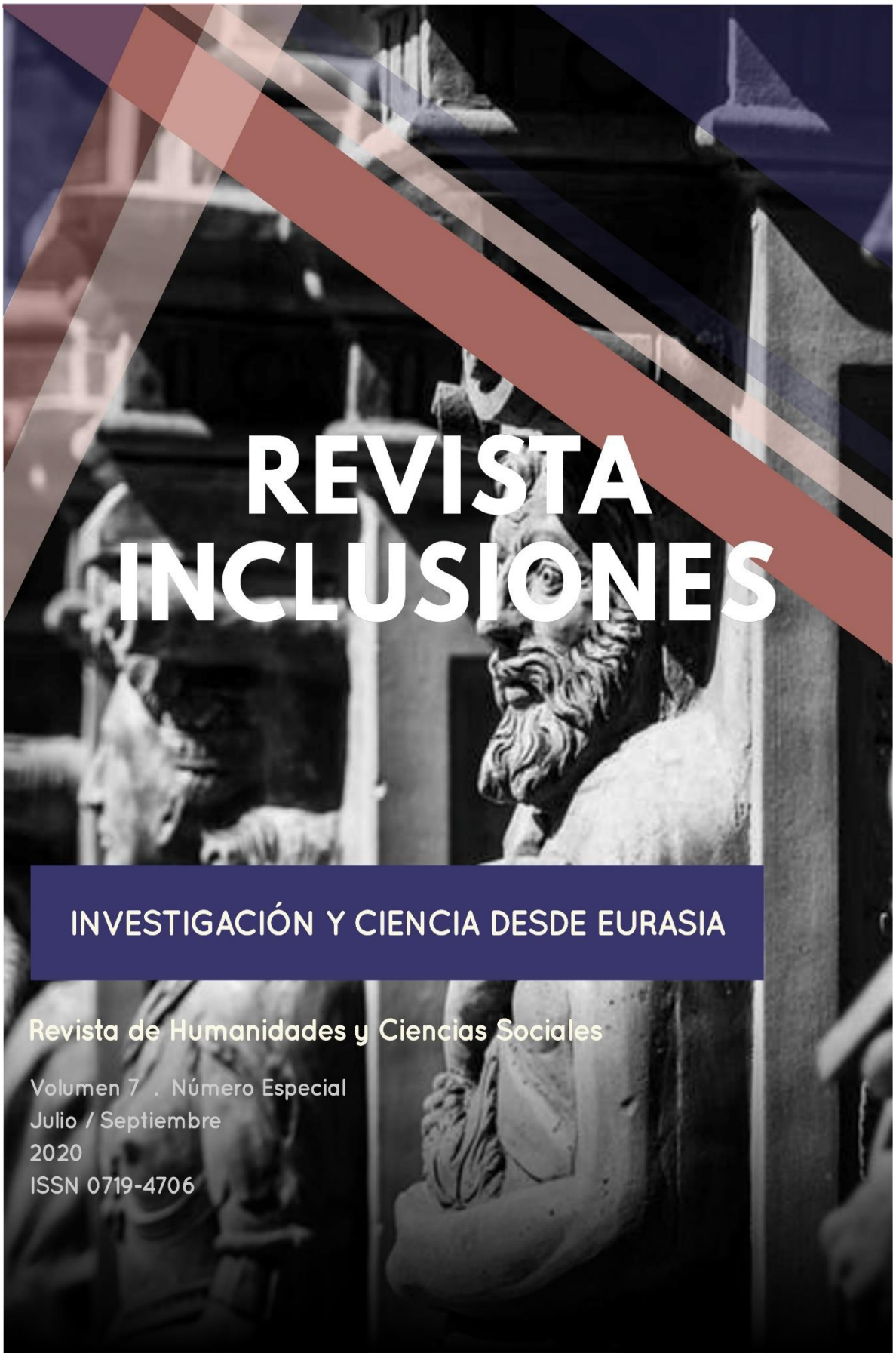

REVISTA INCLUSIONES

INVESTIGACIÓN Y CIENCIA DESDE EURASIA

Revista de Humanidades y Ciencias Sociales

Volumen 7 . Número Especial
Julio / Septiembre
2020
ISSN 0719-4706

REVISTA
INCLUSIONES
REVISTA DE HUMANIDADES
Y CIENCIAS SOCIALES

CUADERNOS DE SOFÍA
EDITORIAL

REVISTA INCLUSIONES
REVISTA DE HUMANIDADES Y CIENCIAS SOCIALES

CUADERNOS DE SOFÍA EDITORIAL

Editorial Cuadernos de Sofía
Santiago – Chile
Representante Legal
Juan Guillermo Estay Sepúlveda Editorial

REVISTA INCLUSIONES
REVISTA DE HUMANIDADES Y CIENCIAS SOCIALES

Indización, Repositorios y Bases de Datos Académicas

Revista Inclusiones, se encuentra indizada en:

BIBLIOTECA UNIVERSIDAD DE CONCEPCIÓN

REVISTA INCLUSIONES
REVISTA DE HUMANIDADES Y CIENCIAS SOCIALES

CUADERNOS DE SOFÍA EDITORIAL

ISSN 0719-4706 - Volumen 7 / Número Especial / Julio – Septiembre 2020 pp. 324-333

BLACK BOX: TRANSPARENCY AND ACCOUNTABILITY OF AUTOMATED DECISION-MAKING SYSTEMS

Ph. D. (C) Dmitry Kuteynikov
Kutafin Moscow State Law University
ORCID: 0000-0003-1448-3085
kuteynikov@me.com
Ph. D. Osman Izhaev
The Moscow City Election Commission, Russian Federation
ORCID: 0000-0003-3777-8927
izhaev.osman@gmail.com
Dr. Valerian Lebedev
Kutafin Moscow State Law University, Russia
ORCID: 0000-0002-7642-1325
lebedev_va@bk.ru
Ph. D. (C) Sergey Zenin
Moscow State Law University, Russia
South-Ural State University, Russia
ORCID: 0000-0002-4520-757X
zeninsergei@mail.ru

Fecha de Recepción: 06 de marzo de 2020 – **Fecha Revisión**: 30 de abril de 2020

Fecha de Aceptación: 22 de junio de 2020 – **Fecha de Publicación**: 01 de julio de 2020

Abstract

The present article analyzes the legal approaches of Europe and the USA, which underlay various measures taken to minimize the risks of human rights violations when using automated decision-making systems in society. The methodological basis of the research included general scientific methods of cognition, namely, the principle of objectivity, consistency, induction, and deduction. In Europe and the USA, different concepts of legal regulation of issues concerning algorithmic accountability and transparency are applied. In Europe, the audit of the automated decision-making systems is conducted through the legislation on protecting personal data. The study concludes that currently, this act does not impose a legal obligation on the controllers to disclose technical information, i.e. to open a black box to the subject of personal data. This may happen in the long run, when the legislative bodies will adopt acts that specify the provisions of the General Data Protection Regulation (GDPR), which now implies that the controller must provide the personal data subject, in respect of which the algorithm is used, with meaningful information about the logic of decisions made. It is concluded that this approach is characterized by the primacy of the interests of personal data subjects over those who derive economic and other benefits from the use of algorithms. The review of US jurisdiction has shown that there is no a comprehensive legal act regulating issues of algorithmic accountability and transparency. Certain regulatory requirements are contained in various anti-discrimination acts that regulate specific areas of human activity. It is concluded that anti-discrimination laws are not a suitable tool for resolving issues arising during the application of algorithms. Also, several current legislative initiatives at the federal and state levels were analyzed, which propose to introduce a mandatory assessment of the impact of the automated decision-making system. These initiatives involve the disclosure of a certain list of information about the operation of the algorithm. It is noted that the USA concept is more preferable for entities that use algorithms for their benefit.

Keywords

Automated decision-making systems – Accountability – Transparency – Human rights – Controllers

Para Citar este Artículo:

Kuteynikov, Dmitry; Izhaev, Osman; Lebedev, Valerian y Zenin, Sergey. Black box: transparency and accountability of automated decision-making systems. Revista Inclusiones Vol: 7 num Especial (2020): 324-333.

Introduction

In the contemporary world, there is a sustainable development trend and mass implementation of computer algorithms that can perceive the environment, analyze the information obtained, and make rational decisions on this basis at varying degrees of autonomy. The peculiarity of these algorithms is that in some cases restoring the course of made decisions with a high degree of confidence is quite problematic[1].

This problem becomes particularly important if using such algorithms in the public sphere. The decisions made by such algorithms may have legal significance or be the basis for a person to make legally significant decisions.

Thus, the subject of the relevant legal relations does not have the opportunity to assess the validity and fairness of such a decision, and thus de facto loses the right to appeal against it. However, this secrecy of decision-making may hide high risks of violation of the individual's rights. To date, sufficient experience in using algorithms has been accumulated. In particular, there are well-known cases of their negative impact on civil[2], political, economic, and social[3] human rights and freedoms.

Literature Review

Many experts have dealt with the problem of transparency and accountability of automated decision-making systems. Thus, J. Burrell[1] in his research examined the issues of understanding the opacity of machine learning algorithms. Angwin, J., J. Larson, S. Mattu, and L. Kirchner considered the possibilities of using the software in the field of crime forecasting. Datta, A., M.S. Tschantz, and A. Datta analyzed automated experiments with advertising privacy settings. Selbst A.D., and J. Pawles[4], when studying the operation of automated systems, paid attention to meaningful information and the protection of rights to confidential information. Mendoza, Isaac, and Beygreyva, studied the issues of automated decisions, as well as the legal status, and features of such decisions. Many other researchers have considered the concerned topic, however, the works performed do not sufficiently analyze the legal approaches of Europe and the USA, which serve frameworks for various measures to minimize the risks of human rights violations when using automated decision-making systems in society.

Methods

The methodological basis of the research included general scientific methods of cognition, namely, the principle of objectivity, consistency, induction, and deduction. General scientific methods of cognition were complimented by partial scientific methods such as descriptive, linguistic, comparative, and legal techniques. The authors of the article aimed

[1] J. Burrell, How the machine "thinks". Understanding opacity in machine learning algorithms. Big Data & Society num 1-12 (2016).

[2] J. Angwin; J. Larson; S. Mattu y L. Kirchner, There's software used across the country to predict future criminals. And it's biased against blacks. (May 23, 2016). Retrieved from: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

[3] A. Datta; M. C. Tschantz y A. Datta, "Automated Experiments on Ad Privacy Settings: A Tale of Opacity, Choice, and Discrimination", Proceedings on Privacy Enhancing Technologies num 1 (2015): 92-112.

[4] A. D. Selbst y J. Powles, "Meaningful information and the right to explanation", International Data Privacy Law Vol: 7 num 4 (2017): 233-242.

at reviewing the legal approaches available in the world to address issues related to accountability and transparency of algorithms. Due to the limited scope of the present work, just two opposite concepts were analyzed, which were most clearly reflected in the approaches used in Europe and the USA.

The first section is dealing with the study of the provisions of the General Data Protection Regulation (further - Regulation), which establishes certain requirements when providing information to personal data subjects in cases of using automated decision-making systems, as well as their assessment, presented in the scientific literature.

The second section shows the key features of the legal approach in the USA towards regulating the use of algorithms at both the federal and individual state levels. Some proposals for improving mechanisms, available in scientific and practical sources, aimed at ensuring accountability and transparency of algorithms were also analyzed.

Results

At the level of the European Union, regulatory requirements for algorithmic accountability and transparency are incorporated into legislation governing the personal data circulation. The controllers are obliged to provide a certain amount of information to the personal data subject related to the decision-making concerning the solution algorithm concerning the personal data subject.

The legal approach is based on the primacy of human rights over the interests of entities that derive economic and other benefits from the use of algorithms. However, this approach has not yet been fully implemented, since there is no Regulation specification regarding the operation of algorithms.

In a strict sense of the current provisions of the Regulation, it follows that they do not impose on the controllers a legal obligation to disclose technical information (open a black box) about the operation of the algorithm corresponding to the personal data subject.

The USA does not have a comprehensive regulatory framework that sets any requirements for algorithmic accountability and transparency, which is a key feature of the US approach. In practice, regulation is carried out ad hoc regarding individual provisions of various anti-discrimination acts. At the federal and state levels, it is proposed to adopt legal arrangements aimed at introducing algorithmic accountability and transparency. The existing legislative initiatives coincide in terms that they propose the establishment of a *mandatory assessment of the impact of an automated decision-making system*, which leads to the disclosure of a certain list of information about the operation of the algorithm.

European approach to information disclosure about the logic of automated decision-making systems

In the European Union, the functioning of algorithms is regulated by Regulation. The requirements of the Regulation to controllers concerning information disclosure to explain to the data subject the logic of the decision made by the algorithm in relation to them have been studied in several scientific articles[5]. The issues of transparency and accountability of

[5] G. Malgieri, “Automated decision-making in the EU Member States: The right to explanation and other “suitable safeguards” in the national legislations”, Computer law & security review num 35

the algorithms include the provisions of Regulation, according to which the controller, in cases of obtaining personal data directly from the data subject or third parties, must provide, and the subject has the right to obtain information “about the presence of automated decision-making system including the profiling referred to in paragraphs 1 and 4 of article 22 of the Regulation, and at least in these cases, to obtain full information about the logic of this system, as well as concerning the significance and expected consequences of such processing for the data subject" (articles 13-15 of the Regulation). At that, paragraph 1 of article 22 of the Regulation states that “the data subject has the right not to be the subject of such a decision that is based solely on automatic processing, including profiling, which creates legal consequences for the data subject, or which similarly significantly affects the data subject". Scientists interpret the above regulations in different ways[6].

Some scientists argue that the relevant provisions imply the obligation of controllers when using the algorithms to disclose full information about the logic of decision-making to the data subject, including technical information, which means the disclosure of the black box[7]. This approach involves a broad interpretation of the Regulation norms.

Others believe that the controllers are not required to provide detailed information, but should only notify the personal data subject (ex-ante, rather than ex-post) about how the algorithm makes a decision in a general context[8]. The reason for this restrictive interpretation is that the Regulation does not specify what information should be provided to the personal data subject. Besides, an important argument here is that article 22 of the Regulation specifies that the provisions are applied in cases where the system has made a decision solely independently, i.e. without any human intervention. In practice, a person often takes a formal part in the decision which is made by the algorithm, which allows the controller to avoid compliance with the relevant provisions of the Regulation.

On the one hand, it follows from the practice of implementing data provisions in national legislation by the European Union member states that the vast majority of these countries did not use a lateral approach, and only a small number of countries have adopted additional acts obliging them to provide detailed information to data subjects about the functioning of algorithms[9]. On the other hand, there is also the standpoint of the Working Group on the issue of human rights in the field of personal data processing, which operates under the auspices of the European Commission. According to this standpoint, from the wording “decisions based solely on automated processing” it does not follow that if a person takes a formal part, the controller is released from the obligation to provide relevant information to personal data subjects[10].

(2019): 6.

[6] A. D. Selbst y J. Powles, “Meaningful information and the right to explanation”, International Data Privacy Law Vol: 7 num 4 (2017): 233-242.

[7] S. Wachter; B. Mittelstadt y L. Floridi, “Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation”, International Data Privacy Law Vol: 7 num 2 (2017): 76–99.

[8] T. Synodinou; P. Jougleux; C. Markou y T. Prastitou, “EU Internet Law: Regulation and Enforcement (Springer, Forthcoming)”, University of Oslo Faculty of Law Research Paper num 20 (2017).

[9] G. Malgieri y G. Comandé, “Why a Right to Legibility of Automated Decision-Making Exists in the General Data Protection Regulation”, International Data Privacy Law, Vol: 7 num 4 (2017): 243–265.

[10] Guidelines on Automated individual decision-making and Profiling for the purposes of Regulation 2016/679. (August 22, 2018). Retrieved from: https://ec.europa.eu/newsroom/article29/item-detail.cfm?item_id=612053

It seems that literal reading of the Regulation does not imply that the controllers must disclose detailed information about the algorithm. However, this issue remains open and will be resolved in the future by law enforcement practice.

Even though current regulations at the level of the European Union still do not regulate these issues in detail, in the scientific and practical sphere, there is a tendency to search for a balanced legal model that would contain all the necessary legal requirements for controllers and provide a sufficient list of legal remedies for personal data subjects, i.e. persons in respect with whom the algorithm will make legally significant decisions. In particular, this is evidenced by the vast amount of scientific research carried out under the auspices of the European Union and the Council of Europe[11].

Consider one of these comprehensive studies which focus on social, technical, and regulatory issues related to algorithmic transparency and accountability. The report published based on its results notes that it is appropriate to talk about transparency in seven areas of machine learning, namely: data (disclosure of the data source, and methods by which they were verified as unbiased and representative, information about the self-updating and learning processes), directly the algorithm (disclosure of information about software development, code, and operation mode), objectives (transparency regarding the tasks and priorities of the algorithm), results (to require from producers or operators transparency of information about the operation of algorithmic systems and the emerging consequences), compliance (regular verification of the algorithm by the operators and informing about its results), influence (information about the presence of entities in whose interests the algorithm is used), usage (information about exactly what kind of personal data and to what extent are used by the algorithm). However, the authors argue that transparency can relate either to the system in general or to individual decision-making. They conclude that it is impractical to require openness of the entire algorithm since this is a very resource-intensive process and does not lead to a meaningful explanation of the logic of the decisions made. The authors insist that a more reasonable way is to establish measures aimed at disclosing information about the logic of individual decision-making. As proposed, these measures can include input data analysis, statistical explanation, design/code validation, and statistical analysis, as well as determining the sensitivity of individual data (which variables exactly determine the result)[12] .

Thus, the European approach is characterized by a focus on disclosure of information concerning the functioning of the algorithm so that subjects have a full opportunity to obtain a comprehensive justification for the decision made in their regard. However, this approach does not necessarily lead to the disclosure of the black box. Rather, the model is based on the thesis that if there are a real need and expediency, in the absence of alternative ways to explain the logic of decision-making, it is possible to establish requirements for the disclosure of technical features/code of the corresponding algorithm.

[11] White Paper on Artificial Intelligence - A European approach to excellence and trust. (February 19, 2020). Retrieved from https://ec.europa.eu/info/sites/info/files/commission-white-paper-artificial-intelligence-feb2020_en.pdf

[12] E. A. Kirillova; A. V. Pavlyuk; O. E. Blinkov; E. V. Blinkova y E. L. Sidorenko, “Digital inheritance of social media accounts”, International Journal of Engineering and Advanced Technology Vol: 8 num 4 (2019): 963-967.

Algorithmic accountability and transparency in the USA

Currently, in the USA there is no federal law that regulates issues related to the transparency and accountability of algorithms. This is due to numerous factors, including the peculiarities of the US legal system.

While the Regulation is a framework law regulating personal data and involves issues of algorithmic accountability under the frame of such regulation, in the USA, the turnover of personal data is governed by the conceptually another way. In Europe, the well-known rule is applied which states that everything that is not expressly permitted is prohibited, while the USA practices the opposite approach – everything that is not expressly prohibited is permitted. Individual regulations that can be applied in disputes involving the use of automated decision-making systems are contained in various anti-discrimination laws at the federal and state levels.

For example, the federal law named Fair Housing Act requires lenders to indicate in a special notification of the applicant for a loan the factors leading to the refusal to allow of loan, and other negative consequences. The Equal Credit Opportunity Act requires that lenders provide a statement[s] of reasons for negative decisions. Following similar acts, it is prohibited to take into account characteristics that are prohibited for processing (for example, gender and race) when making decisions. However, this approach was developed to prevent discrimination resulting from decisions made by a human, and therefore does not provide an exhaustive set of remedies against discrimination by automated decision-making systems[13].

One of the attempts to resolve this issue at the federal level is the adoption by the United States Congress in 2019 of the Algorithmic Accountability Act. This Act provides for the obligation of certain persons, who use automated decision-making systems in their activities, to conduct automated decision system impact assessment during its development and use, including design and training data for compliance with the requirements of accuracy, fairness, impartiality, nondiscrimination, confidentiality, and security. Besides, the supervisory authority must be provided with a detailed description of the automated decision-making system, its design, training, data, and its purpose[14]. At the same time, it is not clear what is meant by a detailed description. The results of the conducted assessment are published at the operator's discretion.

A similar act is being proposed in the state of California. The act contains the concept of an Automated Decision System, which is a computational process, including a process based on machine learning, statistics, or other data processing methods, artificial intelligence techniques that can make decisions or facilitate human decision-making that affects people. Besides, it is also proposed to introduce the term of Automated Decision System Impact Assessment, which means a study that assesses the automated decision system and its development process, including, but not limited to data on the design and training of the system regarding its accuracy, fairness, bias, nondiscrimination, confidentiality, and security.

[13] T. B. Gillis y J. Spiess, “Big Data and Discrimination”, University of Chicago Law Review num 459 (2019): 462.

[14] M.A. Bruckner, “Promise and Perils of Algorithmic Lenders’ Use of Big Data”, Chicago-Kent Law Review Vol: 93 num 1 (2019): 56-57.

It is also proposed to create an Advisory Task Force on automated decision-making systems by March 1, 2022, to review and provide recommendations on their use in business, government, and other areas.

In the states of New York and Washington, legislation is being developed for the use of automated decision-making systems in the public sector (government decisions and legal proceedings).

Discussion

The problem of algorithmic accountability has been repeatedly studied at the intersection of legal and computer sciences. American scientists highlight both legal and technical requirements for transparency and nondiscrimination of algorithms in a single study. In their opinion, the transparency of algorithms and their subsequent audit can only help to prevent undesirable results. Ideally, when developing the algorithm, these types of ex-post analysis should be used in tandem with powerful ex-ante methods.

These technical methods include Software Verification, Cryptographic Commitments, Zero-Knowledge Proofs, and Fair Random Choices. Using these methods from the very beginning at developing the algorithm will allow achieving the necessary results due to algorithm design. For machine learning related algorithms, the following methods can also be used: Learning from Experience, that is, the assumption of a random sample from real life, on which the algorithm could be trained if the initial data are not representative enough; and Fair Machine Learning, which involves creating algorithms based on the concept of human justice. Legal requirements include reducing ambiguity or double meaning in legislation, introducing a permanent expert position in the courts to explain the application of the law by algorithms as a recommendation body, and the need to develop legislation in the field of algorithmic accountability without requiring full transparency of algorithms and free access to them. Also noteworthy is the disclosure of information about the algorithm's operation following the concept of counterfactual explanations[15]. The authors of this method assume that providing a person with technical (internal) information is of no value since the average person is not able to understand the logic of the decision based on such information. The concept of counterfactual explanations involves familiarizing a person with the variables (income, gender, age, etc.) which most influenced the decision made by the algorithm, as well as what exactly needs to be changed to achieve the desired result[16]. The simple counterfactual explanations method can be illustrated by the following example:

If X earned USD 500,000 a year, X would get approved for a loan of Y.

Thus, a hypothesis is put forward that contradicts the actual circumstances (in fact, X does not earn USD 500,000), but X understands what circumstance was the main reason for making a specific decision. This allows X not only to understand why he was not given a loan but also to know what needs to be changed to achieve the desired result. The value of this method is that it does not oblige providing technical information (disclosing the black box), but at the same time allows getting the necessary information that is important for evaluating the fairness and legality of the made decision. The drawback of this method is

[15] S. Wachter; B. Mittelstadt & C. Russell, “Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR”, Harvard journal of law & technology num 31 (2018): 841-887.

[16] T. Wischmeyer, “Artificial Intelligence and Transparency: Opening the Black Box”, Regulating Artificial Intelligence (2020): 75–102.

that sometimes the algorithm can use a very large number of variables to make a decision, and it becomes impossible to make an unambiguous conclusion about what specific factors have influenced a made decision.

Conclusion

Europe and the USA apply different concepts of legal regulation of algorithmic accountability and transparency issues. In Europe, the audit of automated decision-making systems is implemented through data protection legislation. Currently, this act does not impose a legal obligation on controllers to disclose technical information, i.e. to open a black box, to the personal data subject. This may happen in the long run, when the legislative bodies will adopt acts that specify the provisions of the GDPR, which currently imply that the controller must provide the personal data subject, in respect of whom the algorithm is used, with meaningful information about the logic of decisions made. It is concluded that this approach is characterized by the primacy of the interests of personal data subjects over those who derive economic and other benefits from the use of algorithms. The review of USA jurisdiction has shown that there is no comprehensive legal act regulating issues of algorithmic accountability and transparency. Certain regulatory requirements are contained in various anti-discrimination acts that regulate certain areas of human activity. Anti-discrimination laws are not a suitable tool for resolving issues that arise during the application of algorithms. The authors have analyzed several current legislative initiatives at the federal and state levels, which propose to introduce a mandatory impact assessment of the automated decision-making system, which involves the disclosure of a certain list of information about the operation of the algorithm. It is noted that the US concept is more preferable for subjects that use algorithms for their benefit.

Acknowledgments

The reported study was funded by the Russian Foundation for Basic Research (RFBR) according to the research project No. 18-29-16193.

References

Angwin, J.; Larson, J.; Mattu, S.; Kirchner y L. Machine Bias, There's software used across the country to predict future criminals. And it's biased against blacks. (May 23, 2016). Retrieved from: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Bruckner, M. A. "Promise and Perils of Algorithmic Lenders' Use of Big Data". Chicago-Kent Law Review Vol: 93 num 1 (2019): 56-57.

Burrell, J. "How the machine "thinks". Understanding opacity in machine learning algorithms". Big Data & Society num 1-12 (2016).

Datta, A.; Tschantz, M. C. y Datta, A. "Automated Experiments on Ad Privacy Settings: A Tale of Opacity, Choice, and Discrimination". Proceedings on Privacy Enhancing Technologies num 1 (2015): 92-112.

Gillis, T. B. y Spiess, J. "Big Data and Discrimination". University of Chicago Law Review num 459 (2019): 462.

Guidelines on Automated individual decision-making and Profiling for the purposes of Regulation 2016/679. (August 22, 2018). Retrieved from: https://ec.europa.eu/newsroom/article29/item-detail.cfm?item_id=612053

Kirillova, E. A.; Pavlyuk, A. V.; Blinkov, O. E. Blinkova; E. V. Sidorenko y E. L. "Digital inheritance of social media accounts". International Journal of Engineering and Advanced Technology Vol: 8 num 4 (2019): 963-967.

Malgieri, G. "Automated decision-making in the EU Member States: The right to explanation and other "suitable safeguards" in the national legislations". Computer law & security review num 35 (2019).

Malgieri, G. y Comandé, G. "Why a Right to Legibility of Automated Decision-Making Exists in the General Data Protection Regulation". International Data Privacy Law, Vol: 7 num 4 (2017): 243–265.

Selbst, A. D. y Powles, J. "Meaningful information and the right to explanation". International Data Privacy Law Vol: 7 num 4 (2017): 233-242.

Synodinou, T.; Jougleux, P.; Markou, C. y Prastitou, T. "EU Internet Law: Regulation and Enforcement (Springer, Forthcoming)". University of Oslo Faculty of Law Research Paper num 20 (2017).

Wachter, S.; Mittelstadt, B. y Russell, C. "Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR". Harvard journal of law & technology num 31 (2018): 841-887.

Wachter, S.; Mittelstadt, B. y Floridi, L. "Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation". International Data Privacy Law Vol: 7 num 2 (2017): 76–99.

White Paper on Artificial Intelligence - A European approach to excellence and trust. (February 19, 2020). Retrieved from https://ec.europa.eu/info/sites/info/files/commission-white-paper-artificial-intelligence-feb2020_en.pdf

Wischmeyer, T. "Artificial Intelligence and Transparency: Opening the Black Box". Regulating Artificial Intelligence (2020): 75–102.

CUADERNOS DE SOFÍA
EDITORIAL